Spring 2016

Individual differences in speech perception: sources, functions, and consequences of phoneme categorization gradiency

Efthymia Evangelia Kapnoula
University of Iowa

INDIVIDUAL DIFFERENCES IN SPEECH PERCEPTION: SOURCES, FUNCTIONS, AND CONSEQUENCES OF PHONEME CATEGORIZATION GRADIENCY

by

Efthymia Evangelia Kapnoula

A thesis submitted in partial fulfilment
of the requirements for the Doctor of Philosophy
degree in Psychology in the
Graduate College of
The University of Iowa

May 2016

Thesis Supervisor: Professor Bob McMurray

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Efthymia Evangelia Kapnoula

has been approved by the Examining Committee for
the thesis requirement for the Doctor of Philosophy degree
in Psychology at the May 2016 graduation.

Thesis Committee: _____
Bob McMurray, Thesis Supervisor

_____
Jan Edwards

_____
Susan Wagner Cook

_____
Prahlad Gupta

_____
Teresa Treat

Στους γονείς μου,

ACKNOWLEDGEMENTS

First, I would like to thank my advisor, Bob McMurray, who has been an extraordinarily boundless source of support and inspiration. I feel very fortunate to have had him as a mentor and role model of an outstanding member of the community within and outside academia. I also thank my committee members, Teresa Treat, Prahlad Gupta, Thomas Farmer, Susan Wagner Cook, and, especially, Jan Edwards for her invaluable help in this project, as well as our collaborators, Eunjong Kong and Matt Winn, for their vital contributions. I am also grateful to the members of the University of Iowa Language Discussion Group and the DeLTA center, as well as Gregg Oden and Jan Wessel for their feedback.

I would like to thank the wonderful members and friends of the MACLab for all their help and support. The daily MACLab routine was a core part of my graduate training and has deeply shaped the way I aim to practice scientific research. Special thanks go to Tanja Römbke, one of the most perceptive and talented people I have met; Jamie Klein, for her help, especially with recruiting and running participants; Ariane Rhone, for helping me with stimulus construction as the fricative manipulation wizard that she is, but also for being a brilliant scholar and role model, along with Ashley Farris-Trimble; Marcus Galle, for being an amazing teacher during my first years at the MACLab; Joe Toscano, for introducing me to the magical world of ERPs and VOT manipulation; Keith Apfelbaum, Dan McEchron, Aimee Marino, Emily Shaw, Kathryn Hiolski, Libo Zhao, Kayleen Schreiber, and Claire Goodwin, for all the small and big ways in which you helped me – thank you.

I feel very grateful to my family and friends in Greece; especially my sister for her unconditional support, my wonderful parents for their love, support, and patience, and my grandparents for inspiring me to be my best self. Angela Kiofiri, thank you for being a loyal and supportive friend despite the distance. Also, a big thank you goes to Bret Feddern for his unlimited patience and support and, of course, a very special thank you to the one and only Litsa Cheimariou for all the ways in which she has stood by me in this journey.

Finally, Thanassi Protopapas, thank you for inspiring and encouraging me to follow my dreams. Gerry Altmann, thank you for your guidance, support, and the occasional words of wisdom.

Oh, and coffee – thank you, I could not have done it without you.

ABSTRACT

During spoken language comprehension, listeners transform continuous acoustic cues into categories (e.g. /b/ and /p/). While longstanding research suggests that phoneme categories are activated in a gradient way, there are also clear individual differences, with more gradient categorization being linked to various communication impairments like dyslexia and specific language impairment (Joanisse, Manis, Keating, & Seidenberg, 2000; López-Zamora, Luque, Álvarez, & Cobos, 2012; Serniclaes, Van Heghe, Mousty, Carré, & Sprenger-Charolles, 2004; Werker & Tees, 1987).

Crucially, most studies have used two-alternative forced choice (2AFC) tasks to measure the sharpness of between-category boundaries. Here we propose an alternative paradigm that allows us to measure categorization gradiency in a more direct way. We then use this measure in an individual differences paradigm to: (a) examine the nature of categorization gradiency, (b) explore its links to different aspects of speech perception and other cognitive processes, (c) test different hypotheses about its sources, (d) evaluate its (positive/negative) role in spoken language comprehension, and (e) assess whether it can be modified via training.

Our results provide validation for this new method of assessing phoneme categorization gradiency and offer valuable insights into the mechanisms that underlie speech perception.

PUBLIC ABSTRACT

Understanding spoken language is something we may take for granted. However, it is a quite remarkable skill, especially if we consider how often we deal with noise and ambiguities in everyday interactions (due to background noise, unfamiliar accents etc.). Even though listeners typically cope with such difficulties in an effortless manner, we do not yet have a comprehensive understanding of the perceptual and cognitive mechanisms that allow for this.

One core issue that remains unclear is how do listeners distinguish between similar speech sounds (e.g. between the words *beach* and *peach*). On the one hand, there is robust evidence showing that typical listeners perceive speech in great detail and they use this information in a gradient manner. However, according to an alternative account, listeners are better off focusing only on that portion of the speech signal that is relevant for the ultimate categorization decision. Furthermore, divergence from this latter pattern has been considered a marker of atypical or non-optimal language processing.

Interestingly, recent findings suggest that there are substantial differences between listeners in how they process the speech signal. By studying these differences we can achieve a better understanding of how listeners process spoken language, but also identify situations in which maintaining detailed speech information is advantageous or detrimental for language comprehension.

The goal of the present study is to develop a novel way of studying such individual differences in order to address fundamental questions about speech perception processes. Ultimately, the study of such differences will lead to a more comprehensive understanding of both typical and atypical patterns of language processing.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: GRADIENCY IN SPEECH PERCEPTION

1.1 Phoneme categorization gradiency

When comprehending spoken language, auditory input varies along multiple acoustic dimensions (e.g. formant frequencies) that are continuous and highly variable. Listeners process this signal to extract linguistically relevant information like phonemes and features, which they use to recognize words. This process represents a transformation from *continuous* input that is both ambiguous and redundant, into relatively *discrete* categories, such as features, phonemes, and words.

During this process, listeners are faced with a critical problem: the same cue[1] values (e.g. the same formant frequencies) do not always map onto the same phonemic category. That is, stimuli with the same acoustic cue values may correspond to different phonemic categories depending on the context (e.g., speech rate or talker's gender). For example, a stimulus with a voice onset time (VOT) of 20 ms could be a /b/ in slow speech or a /p/ in fast speech. In fact, despite over 40 years of research, phoneticians and speech scientists have identified few (if any) acoustic cues that unambiguously identify a phoneme across different contexts (e.g., McMurray & Jongman, 2015; Ohala, 1996).

Traditional approaches have suggested that this problem of lack of invariance is solved via the use of specialized mechanisms that discard irrelevant (i.e. within-category) information, leading to the perception of distinct phonemic categories (Liberman & Whalen, 2000; Liberman, Harris, Hoffman, & Griffith, 1957). In contrast to this

[1] Even though we use the term "cue" here, we do not make a strong theoretical commitment as to the kind of auditory information this term entails.

hypothesis, recent studies have shown that the modal[2] listener maintains fine-grained information that is seemingly irrelevant for discriminating between phonemic categories (i.e. within-category information; Massaro & Cohen, 1983a; McMurray, Tanenhaus, & Aslin, 2002; Toscano, McMurray, Dennhardt, & Luck, 2010).

Despite the robust evidence that gradiency in phoneme categorization is a fundamental aspect of speech perception, we do not fully understand the mechanisms that subserve it, and we do not have a clear view of the functional role of maintaining within-category information. For example, while there are theoretical reasons why a gradient representation may be useful (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Kleinschmidt & Jaeger, 2015; McMurray & Farris-Trimble, 2012; Oden & Massaro, 1978), there is little empirical data that speaks to the issue of why listeners would want to maintain such detail (though see McMurray, Tanenhaus, & Aslin, 2009) and what they might do with it. In order to address these issues, we need to achieve a better understanding of the nature of phoneme categorization gradiency on a mechanistic level.

Here we address these issues within an individual differences approach. We next describe the basic problem (lack of invariance) that first sparked the question of whether phoneme categorization is gradient or categorical, then we review the literature for and against the contrasting views, and describe findings showing evidence for substantial differences between typical and atypical populations, as well as some preliminary results showing individual differences within typical populations. At the end of this chapter, we present the goals of the present work and the specific issues addressed by each experiment.

[2] By "modal listener" we refer to the most common pattern of behavior among typical listeners.

1.2 The problem of lack of invariance

During speech perception, listeners use whatever acoustic information is available at each point in time to recognize the words produced by a talker (McMurray & Jongman, 2011; McMurray, Clayards, Tanenhaus, & Aslin, 2008; Warren & Marslen-Wilson, 1987). This information can be commonly described (at least in part) in terms of classic acoustic/phonetic cues. For example, voice onset time (VOT) is the time between the onset of the release burst and the onset of vocal-cord vibration and it is the primary cue for distinguishing between voiced and unvoiced stop consonants (e.g. labial stops with VOTs below 20 ms are typical of a /b/ sound, while those with VOTs over 20 ms are more frequently perceived as a /p/). Critically, while the underlying acoustic cues are *continuous,* our conscious percept, as well as linguistic analyses of language, seem to reflect more or less *discrete* categories (/b/ and /p/).

Mapping continuous cues into discrete categories is quite complex. This is mainly because the same set of cue values can map onto different phoneme categories, depending on multiple factors, such as the talker's gender (Hillenbrand, Getty, Wheeler, & Clark, 1995), the neighboring speech sounds (Hillenbrand, Clark, & Nearey, 2001), and speaking rate (Miller, Green, & Reeves, 1986). For example, a fricative with 4,000 Hz peak frequency could be an /s/ spoken by a woman or an /ʃ/ spoken by a man. This is the problem of *lack of invariance*; phoneme categories do not have invariant acoustic attributes, and a single acoustic attribute cannot reliably be mapped to a single speech sound.

1.3 Categorical perception of phoneme categories

One solution to the lack of invariance problem stems from the classic phenomenon of *Categorical Perception* of speech (CP; Liberman et al., 1957). CP describes the finding that discrimination within a category (e.g. between two instances of a /b/) is poor, but discrimination of an equivalent acoustic difference that spans a category boundary is quite good. This is a behavioral phenomenon that has been extremely well replicated (e.g., Liberman & Harris, 1961; Pisoni & Tash, 1974; Repp, 1984; Schouten & Hessen, 1992 for a review).

Recent research also suggests a neural basis for CP: neuroimaging techniques reveal differences in the processing of within- versus between-category pairs of speech sounds. This has been seen using event-related potentials (ERPs) and magneto-encephalography (MEG) in the mismatch negativity (MMN) paradigm; a larger MMN is elicited when listeners hear a syllable that falls into a different category from previous syllables, than when the same acoustic discrepancy does not cross phonemic boundaries (Dehaene-Lambertz, 1997; Phillips et al., 2000; Sams, Aulanko, Aaltonen, & Näätänen, 1990; see also Chang et al., 2010).

The aforementioned behavioral and neuroimaging findings can be viewed as evidence for some kind of warping of the perceptual space that amplifies the influence of categories. For example, an instance of a [b] with a VOT of 15 ms is perceived as more similar to a [b] with a VOT of 0 ms than a [p] with a VOT of 30 ms, because the first two both map onto the same category. Such non-linear perceptual representations are particularly difficult to explain, if we assume that the basic units of speech perception are auditory, as this kind of warping would appear to violate Weber's law.

A common interpretation of these findings is that listeners are equipped with specialized processes that rapidly discard within-category variation in favor of discrete encoding at both the auditory/cue level and at the level of phoneme categories. This view of speech perception—perhaps best exemplified by *motor theory* (Liberman & Whalen, 2000)—suggests that auditory encoding is aligned to the discrete goals of the system (i.e. phoneme categorization). These mechanisms were thought to rapidly sort through the variance, discarding unnecessary detail to extract the invariant kernels of the signal. As a result, acoustic variations, arising from talker differences and/or co-articulation (i.e. the natural overlap of articulatory gestures), do not pose significant issues for speech perception, because the *underlying* representations (gestures or phonological units) can be rapidly extracted by these specialized (however ill-specified) mechanisms.

1.4 The gradient alternative

According to CP, listeners' encoding of acoustic cues is somewhat discrete, and, consequently, this information can be mapped to fairly discrete categories. However, neither of these claims have held up to scrutiny.

A number of concerns has been raised with regard to the discrimination tasks used to establish CP. Indeed, a wealth of work suggests that the degree to which discrimination shows a categorical pattern (i.e. better discrimination across a boundary) depends heavily on the task used to assess it (Carney, Widin, & Viemeister, 1977; Gerrits & Schouten, 2004; Pisoni & Lazarus, 1974; Schouten, Gerrits, & Hessen, 2003). Pisoni and Lazarus (1974), for example, showed that different discrimination tasks reveal different degrees of sensitivity to within-category differences, while Gerrits and Schouten, (2004) and

Schouten et al. (2003) investigated factors that may moderate CP effects and found that the most significant parameter is the discrimination task itself.

The key results from these investigations is that certain tasks have higher working memory demands, which in turn can lead listeners to rely on subjective labels (rather than auditory codes, which may decay more rapidly). For example, when listeners have to discriminate between sounds presented with a lengthy delay between them, the auditory traces may have faded away by the time they need to make a decision. In such cases, listeners are in a way forced to rely more heavily on phonetic labels, which could lead to a more categorical-like pattern of responses, even if the pre-categorized perceptual representation is continuous (see also Carney et al., 1977; Gerrits & Schouten, 2004; Pisoni & Tash, 1974). This suggests that CP may in fact reflect the influence of categories on participants' memory and decision processes, not on the perceptual processes per se. Indeed, when less biased discrimination measures are employed, CP-like effects disappear (Gerrits & Schouten, 2004; Massaro & Cohen, 1983; Pisoni & Lazarus, 1974).

This dependence of CP on task properties implies that encoding of speech cues may not be warped at all, but rather it may veridically reflect the input. The idea that listeners maintain within-category information about speech sounds, is featured in a number of alternative theoretical approaches (Goldinger, 1998; Kleinschmidt & Jaeger, 2015; McMurray & Jongman, 2011; Oden & Massaro, 1978), which argue that the system does in fact preserve fine-grained detail. In support of this, ERP and MEG responses to isolated words from VOT continua reflect a systematic and linear response to changes along the continuum with no evidence of warping (Frye, Fisher, & Coty,

2007; Toscano et al., 2010), and this can be observed in early phonological processing areas using fMRI (Myers & Blumstein, 2009).

Furthermore, beyond the level of auditory encoding, there is now substantial evidence that fine-grained detail is preserved at higher levels of the auditory-cognitive pathway. That is, listeners are not only sensitive to within-category differences in early stages of processing, but this information is still available downstream, at the level of lexical processing as it modulates the continuous degree to which lexical competitors (e.g., the word *beach* vs. the word *peach*) are active (Andruski, Blumstein, & Burton, 1994; McMurray et al., 2002; Utman, Blumstein, & Burton, 2000). For example, *beach* will be slightly more active (and *peach* less so) with a VOT of 0 ms than a VOT of 10 ms – even though both VOTs are clearly indicative of a /b/.

1.5 The functional role of gradiency

Maintaining within-category, continuous differences throughout levels of speech processing may allow for more flexible and efficient speech processing. There are a few ways in which this can happen.

First, processes like coarticulation and assimilation leave fine-grained, sub-categorical traces in the signal (e.g., Gow, 2001), which can be used to anticipate upcoming speech sounds speeding up processing. Multiple studies suggest that listeners take advantage of anticipatory coarticulatory information in this way (Gow, 2001; Mahr, McMillan, Saffran, Ellis Weismer, & Edwards, 2015; McMurray & Jongman, 2015; Salverda, Kleinschmidt, & Tanenhaus, 2014; Yeni–Komshian, 1981). However, as these

modifications are largely within-category, such anticipation is only possible if listeners are sensitive to this fine-grained detail.

Second, at low levels, a more or less linear response to cues (e.g., Massaro & Cohen, 1983; Toscano et al., 2010) may allow for greater flexibility in how cues map onto categories. Continuous encoding of cues may make it easier for listeners to combine multiple cues in a more sensitive way. That is, it may permit for the values of one cue to be interpreted in light of the values of other cues. Such processes may underlie the well-known trading relations that have been documented in speech perception (Repp, 1982; Summerfield & Haggard, 1977; Winn, Chatterjee, & Idsardi, 2013). This kind of combinatory process would also be necessary for accurately compensating for higher level contextual expectations—for example recoding pitch relative to the talker's mean pitch (McMurray & Jongman, 2011, 2015).

Third, gradient responding at higher levels, for example, at the level of phonemes (Miller, 1997; McMurray, Aslin, Tanenhaus, Spivey & Subik, 2008) and at the lexical level (McMurray et al., 2008; Andruski, Blumstein & Burton, 1994), suggests that the degree to which the perceptual system commits to one representation over another (e.g., /b/ vs. /p/) is a function of continuous changes in the signal. For example, a labial stop with a VOT of 5 ms activates /b/-onset words *more* than a labial stop with a VOT of 15 ms, even though both map onto the same category. Superficially, this may appear disadvantageous for speech perception, as it could slow an efficient decision. However, when we consider the variability, noise, and non-relevant information present in the speech signal, this gradiency may allow the listener to "hedge" their bets in the face of ambiguity. That is, it is precisely in these situations of ambiguity when a listener may not

want to commit too strongly and to keep their options open until more information arrives (Clayards et al., 2008; McMurray, Tanenhaus, & Aslin, 2009).

In sum, gradiency may allow the system to (a) harness fine-grained (within-category) differences that may be helpful, (b) more flexibly integrate information from multiple sources, and (c) avoid making a commitment when insufficient information is available so that listeners can flexibly adjust when new information arrives. Thus, while the somewhat empirical question of the gradient versus discrete nature of speech representations has been hotly debated (Chang et al., 2010; Gerrits & Schouten, 2004; Liberman & Whalen, 2000; Massaro & Cohen, 1983; McMurray et al., 2002; Myers & Blumstein, 2009; Toscano et al., 2010), it also has important theoretical ramifications for how effectively listeners solve a fundamental perceptual problem.

1.6 Individual differences in phoneme categorization

Despite the substantial evidence for gradiency in the modal listener, it is less clear whether there are *individual differences* in the tendency to show gradiency in speech perception. There is now mounting evidence in neuroscience for multiple pathways of speech processing (Blumstein, Myers, & Rissman, 2005; Hickok & Poeppel, 2007; Myers & Blumstein, 2009) that can be flexibly deployed under different conditions (Du, Buchsbaum, Grady, & Alain, 2014). Given this, different listeners may adopt different solutions to this problem, perhaps providing more weight to either dorsal or ventral pathways (see Ojemann, Ojemann, Lettich, & Berger, 1989 for analogous evidence in word production). Similarly, the Pisoni and Tash model of categorical perception (Pisoni & Tash, 1974) suggests that listeners have simultaneous access to both continuous

acoustic representations *and* discrete categories. Once again, this raises the possibility that listeners may weigh these two sources of information differently during speech perception.

With respect to the function of gradiency in speech perception, the possibility of individual differences raises three important questions: 1) Are different listeners gradient to varying degrees? 2) What are the underlying sources of these differences? 3) Do such differences have a positive/negative impact for speech perception as a whole?

Much of the debate around categorical versus gradient modes of perception in typical listeners concerns the degree to which a gradient representation of fine-grained detail may be adaptive (or maladaptive). In this regard, and particularly in light of our core question of individual differences, a consideration of listeners with communication disorders would be useful.

Classic and ongoing work on language-related disorders like specific language impairment (SLI) and dyslexia suggests that there are indeed significant differences between populations in the gradiency or discreteness of categorization (Coady, Evans, Mainela-Arnold, & Kluender, 2007; McMurray, Munson, & Tomblin, 2014; Robertson, Joanisse, Desroches, & Ng, 2009; Serniclaes, 2006; Sussman, 1993; Werker & Tees, 1987, but see Coady, Kluender, & Evans, 2005). Much of this work has addressed these questions by examining phoneme categorization in a 2AFC task. In this task, participants hear a word (or phoneme sequence; e.g. *ba* or *pa*) that comes from a continuum ranging in small steps from one endpoint to another and their task is to assign the stimulus to one of two categories. Listeners typically show a sigmoidal response function transitioning sharply from one phoneme to the other somewhere in the middle of the continuum.

Critically, the steepness of the slope of the categorization function is used as a measure of the discreteness of the categories.

Studies have used this measure to show that a variety of impaired populations generally show shallower transitions between categories (but see Blomert & Mitterer, 2004; Coady, Kluender, & Evans, 2005; McMurray et al., 2014). For example, Werker and Tees (1987) found that children with reading difficulties had shallower slopes on a /b/-to-/d/ continuum relative to typical children (see also Godfrey & Syrdal-Lasky, 1981; Serniclaes & Sprenger-Charolles, 2001). Joanisse, Manis, Keating, and Seidenberg (2000) found a similar pattern for language impaired (LI) children. More recently, López-Zamora et al. (2012) found that shallower slopes in a phoneme identification task predict atypical syllable frequency effects in visual word recognition, suggesting some kind of atypical pattern of sublexical processing. Lastly, Serniclaes, Ventura, Morais, and Kolinski (2005) found that literate adults have sharper identification slopes compared with illiterate adults.

These findings are typically attributed to some form of non-optimal categorical perception, an approach that assumes a sharp, discrete category boundary as the optimal response function. Consequently, if impaired learners encode cues inaccurately (e.g., they hear a VOT of 10 ms occasionally as 5 or 15 ms), then tokens near the boundary are likely to be encoded with cue values on the other side, flattening the function. This assumes a highly categorical response as the goal, which is, however, corrupted by internal noise. That is, impaired listeners may be equally categorical (in terms of mapping cue values to phoneme categories) as non-impaired listeners, but show noisier auditory encoding. This is a likely possibility in the case of individuals with certain

communication disorders, such as hearing loss (Moberly, Lowenstein, & Nittrouer; Winn & Litovsky, 2015). However, in cases like dyslexia or specific language impairment, this is less clear. An alternative explanation is that children with dyslexia have *heightened* within-category discrimination (Werker & Tees, 1987). This would mean that dyslexia is linked to a difficulty in discarding acoustic details that are linguistically irrelevant (Bogliotti, Serniclaes, Messaoud-Galusi, & Sprenger-Charolles, 2008; Serniclaes et al., 2004), which is a failure of a functional goal of categorization. In either case, the assumption is that sharp, discrete categorization, and a reduction of within-category sensitivity are to be desired, and a failure in any aspect of this process drives the shallower response slope.

However, a recent study by Messaoud-Galusi, Hazan, and Rosen (2011) challenges this foundational assumption. They tested typically developing and children with dyslexia using a large battery of phoneme discrimination tasks along with the standard 2AFC phoneme identification tasks. They found that, if anything, typically developing children were better at within-category discrimination (less categorical), and they only found significant group differences in between-category discrimination in some of the discrimination measures. Perhaps more importantly, discrimination measures (both within- and between-category) showed very little correlation with each other, undermining the validity of this important source of evidence for categorical perception. According to the authors, this suggests that any significant between-group differences in phoneme categorization are more probably due to task-related factors and not because of differences in sensitivity to phonological/allophonic contrasts.

Few studies have examined individual differences in speech perception from the perspective that more gradient responding may be beneficial (though see McMurray et al., 2014). An exception to this is recent work by Clayards et al., (2008). They manipulated the variability of VOTs during a brief training session and found that when VOTs were more variable, listeners' response patterns followed shallower (i.e. more gradient) 2AFC slopes (this was also observed in an eye-tracking paradigm similar to that reported by McMurray et al., 2002). This raises the possibility that a shallower identification slope may reflect a different (and Clayards et al. argue more useful) way of mapping cue values onto phoneme categories. Of course, such an explanation need not conflict with an account based on differences in perceptual sensitivity (or internal noise), as the former reflects the mapping from cues to categories, while the latter reflects how those cues are encoded.

In sum, there appear to be two quite different views on this issue of whether gradiency in speech perception is useful. On one side, there is the classical view that favors steeper, more categorical phoneme labeling that comes mainly from studies on atypical populations. The other, more recent view, largely supported by basic research on speech perception, argues that shallower, more gradient response functions are not only the typical pattern of phoneme categorization, but may also be favorable in some aspects.

To some extent both sides may hold some truth; shallower functions may derive from both noisier response pattern *and* a more graded mapping of cues to speech categories. However, what is clear is that there are group differences in phoneme categorization that relate to differences in language processing. Perhaps more importantly, our review thus far suggests that measures like the 2AFC phoneme

identification task may not do a good job measuring these differences, because it is difficult to distinguish different sources of noise from more gradient categorization.

1.7 Towards a new measure of phoneme perception gradiency

The foregoing review reveals a fundamental limitation of the 2AFC task as a way of measuring gradiency of phoneme categorization; a shallower slope in a 2AFC task is ambiguous. That is, the systematicity with which a listener identifies acoustic cues and maps them to phoneme categories (i.e. noise) may be *orthogonal* to the degree to which they are sensitive to and maintain within-category information (see also López-Zamora et al., 2012; Messaoud-Galusi et al., 2011). This is partly because the 2AFC task only provides binary responses. Therefore, when a listener reports a stimulus 30% of the time as /b/ and 70% as /p/, it is unclear whether they do so because they discretely thought the stimulus *was* a /b/ 30% of the time, or because they genuinely thought it had some likelihood of being either or both and the responses reflect the probability distributions of cues-to-categories mappings.

Instead, a continuous measure (e.g., of the category goodness of /ba/ versus /pa/) may offer a more precise way to address this problem. In the example above, if listeners hear the stimulus categorically as /b/ 30% of the time (and as /p/ 70% of the time) the trial-by-trial data should reflect a fully /b/-like response (or /p/-like response) with a different likelihood of choosing one or the other. In contrast, if listeners' representations actually reflect the partial ambiguity, they should respond somewhere in between, with variance clustered around the mean rating. As pointed out by Massaro and Cohen (1983),

"relative to discrete judgement, continuous judgments may provide a more direct measure of the listener's perceptual experience".

One such task is a visual analogue scaling (VAS) task. In this task, participants hear a stimulus and select a point on a line to indicate how close the auditory stimulus was to the word shown on each side (Figure 1.1; see Massaro & Cohen, 1983, for an analogous task in discrimination). This sort of continuous response (instead of a forced binary choice) permits a much more direct measure of gradiency or discreetness.

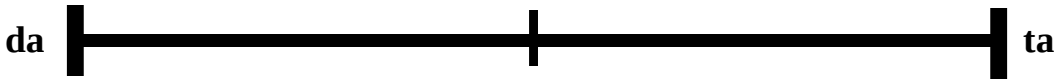


Figure 1.1 Visual analogue scaling task used by Kong & Edwards (2011; submitted)

For example, if we assume a relatively step-like categorization function, plus noise in the cue-to-category mapping, one might expect listeners' responses to cluster close to the extremes of the scaling measure, though for stimuli that are near the boundary, participants might choose the *wrong* extreme because noise would cause them to misclassify the stimulus (e.g., they may choose the left end of the continuum for some rather ambiguous /p/-initial stimuli), even as the categorization process is discrete. On the other hand, if listeners respond more gradiently, one should observe a more *linear* relationship between the acoustic cue value (e.g. the VOT) and the VAS response, with participants using the whole range of the line and variance across trials clustered around the line. In contrast to the VAS task, under either model, a 2AFC would give us an identical response function: a shallower slope.

These kinds of visual analogue scales have been used for quite some time and generally support a gradient perspective on speech. Massaro and Cohen (1983a), for example, used a VAS task to evaluate categorical versus continuous models of perception of acoustic features and found that the distributions of rating responses were better described by the latter type of model. Similarly, a large number of studies by Miller and colleagues (e.g., Allen & Miller, 1999; Miller & Volaitis, 1989) used a VAS goodness scale task (e.g. asking the participant "*How good of a /p/ was this?*") to characterize the graded prototype structure of phonetic categories. However, none of these lines of work examined individual differences, nor related such measures to variation in 2AFC categorization.

Recent work by Kong and Edwards (2011; submitted), building on related work by Schellinger, Edwards, Munson, and Beckman (2008) and Urberg-Carlson, Kaiser, and Munson (2008), offers some evidence for such differences. They tested adults on a /da/-/ta/ continuum, asking them to rate each token on a continuous scale. Participants varied substantially in the pattern of their ratings; some exhibited a more categorical pattern, preferring the endpoints of the line, while others were more gradient, using the entire range of available responses. In addition, listeners who responded gradiently showed a stronger reliance on a secondary acoustic cue in a separate categorization task and this pattern of results was consistent across two separate testing sessions. Lastly, Kong and Edwards observed a correlation between gradiency and cognitive flexibility (assessed by the switch version of the Trail Making task), which suggests there may be a link between speech perception and executive function processes. These findings speak to the potential

strengths of an individual differences approach for addressing these fundamental questions using continuous (VAS) ratings.

The results of Kong and Edwards demonstrate the reliability of VAS measures, and provide preliminary support for a link between gradiency in the VAS task to the use of secondary cues (a key prediction of accounts suggesting gradiency could be functionally beneficial to speech perception). In this way, they provide a methodological basis that we can use to study typical and atypical speech perception. However, certain methodological refinements and experimental extensions are necessary to fully address the key questions we ask here. We outline these issues and present the reasoning behind them in the next paragraphs.

First, to assess secondary cue use, Kong and Edwards used the anticipatory eye movement (AEM) task (McMurray & Aslin, 2004). This is a fairly non-traditional measure of phoneme categorization processes which makes it difficult to evaluate their results in relation to findings from other studies using more traditional measures of phoneme categorization (e.g., 2AFC tasks, such as those used in the SLI studies). It is, therefore, unclear how the same individual may perform the more traditional 2AFC task versus a task like the VAS, and the differences between the two patterns of performance would inform our understanding of the speech perception processes these two tasks tap into.

The previous point is particularly important given the discrepancy between studies of language disorders that have found shallower 2AFC slopes in SLI and dyslexia (e.g., Werker & Tees, 1987), and the newer view from basic research showing that gradiency is not only the typical pattern among non-impaired listeners, but may be highly

adaptive (Clayards et al., 2008). The VAS task may offer unique insight into the relationship between the 2AFC task and these contrasting theoretical notions of gradiency, which is crucial for resolving this mismatch.

Another important aspect of the Kong and Edwards study is their statistical measure of gradiency (from the VAS data), which captured the overall distributions of individuals' ratings (e.g., how often participants use the VAS endpoints) independently of the stimulus characteristics. This is important to point out for two reasons. First, it leaves open the possibility that such individual differences may also be sensitive to other aspects of speech perception (e.g. multiple cue integration or noise). For example, a flatter, more uniform distribution could be obtained if listeners matched their VAS ratings to the VOT, or if they showed a large effect of F_0 (which would spread out their responses), or even if they simply guessed. In contrast, analysis of a VAS-based measure that is sensitive to the stimulus characteristics would allow us to estimate categorization gradiency independently of other potentially confounding facets of speech perception. Second, by developing a stimulus-dependent measure we can also compute an estimate of trial-by-trial noise in the encoding of the stimuli independently of gradiency, thus addressing a main critique of the 2AFC task.

Finally, executive function is a complex and multi-faceted construct. Kong and Edwards used two measures (a Trail Making task and a color-word Stroop task), which possibly load on different aspects of executive function, but only found a correlation between the former one and VAS gradiency (though this should be qualified by their moderate sample size of 30).

1.8 Remaining questions and present study

The literature review reveals limitations in our understanding of speech perception. Despite the evidence that typical listeners maintain within-category information, there are several still unresolved questions:

- What are the perceptual bases and cognitive mechanisms that underlie phoneme categorization gradiency?
- How is categorization gradiency linked to other aspects of speech perception (e.g. multiple cue integration)?
- Can a gradient approach in phoneme categorization be beneficial or detrimental for speech perception in different situations?
- Is speech gradiency a stable characteristic of listeners' perceptual systems, and to what degree can it be modified via experience?

The present study sought to address these questions within an individual differences approach. Crucially, we set out to get at these issues in a comprehensive way by: (1) developing and testing a theoretically-grounded measure of phoneme categorization gradiency, (2) exploring a wide variety of processes (both within and outside the language system) that may be linked to gradiency, (3) assessing the impact of gradiency on spoken language comprehension, and (4) examining whether the way in which a listener categorizes speech sounds can be adjusted via experience.

In Chapter 2, we describe our methods and present a novel VAS-based paradigm for measuring phoneme categorization gradiency. In contrast to previously used VAS

paradigms, we complemented the typical behavioral measure with a statistical approach that was specifically designed to dissociate between gradiency and multiple cue use. This allowed us for the first time to extract an assessment of phoneme categorization gradiency independently of multiple cue integration. In addition, we report results from Monte Carlo simulations that display the ability of our measure to accurately reflect the underlying structure of the data and, thus, speak to its content validity.

In Chapter 3, we report a few preliminary findings using this new approach, while replicating the individual differences in the use of the VAS task reported by Kong and Edwards (submitted). In addition, we assess the role of stimuli characteristic and evaluate the degree to which the VAS task is sensitive to them, in comparison to a more traditional speech perception measure like the 2AFC task.

Next, we move on to explore how gradiency is linked to other aspects of speech perception like internal noise and multiple cue integration, using a variety of different speech cue combinations, in Chapters 4 through 6.

To investigate possible sources of gradiency, we assess the role of broader cognitive processes like different aspects of executive function (see Chapters 4 and 5), as well as various aspects of language processing, such as inter-lexical inhibition (see Chapter 5). Crucially, we also examine the possibility that individual differences in gradiency are due to differences in the early perceptual encoding of acoustic cues (see Chapter 5).

In order to explore different ways in which higher or lower levels of sensitivity to within-category information may affect the efficiency of speech processing, we use two kinds of tests: 1) listeners' ability to perceive speech in noise (Chapters 4 and 6) and 2)

listeners' ability to deal with ambiguities and recover form erroneous interpretations when needed (see Chapter 6).

Then, in Chapter 7, we report the results of a preliminary test of the hypothesis that the way in which listeners use within-category information can be adjusted with experience using an experimental between-group training manipulation.

Lastly, in Chapter 8, we discuss the findings cumulatively across studies and tasks, draw parallels, and point out systematic patterns of results that speak to the key questions of interest.

CHAPTER 2: GENERAL METHODS

Chapter 2 describes in detail our methodological approach for measuring phoneme categorization gradiency (using the VAS task) and for quantifying the use of secondary cues (using the 2AFC task).

2.1 Measuring phoneme categorization gradiency via the VAS task

To measure individual differences in the gradiency of phoneme categorization, we used the visual analogue scaling (VAS) task (with different stimuli for each Experiment). In this task, participants are presented with auditory stimuli varying along two dimensions (e.g. VOT and F_0) and are asked to indicate what they heard by choosing a point on a line. In this way, instead of being forced to choose between two options, participants are given the opportunity to give responses that match more closely the continuity of the stimuli.

2.1.1 Basic stimulus manipulation in the VAS task

Our VAS task requires two-dimensional continua (e.g., VOT × F_0) for the construction of which we used the Praat software (Boersma & Weenink, 2016, 2012 [version 5.3.23]). For the voicing manipulation, stimuli were constructed from natural speech using the progressive cross-splicing method described by Andruski, Blumstein, and Burton (1994) and McMurray, Aslin, Tanenhaus, Spivey, and Subik (2008). Progressively longer portions of the onset of a voiced sound (e.g., /b/) were replaced with analogous amounts taken from the aspirated period of the corresponding voiceless sound (e.g., /p/). This creates a VOT continuum in which acoustic cues other than voicing are

also present (e.g., aspiration, pitch, and first formant frequency). Then, for each VOT step, the pitch contour was extracted from the recording and was modified using the pitch-synchronous overlap-add (PSOLA) algorithm in Praat. For Experiment 1, pitch level was kept steady over the first two pitch periods of the vowel and fell (or rose) linearly until returning to the original contour at the 80-ms point in the vowel. For Experiments 2-4, pitch level varied throughout the entire duration of the stimuli (see *Methods* section of individual experiments for further details on stimulus manipulation).

2.1.2 Basic VAS task procedure

On each trial, participants saw a line with a printed word at each end (e.g. *bull* on the one end and *pull* on the other, see Figure 1.1). Across participants and experiments, voiced-initial stimuli were always presented on the left side. For Experiment 1, in the middle of the line there was a rectangular bar and participants were instructed to use the computer mouse to drag that bar onto a point on the line that indicated where they think the sound falls in between the two words. In Experiments 2-4, the task was the same, but the rectangular bar only appeared after the participant clicked on the line.

2.1.3 Statistically dissociating gradiency from secondary cue use: The rotated logistic function

An obvious way of extracting a measure of gradiency would be to fit a sigmoid or a logistic function to the VAS data of each participant and use the steepness of the slope as a measure of gradiency. However, since stimuli also varied along a secondary dimension, this method is problematic. For example, if each listener has a perfectly

discrete boundary in VOT space, but the location of this boundary varies with F_0, then the average boundary (across F_0s) would look quite gradient. Instead, what is needed is an estimate of the slope in two dimensions.

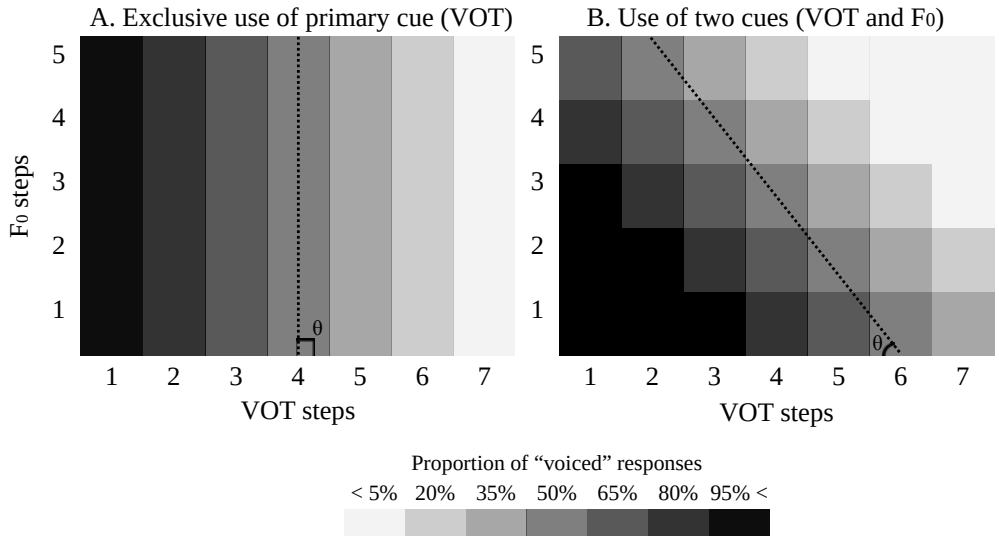


Figure 2.1 Hypothetical response patterns based on unidimensional (left) and bi-dimensional (right) category boundaries

To accomplish this, we developed a new function (Eq.1), which we term the *rotated logistic*. This assumes a diagonal boundary in two-dimensional space that can be described as a line with some cross-over point (along the primary cue) and an angle, θ (Figure 2.1). Here, a θ of 90° would indicate that the listener only used the primary cue (see Figure 2.1.A), while a θ of 45° would indicate relatively equal use of both cues (see Figure 2.1.B). Once the angle of the boundary is identified, we rotate the coordinate space to be orthogonal to this boundary and estimate the slope of the response function perpendicular to this diagonal boundary. This allows us to model the gradiency of the function with a single parameter that indicates that derivative of the function orthogonal

to the (diagonal) boundary; the steeper the slope the more categorical the response pattern independently of cue use[3].

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

Here, b_1 is the lower asymptote, b_2 is the upper asymptote, and s is the slope (much like the standard four parameter logistic). The new parameters are: θ, which is the angle of the boundary (in radians), and x_0, *which* is the x-intercept of the diagonal boundary. The two independent variables or cues are represented by *VOT* and F_0.

The function $\upsilon(\theta)$ (in the denominator) simply switches the direction of the slope if θ is less than 90° to keep the function continuous (see Eq.2).

$$\upsilon(\theta) = \begin{cases} 1 & \text{if } \theta <= (\pi/2) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

For each participant, we calculated the average of their responses for each of the stimuli they heard during the VAS task. Next, the rotated logistic equation was used to fit to each participant's averaged VAS data using a constrained gradient descent method implemented in Matlab (using the *fmincon*() function) that minimized the least squared

[3] For a test of the independence between the slope (*s*) and the angle theta (θ), see Section 2.1.4 Testing the independence of VAS slope and theta angle with Monte Carlo simulations.

error. The data for each participant and each place of articulation were fitted separately. Starting parameters for the fits were estimated directly from the data. The starting values for the upper and lower asymptotes (b_1, b_2) were based on the average values (across repetitions) of individual participants' minimum and maximum VAS rating values, and b_1 and b_2 were constrained during fitting to lie between 0 and 100. Starting value for the crossover was the middle VOT step and was constrained so that it could only take values between the minimum and maximum VOT step (with the exception of Experiment 4 for which this value was unconstrained). For the theta (θ) angle, we used a starting value of 80° and this parameter could only take values between .01° and 179.9°. Finally, the starting value for slope (s) was 0 and permitted values were between -1000 and 1000[4]. We then examined the estimated parameters as measures of each participant's degree of gradiency (s) and multiple cue use (θ). In most cases the other parameters were not of interest and were not analyzed.

2.1.4 Validating the VAS slope and testing its independence from the theta angle with Monte Carlo simulations

We conducted a Monte Carlo analysis to evaluate our curvefitting procedure and the rotated logistic function. This had two goals. First, we wanted to determine the ability of this procedure to estimate the true parameters that generated a dataset (i.e. the validity of this statistical method). Second, we investigated whether parameter estimates were biased, or non-independent from each other. In particular, we needed to verify that the

[4] In rare occasions the fitter would output values at the maximum/minimum of the permitted values. This was problematic because it meant that other values may have been affected in the fitter's effort to reach a better fit. For this reason, such fits were excluded from the main analyses.

estimated slope (*s*) and theta (θ) values were not correlated with each other due to our curve-fitting procedure.

On the first step of this Monte Carlo procedure we generated a set of underlying parameters for 1000 simulated subjects. To ensure that the simulated data roughly reflected the data collected from our participants, we based our simulated data on the mean values and standard deviations of the parameters that were estimated from the behavioral data collected in Experiment 1 (see Table 2.1).

For each simulated subject, we started off with the given mean value (e.g. 5.4 for the crossover; see Table 2.1) and, using the *randn*() Matlab function, we added to that value a random number drawn from a standard normal distribution with a mean equal to the source mean and a standard deviation equal to the standard deviation for that parameter. For any given participant, each parameter was estimated independently, so there was no underlying relationship between any of the five parameters.

Next, we used the rotated logistic equation to generate simulated responses for each of the 1000 simulated subjects. Following our experimental procedure, we generated 3 responses for each of the 35 (7 VOT × 5 F_0) cells. Reponses came from a random normal function with a mean given by the rotated logistic for that subject, and a specified variance. Since the SD of the responses was partially dependent on the VAS rating value (with greater standard deviation for middle ratings, and much smaller SDs near the ends), we estimated the function linking the two and used the resulting equation (see Eq.3) to generate the SD for the random normal function given the mean VAS rating specified by the rotated logistic.

$$St.Dev. = -.00091 \times Rating^2 + .9064 \times Rating + .641 \quad (3)$$

Table 2.1 Parameter values in Monte Carlo simulations

	Low asymptote (b_1)		High asymptote (b_2)		Crossover (x_0)		Theta angle (θ) [sqrt]		Slope (s) [log]	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Source parameters	13.4	8.0	87.6	8.0	5.4	1.0	7.8	.7	-1.7	.3
Simulated parameters	14.0	7.5	86.7	7.1	5.3	.9	7.8	.2	-1.7	.3
Estimated parameters	16.0	8.9	94.9	7.8	5.4	.9	7.8	.2	-1.8	.4

On the third step, we averaged across the three responses for each subject and each VOT/F_0 cell and fitted the simulated response data using the same curvefitting procedure as the one used to fit the behavioral data. Then we calculated (1) the correlations between the simulated parameters for that subject (see Table 2.1, second row) and the estimated parameters from the curvefitting procedure (see Table 2.1, third row); and (2) the correlation between the slope (s) and the theta (θ) values of the estimated parameters.

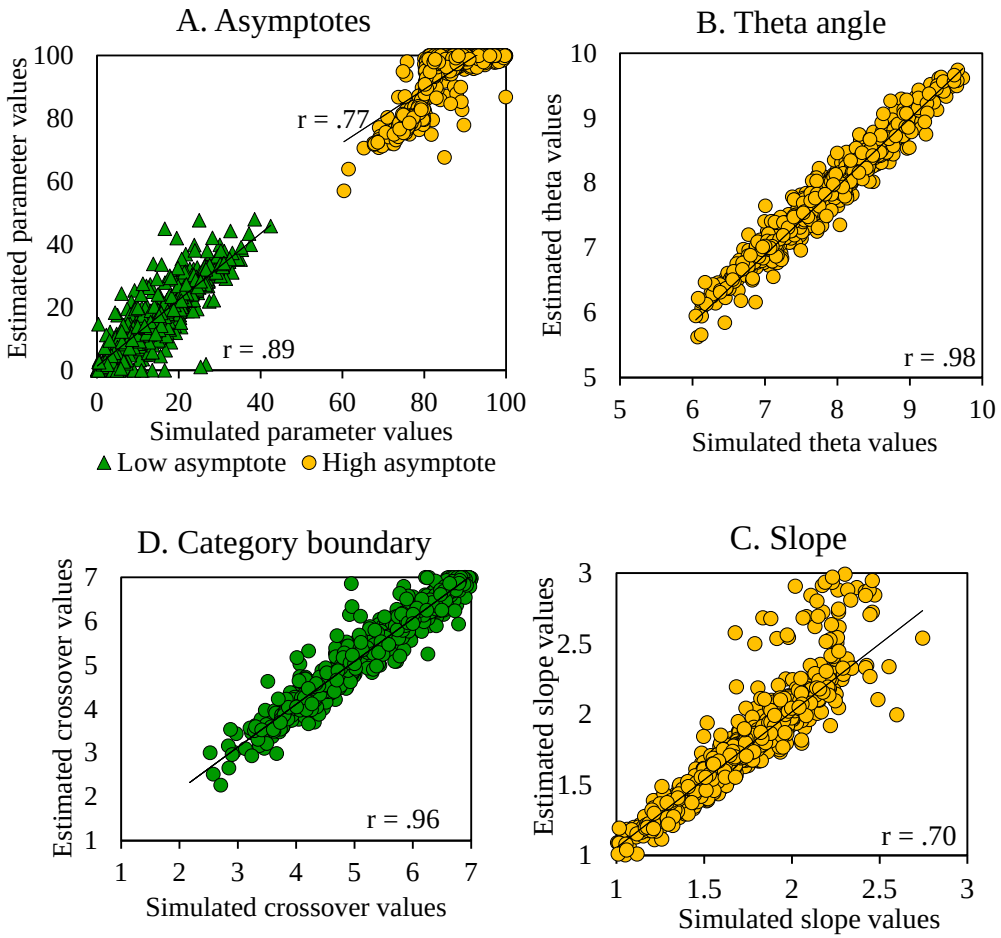


Figure 2.2. Simulated parameter values (x axes) by estimated parameter values (y axes)

Figure 2.2 shows the correlations between the estimated and underlying parameter values. For all five parameters, underlying values were very close to the simulated data. All correlations were above .7 and some were extremely high (e.g., r = .98! for θ). This validates the accuracy of the curve-fitting. It also suggests that three repetitions per VOT/F_0 step are sufficient to obtain good parameter estimates with this procedure.

We next examined the relationship between the estimated values of slope (*s*) and theta angle (θ). Recall that the underlying parameters for each subject were generated independently, such that any correlation among the estimated parameters would have been imposed by this procedure. However, as Figure 2.3 shows, no such correlation was

observed ($R^2 < .001$). Consequently, any correlation found in the empirical data was not due to a bias of the curve-fitter, but due to a correlation in underlying properties of the individuals being tested.

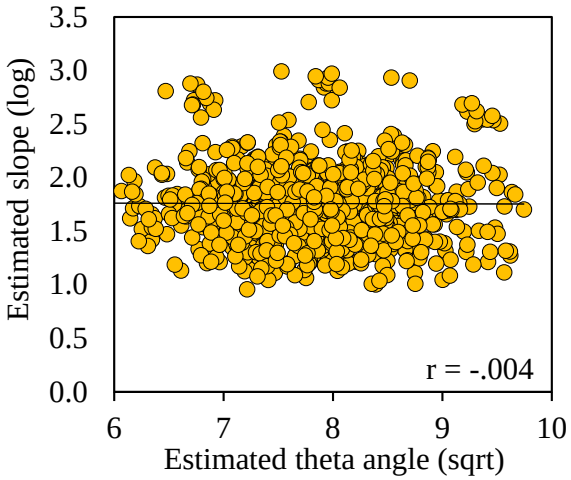


Figure 2.3 Correlation between estimated theta angle (sqrt) and slope (log)

2.2 Dissociating gradiency in phoneme categorization from gradient response bias: The visual VAS task

In Experiments 2 and 3, we evaluated the degree to which participants were inclined to use the whole line versus the endpoints using a visual version of the VAS task. This was important for determining whether individual differences in gradiency are due to differences in how people approach the VAS task, differences in general cognitive factors (e.g., an overall more gradient approach to categorization), or whether they are due to differences specifically in speech perception.

To assess this, we used a task that was similar to the phoneme VAS task described above (see Section 2.1), but instead of listening to words, participants saw pictures of objects varying between a picture of an apple and that of a pear and were asked to

evaluate the degree to which each picture was visually closer to an apple versus a pear. This allowed us to extract a baseline of non-speech-related categorization gradiency (i.e. visual VAS slope). Then, we partialed out the phoneme VAS slope variance explained by the visual VAS slope to compute the *residualized VAS slope*, which was used in the main analyses in Experiments 2 and 3 along with the VAS slope.

2.2.1 Visual VAS task design and materials

For the endpoints of the visual VAS task, we used two pictures downloaded from a commercial clipart database, which we edited in order to intensify prototypical characteristics. We subsequently morphed these pictures using the Fantamorph (ver. 5) software to create 35 stimuli varying orthogonally in shape and color. Following the 7-by-5 structure of the auditory stimuli, we had 7 shape steps and 5 color steps (see Figure 2.4). Each picture was presented 5 times, resulting in 175 trials.

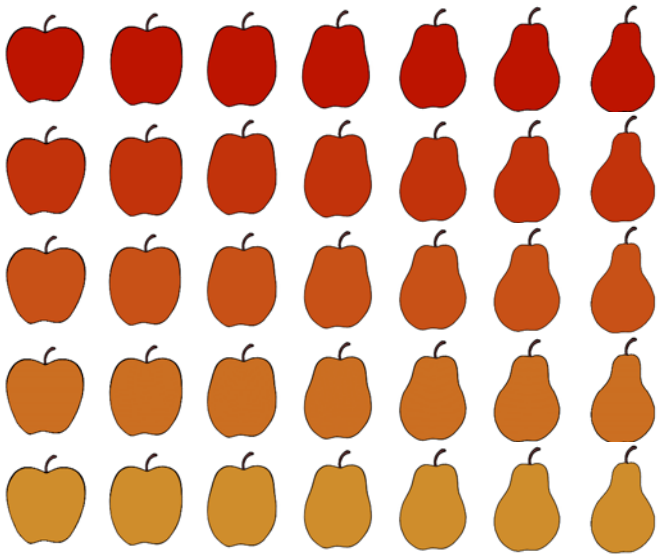


Figure 2.4 Picture stimuli used in the visual VAS task in Experiments 2 and 3

2.2.2 Visual VAS task procedure

Similarly to the phoneme VAS task, in each trial, participants saw the morphed picture at the center of the screen as well as a line the endpoints of which were named *apple* and *pear*. The instructions to the participants were similar to those presented in the phoneme VAS task: "*Click on the line to indicate where you think what you see falls on the line*". When they clicked on the line the rectangular bar would appear at the point where they clicked and then they could either change their response or press the space bar to verify it.

2.2.3 Extraction of visual gradiency measure

As in the phoneme VAS task, we used the rotated logistic equation to fit individual participants' responses and collect a measure of visual categorization gradiency. As mentioned earlier, shape was used as the primary categorization cue and color as the secondary.

2.3 Measuring secondary cue use via the 2AFC task

For Experiments 1, 3, and 4, we used 2AFC phoneme identification tasks to measure multiple cue integration. The 2AFC task offers a convenient and standard way of assessing multiple cue integration (which we hypothesized to be better in more gradient listeners) as the degree to which the category boundary along the primary cue continuum shifted as a function of secondary cue. Crucially, this measure was extracted from a different task than the one used to measure gradiency (VAS task), thus providing us with an independent measure of secondary cue use.

In addition to its primary purpose, as a standard measure of speech categorization, this task offers a way to evaluate our slope measure, even though, as we described, this is an imperfect measure, because shallower slopes can result from both nosier cue encoding and a more gradient categorization. Thus, examining the relationship between these measures can tell us whether individual differences in the normal range are primarily due to one or the other.

2.3.1 Basic 2AFC task design

For each experiment in which a 2AFC task was used, a subset of the VAS stimuli were used in this task. Specifically, all primary cue steps were presented, but only the two extreme secondary cue values. This was done to simplify our quantification of listeners' use of secondary cues as the difference between boundaries for each secondary cue value.

2.3.2 Basic 2AFC task procedure

On each trial, participants were presented with two squares (one on the left and one on the right side of the screen), each containing one of two printed words (e.g. *bull* in one square and *pull* in the other). Across participants and experiments, the voiced-initial stimuli were always presented in the left square. Participants were prompted to listen carefully to each stimulus and then click in the box that contained the word they thought best matched what they heard.

2.3.3 Pre-processing of 2AFC data

To quantify how much each participant used the secondary cue, we fitted participants' response curves using a four parameter logistic function (see McMurray et al., 2010), which provides us with estimates for minimum and maximum asymptotes, slope, and crossover (see Eq. 4). In this equation, b_1 is the lower asymptote, b_2 is the upper asymptote, s is the slope, and co is the x-intercept.

$$p(resp) = b_1 + \frac{b_2 - b_1}{1 + e^{\left(\frac{-4 \cdot s}{(b_2 - b_1)}(x - co)\right)}} \tag{4}$$

This function was fitted to each participant's responses separately for each secondary cue value (and stimulus type, wherever applicable), thus extracting at least two sets of parameters for each participant (see Figure 2.5).

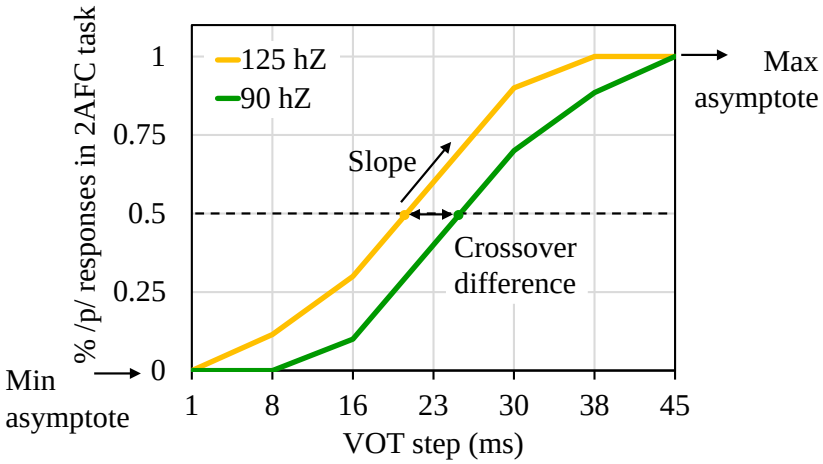


Figure 2.5 Hypothetical response curves in the 2AFC (green: low pitch; yellow: high pitch)

The starting values for the upper and lower asymptotes (b_1, b_2) were the average (across repetitions) of individual participants' minimum and maximum 2AFC rating values. We used 0 and 1 as minimum and maximum values for the lower and higher asymptotes respectively. Starting value for the crossover was the middle VOT step (3). The crossover value was not constrained. The starting value for slope (s) was calculated based on the correlation coefficient between the VOT values and the participant's responses and permitted values for slope were between this starting slope and 20. Curves were fitted using a constrained gradient descent method implemented with *fmincon*() in Matlab (similar to that used for the VAS task).

2.4 Summary of methods

In Chapter 2, we presented our methodological tools for assessing basic aspects of speech perception such as phoneme categorization gradiency and multiple cue integration. Crucially, our approach takes advantage of the VAS task (which allows for continuous responses), which we paired with a novel way of dissociating phoneme categorization gradiency from multiple cue integration using the *rotated logistic* equation (see Eq.1).

In order to evaluate our measure, we ran Monte Carlo simulations, which demonstrated that (1) the curve-fitting procedure was unbiased and generated truly independent fits of gradiency and multiple cue integration, and (2) the fits accurately represented the underlying structure of the data to-be-fit even with as few as three repetitions per stimulus step.

This novel paradigm was used in Experiments 1-4 presented next.

CHAPTER 3: EFFECTS OF STIMULI CHARACTERISTICS ON GRADIENCY AND SECONDARY CUE USE (EXPERIMENT 1A)

Experiment 1 aimed at (1) providing some preliminary results using our novel measure of phoneme categorization gradiency (see Experiment 1a), but also (2) addressing some of our theoretical questions about the role of speech gradiency presented in Chapter 1 (see Experiment 1b). Even though the data for Experiments 1a and 1b were collected via the same tasks and from the same participants, the motivation is different between these two sub-experiments, and for this reason they are presented in separate chapters.

3.1 Introduction

Experiment 1a examined phoneme categorization gradiency using both word and nonword continua, as well as both labial- and alveolar-initial stimuli. This allowed us to assess any possible effects of lexical status and place of articulation respectively.

While these manipulations were somewhat exploratory, prior eye-tracking results suggest that listeners may be more sensitive to subphonemic detail with lexical tasks rather than phoneme decision tasks (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008). This raises the possibility that the individual differences reported by Kong and Edwards are only seen with nonwords, while most listeners show a gradient response pattern with words. This was particularly important for us to test, because one of our goals was to examine possible correlations between our VAS-based measure of gradiency and other measures extracted from tasks that use real words. Therefore, we decided to

start by carefully assessing any potential effects of stimulus-specific characteristics on gradiency.

3.2 Methods

3.2.1 Participants

Participants were 131 adult monolingual speakers of American English. All participants completed a hearing screening at four octave-spaced audiometric test frequencies for each ear; one participant was excluded on this basis because of thresholds greater than 25 dB HL. This left 130 participants for analysis. Participants received course credit for participation in the study, and underwent informed consent in accord with University of Iowa IRB policies.

Technical problems with several of the tasks led to their results not being available for one or more participants. Consequently, across Experiments 1a and 1b, between two and 11 participants were excluded from the analyses of the specific tasks for which there were missing data.

3.2.2 Design and tasks

A hearing screening was performed at the beginning of the session and lasted approximately 3 mins. Immediately after that, participants performed a series tasks including a VAS and a 2AFC task. To explore any stimulus-driven effects on gradiency, we included voicing continua in both labials and alveolars (within subject), in words, nonwords, and phonotactically impermissible nonwords that featured lax vowels with no word-final consonant (with consonant place of articulation and syllable type being

between subjects). This was assessed for stimuli that varied on seven steps of VOT (the primary cue for word-initial voicing) and five steps of F_0 (a secondary cue).

The 2AFC task was conducted on continua that varied on seven steps of VOT and only two steps of F_0; this allowed us to extract an independent estimate of secondary cue use measured as the difference in the category boundary between the two VOT continua.

3.2.3 VAS task

3.2.3.1 VAS task design and materials. To measure individual differences in the gradiency of phoneme categorization, we used the VAS task with several different continua. Specifically, to test whether the individual differences in categorization reported by Kong and Edwards (2011), can also be observed with real words, we used three types of stimuli (*stimulus-types*): 1) CVC real words, henceforth RW; 2) CVC nonwords, henceforth NW; and 3) phonotactically impermissible nonword CVs[5], which violated constraints of English lax vowels being permitted only in closed syllables (i.e. in syllables that end with a consonant). Each participant was only tested on one stimulus-type. We also used two places of articulation (henceforth PoA): one stimulus set with labial-initial phoneme (e.g. *bull-pull*) and one with alveolar-initial phoneme (e.g. *den-ten*; see Table 3.1). Each participant was tested on both labials and alveolars (of the same stimulus type).

[5] Similar to those used by Kong and Edwards (2011, submitted)

Table 3.1 Stimuli used in the VAS and the 2AFC tasks in Experiment 1

	Stimulus type		
	Real word	Nonword	CV
Labial	bull – pull	buv – puv	buh – puh
Alveolar	den – ten	dev – tev	deh – teh

All participants performed the VAS task first. For each of the six continua, we created a two-dimensional continuum by orthogonally manipulating VOT (seven VOT steps; 1 to 45 ms) and F_0 (five F_0 steps; 90 to 125 Hz) using Praat (Boersma & Weenink, 2012 [version 5.3.23]). All stimuli were constructed from natural speech using a modified version of the progressive cross-splicing method described by Andruski et al. (1994) and McMurray et al. (2008). Progressively longer portions of the onset of a voiced sound (/b/ or /d/) were replaced with equivalent amounts taken from the aspirated portion of the corresponding voiceless sound (/p/ or /t/). This creates variations in VOT in which multiple additional cues to voicing are maintained (e.g., pitch, first formant frequency), as well as differences in vowel onset intensity consistent with elision of varying amounts of time from the original onset. Each vowel excised from the original recording was multiplied by a 3 ms onset ramp, and was cross-spliced with the consonant burst/aspiration segment using a symmetrical 2-ms cross-fading envelope, in order to remove any waveform discontinuities at the boundary between aspiration and vocalic segment.

At each step of the VOT continuum, the pitch contour was extracted from each stimulus and modified using the pitch-synchronous overlap-add (PSOLA) algorithm in Praat. Pitch onset varied in five steps spaced equally within a 30 Hz range spanning from

190 Hz to 125 Hz. Pitch level was kept steady over the first two pitch periods of the vowel and fell (or rose) linearly until returning to the original contour at the 80-ms point in the vowel. Following the 80-ms time point, all pitch contours were identical within each continuum.

Each participant was presented with all 35 stimuli from each of the two PoA series with three repetitions of each stimulus resulting in 210 trials (7 VOTs × 5 F_0's × 2 PoA × 3 repetitions). Stimulus presentation was blocked by PoA, and the order of PoA was counterbalanced between participants (i.e. some heard labial-initial stimuli first while others heard alveolar-initial stimuli first).

3.2.3.2 VAS task procedure. On each trial, participants saw a line with a printed word at each end (e.g. *bull* on the one end and *pull* on the other). Across blocks and participants, voiced-initial stimuli were always presented on the left side. In the middle of the line there was a rectangular bar and participants used the computer mouse to click on that bar and drag it onto a point on the line that indicated where they thought the sound fell in between the two words. At the beginning of the task, the participant performed a few practice trials with the experimenter in the room to ensure the participant understood the task. Unless the participant had clarifying questions, no further instructions were given. The VAS task took approximately 15 mins.

3.2.4 2AFC phoneme identification task

3.2.4.1 2AFC task design and materials. The 2AFC task was always performed after the VAS task because we wanted to minimize any step-like bias possibly induced by the 2AFC on the VAS task.

A subset of the VAS stimuli were used in the 2AFC task. Specifically, all 7 VOT steps were presented, but only the two extreme F_0 values. This was done to simplify our quantification of listeners' use of F_0 as the difference between boundaries for each F_0. This led to 7 VOTs × 2 F_0's × 2 PoA (28 stimuli). Each stimulus was presented 10 times for a total of 280 trials. Similarly to the VAS task, stimuli were presented in two separate blocks, one for each place of articulation, and the order of the blocks was counter-balanced between participants.

3.2.4.2 2AFC task procedure. On each trial, participants were presented with two squares (one on the left and one on the right side of the screen), each containing one of two printed words (e.g. *bull* in one square and *pull* in the other). The voiced-initial stimulus was always presented in the left square. Participants were prompted to listen carefully to each stimulus and then click in the box that contained the word they thought best matched what they heard. At the beginning of the task participants performed a few practice trials. The 2AFC task took approximately 11 mins.

3.3 Results

3.3.1 VAS task results

Participants performed the VAS task as instructed with the exception of three participants, who slid the response bar to random locations on the line and were excluded from analyses. In addition, technical problems led to missing data for 5 participants, leaving 123 participants with valid data for this task.

Participants used both VOT and F_0 to categorize stimuli. As expected, participants rated stimuli with higher VOT (see *VOT step* axis in Figure 3.1) and higher F_0 values (see *F_0 step* axis in Figure 3.1) as more /p/- (or /t/-) like.

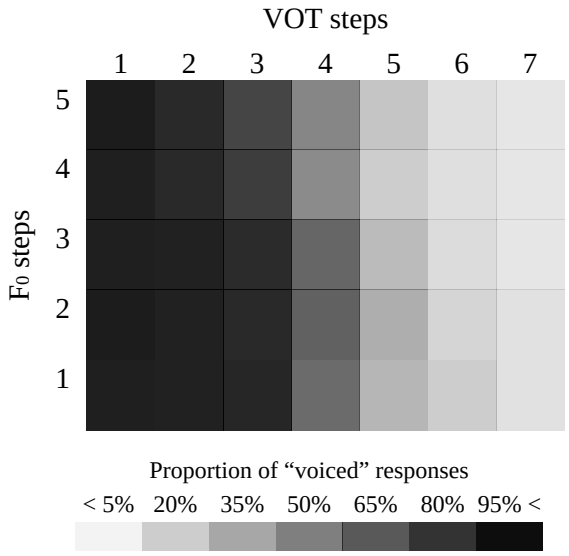


Figure 3.1 VAS responses by VOT and F_0 steps

Replicating Kong and Edwards (2011), we found that participants differed substantially in how they performed the VAS task. This can be clearly seen by computing simple histograms of the points that were used along the visual analogue scale. As Figure 3.2 shows, some participants primarily responded using the endpoints of the VAS line (Figure 3.2.A), suggesting a more categorical mode of responding, while others used the entire range of the response continuum (Figure 3.2.B), suggesting a more gradient pattern of responses.

A. Individuals preferring endpoints B. Individuals using entire range of responses

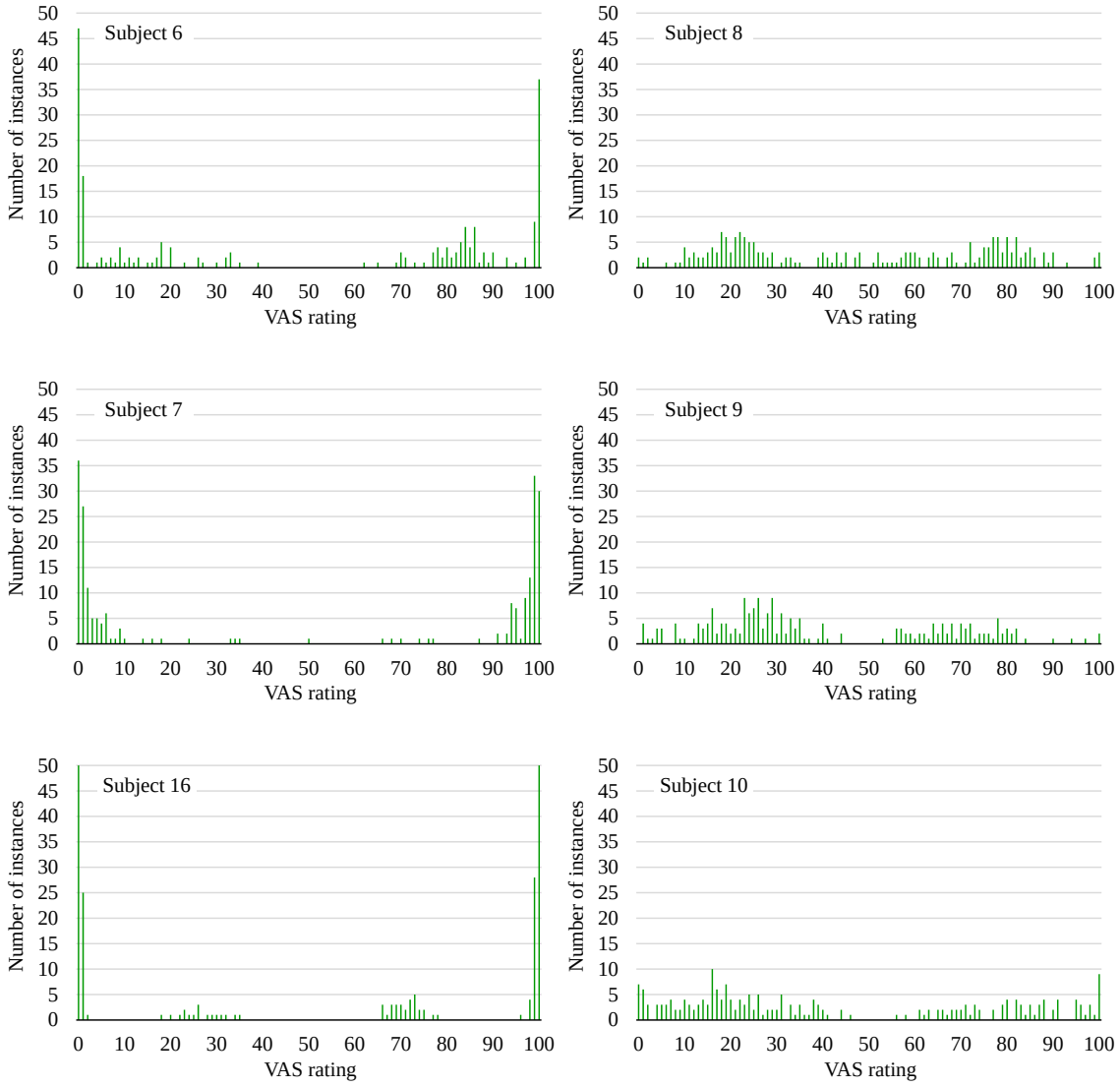


Figure 3.2 Histograms of sample individual VAS responses

Figure 3.2 implies fairly striking individual differences in listeners' categorization pattern, but this approach is insufficient for addressing our primary questions because it ignores the actual stimulus (e.g., the VOT and F_0 values). For example, a participant like subject 10 could show a uniform distribution of VAS scores because they were guessing or because they were closely aligning their VAS ratings with the stimulus.

A better approach is to consider the relationship between stimulus and response. Figure 3.3 shows representative results for two participants plotting the individual (trial by trial) VAS responses as a function of VOT and F_0.

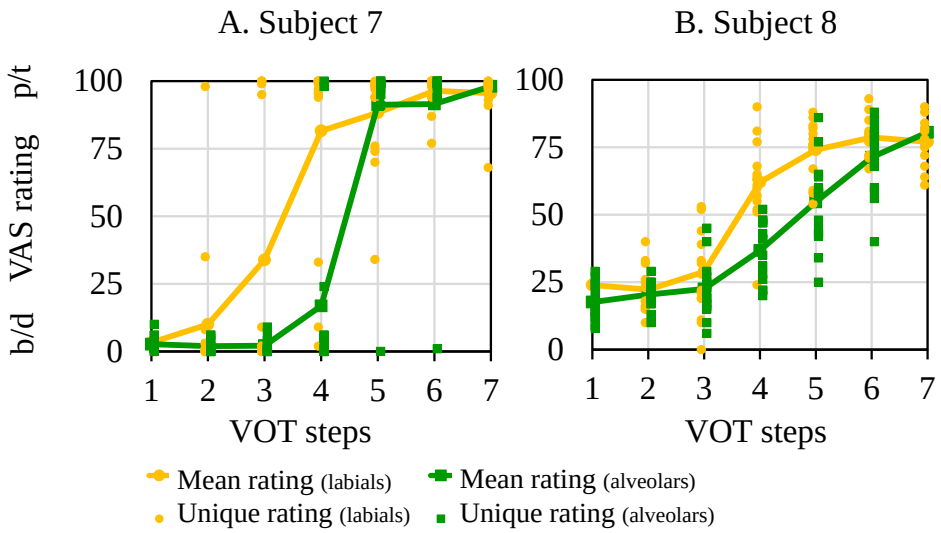


Figure 3.3 Sample VAS ratings per VOT and F_0 value

Figure 3.3. shows two highly dissimilar patterns of responding. Here we see that subject 7 seems to give almost binary responses, reporting VAS scores near 0 or 100 on the VAS scale. What differs as a function of VOT is the likelihood of a close-to-0 or close-to-100 response. In this case, intermediate responses seem to reflect random fluctuations between the two endpoints, rather than being tightly clustered around an intermediate VAS value. Thus, this participant appears to have adopted a categorical approach (with discrete categories around 0 and 100, and the occasional intermediate selection). In contrast, subject 8 gave responses that individually closely follow the cue values for each stimulus, and the variation is tightly clustered around the mean. Thus, this participant's responses seem to reflect the gradient nature of the input. While the VAS task appears to capture these individual differences nicely, such a picture would not be

possible in a 2AFC task. In this task, the average categorization function of both participants would likely look similar, and the variance would be uninformative, since trial by trial data is always a 0 or a 1. Thus, we quantified these differences between participants using the rotated logistic curvefitting approach.

We fitted participants' responses in the VAS task using the rotated logistic function provided in Eq.1. Overall fits were good (R^2 = .96). In addition, we evaluated the quality of our curve-fitted data by visually inspecting individual participants' fits. Figure 3.4 shows the actual and fitted response curves for the two types of stimuli (labial and alveolar) across participants.

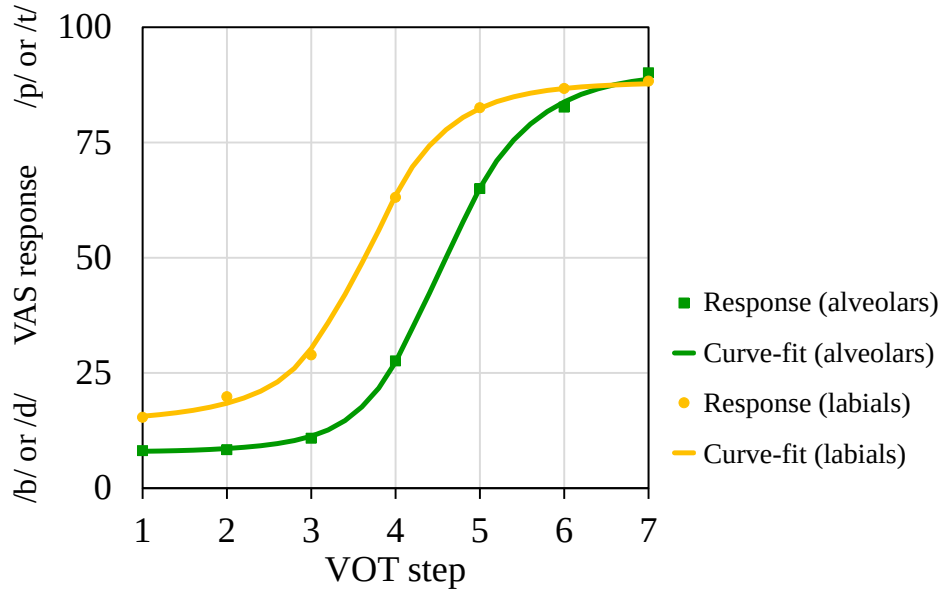


Figure 3.4 Actual and fitted VAS ratings (yellow: labial; green: alveolar)

Because the distribution of raw VAS slopes was substantially positively skewed, we used the log-transformed values in all analyses.

Next, we turned to addressing our first question regarding the effects of stimulus characteristics on phoneme categorization. Effects of place of articulation (PoA) and stimulus type (words, nonwords, CV; henceforth: RW, NW, and CV respectively) were evaluated in a series of two-way analyses of variance with VAS slope, crossover, and theta angle as dependent variables.

Our first analysis examined VAS slope (*s*, see Figure 3.5.A). This found no main effect of PoA, $F < 1$, nor stimulus type, $F < 1$, and the interaction was also not significant, $F(2,120) = 2.60$, $p = .079$.

We next examined crossover (x_0, see Figure 3.5.C). Here the main effect of PoA was significant, $F(1,120) = 129.50$, $p < .001$, with higher crossovers for alveolar-initial stimuli ($M = 4.9$, $SD = .71$) compared to labials ($M = 4.2$, $SD = .67$). Stimulus type was also significant, $F(2,120) = 11.43$, $p < .001$, and interacted with PoA, $F(2,119) = 10.81$, $p < .001$. Post-hoc pairwise (Bonferroni adjusted) comparisons revealed that CV ($M = 4.9$, $SD = .87$) stimuli differed significantly from both NW ($M = 4.4$, $SD = .65$, $p < .001$) and RW stimuli ($M = 4.4$, $SD = .71$, $p < .001$), but RW did not differ from NW stimuli. To investigate the interaction, we split the data by PoA. The effect of stimulus type on crossover was significant for alveolars, $F(2,120) = 21.43$, $p < .001$, but not for labials, $F(2,120) = 2.05$, $p = .133$.

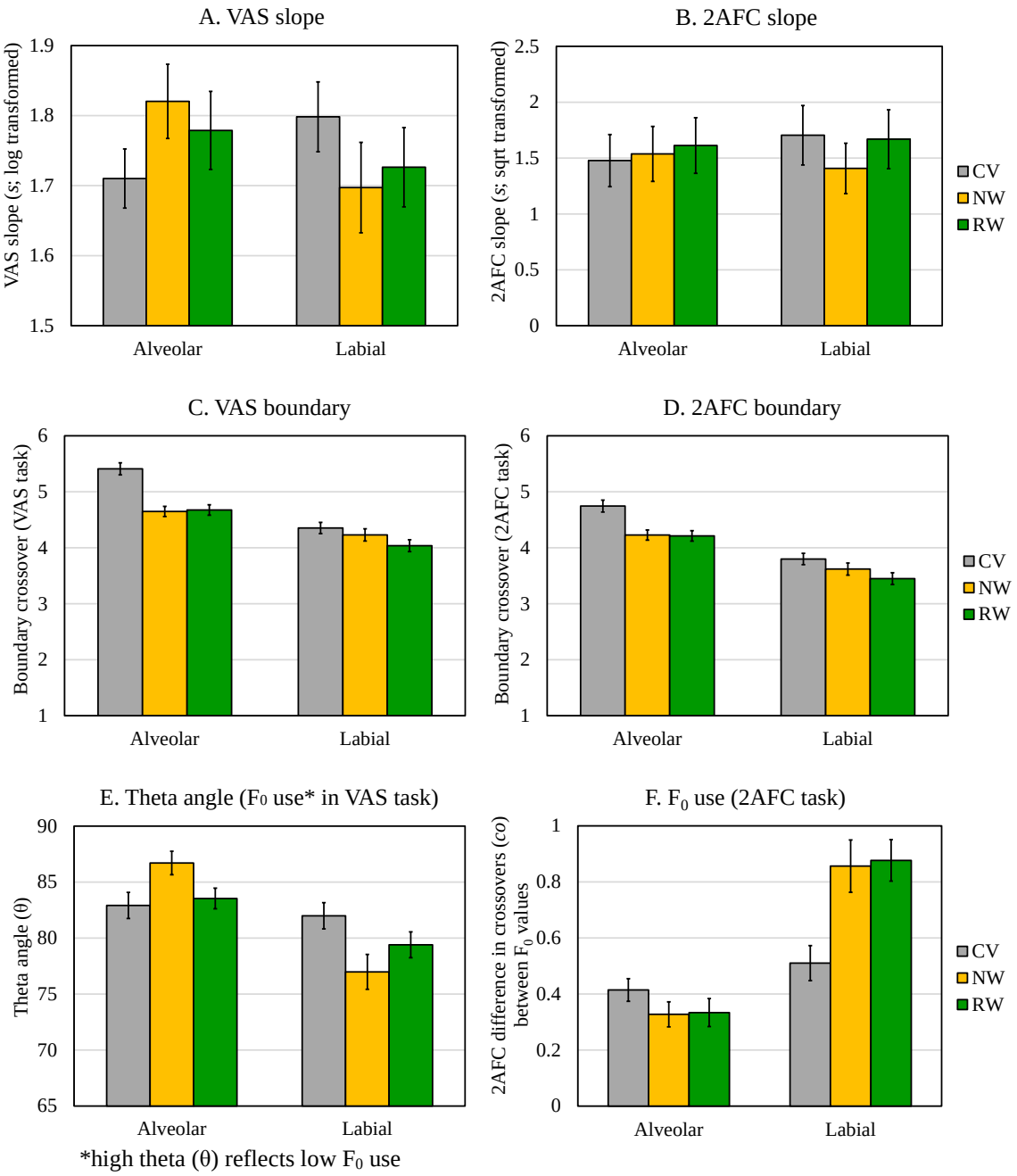


Figure 3.5 Stimulus effects on VAS and 2AFC parameters

Lastly, we examined the degree of multiple cue use (θ, see Figure 3.5.E). Here we found a significant effect of PoA, $F(1,120) = 36.63$, $p < .001$, with higher theta angles (i.e. less secondary cue use) for alveolar-initial stimuli ($M = 84.35°$, $SD = 6.8$), compared to labials ($M = 79.36°$, $SD = 8.5$). Stimulus type was not significant, $F < 1$, but the

interaction term was, $F(2,120) = 11.41$, $p < .001$. To investigate the interaction, we split the data by PoA. Stimulus type was significant for both PoA, however, post-hoc pairwise (Bonferroni adjusted) comparisons revealed that for alveolar-initial stimuli, theta angle was significantly lower in CVs (M = 82.91°, SD = 7.4) compared to NWs (M = 86.70°, SD = 6.59; $p < .05$), but not RW stimuli (M = 83.53°, SD = 6.0), whereas for labial-initial stimuli the opposite was true; theta angle was significantly higher in CVs (M = 81.99°, SD = 7.5) compared to NWs (M = 76.64°, SD = 9.8; $p < .05$), but not RW stimuli (M = 79.40°, SD = 7.3).

The analyses showed that there were no effects of stimulus type or place of articulation (PoA) on VAS slope. We did, however, observe shifts in the category boundary (x_0) driven by stimulus type and PoA. The effect of PoA is quite expected and consistent with production data (e.g., Abramson & Lisker, 1964) showing that the VOT boundary for alveolars is higher compared to labials. However, the rather idiosyncratic effects of stimulus type are more difficult to explain. They may reflect uncontrolled secondary cues (e.g., duration or F_1) that differed among the stimuli. Interestingly, we found a significant difference between labials and alveolars on how much listeners used the secondary cue, with participants showing greater use of F_0 (i.e. smaller theta angle) for labials compared to alveolars). Lastly, there were some significant differences between stimulus types on the theta angle, but they seem to be inconsistent across PoA.

3.3.2 2AFC task results

Participants performed the 2AFC task as instructed. The three participants that were excluded from the VAS analyses (due to failure to perform the task) were also

excluded from the 2AFC analyses. In addition, two additional participants were excluded due to technical issues, leaving 126 participants with valid data for this task.

Participants used both VOT and F_0 to categorize stimuli. As expected, they were more likely to categorize stimuli as /p/ (or /t/) when they had higher VOTs (Figure 3.6) and higher F_0 values (see difference in horizontal positions of two curves). We fitted participants' responses in the 2AFC task using Eq.4 (implemented in Matlab). Overall fits were good ($R^{2} = .99$). Similarly, to the VAS data, we also evaluated the quality of our curve-fits by visually inspecting individual participants' fits.

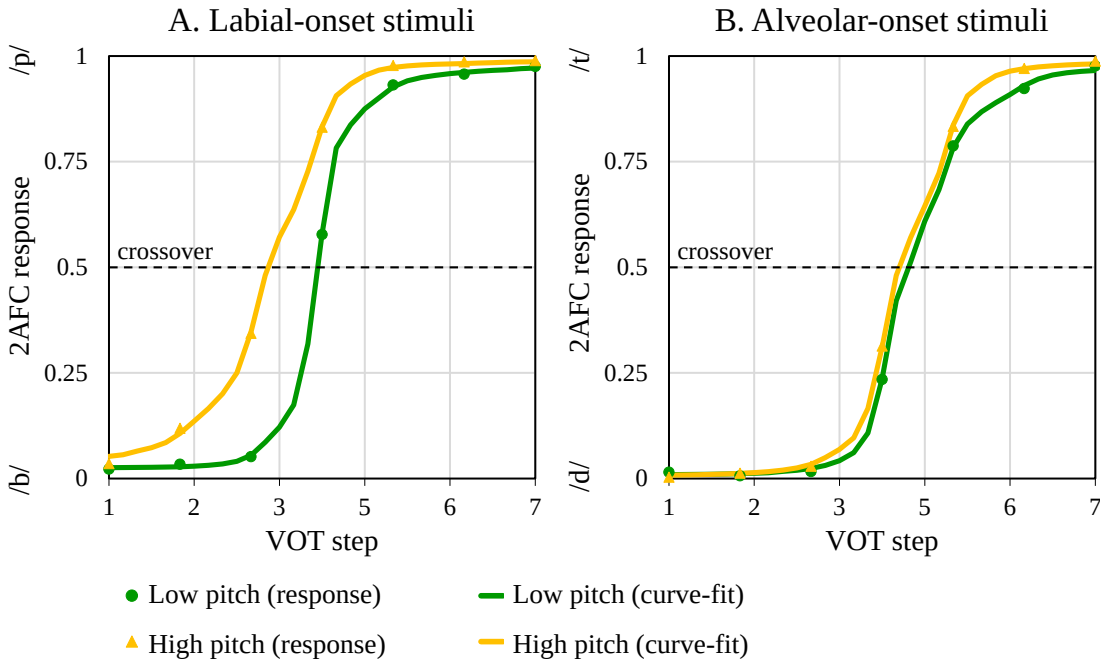


Figure 3.6 Actual and fitted 2AFC responses (green: low pitch; yellow: high pitch)

Figure 3.6 shows the actual and fitted response curves for the two F_0 values across all stimulus types. Because the distribution of raw 2AFC slopes was substantially positively skewed, we used the log-transformed values in all analyses. Similarly,

because the distribution of raw crossover differences (i.e. our measure of F_0 use) was moderately positively skewed, we used the square-root-transformed values in all analyses.

Similarly to the VAS parameters, we then asked whether the 2AFC identification results differed between stimulus conditions by examining the parameters of the curve-fits. We started with two three-way analyses of variance with 2AFC slope and crossover as the dependent variables, and PoA, stimulus type, and F_0 value as the independent variables.

Our first analysis examined 2AFC slope (*s*). It found no significant main effect of PoA, $F < 1$, or stimulus type, $F(2,122) = 1.84$, $p = .164$ (see Figure 3.5.B), but there was a significant effect of F_0, $F(1,122) = 8.25$, $p < .01$, with steeper slopes for low-pitch stimuli ($M = 3.07$, $SD = 3.41$) compared to high-pitch ($M = 2.60$, $SD = 3.54$). The PoA × stimulus type interaction was significant, $F(2,122) = 3.08$, $p < .05$. We split the data by PoA to investigate this interaction. The effect of stimulus type on 2AFC slope was significant for labial-initial stimuli, $F(2,122) = 5.10$, $p < .01$, but not alveolars, $F < 1$. Post-hoc pairwise (Bonferroni adjusted) comparisons revealed that for labial-initial stimuli, NW stimuli showed significantly shallower 2AFC slope values ($M = 2.34$, $SD = 2.19$; $p < .05$) compared to CV stimuli ($M = 3.23$, $SD = 2.17$) and were also significantly shallower compared to RW stimuli ($M = 3.05$, $SD = 2.79$; $p < .05$), but 2AFC slopes did not differ between RW and CV stimuli.

The PoA × F_0 interaction was also significant, $F(1,122) = 11.31$, $p < .01$. When split by PoA, we found that the effect of F_0 on 2AFC slope was significant for labial-

initial stimuli, $F(1,122) = 17.52$, $p < .001$, but not alveolars, $F < 1$. The three-way interaction was not significant, $F(2,122) = 2.70$, $p = .071$.

For the crossover (*co*), the main effect of PoA was significant, $F(1,122) = 303.09$, $p < .001$, with higher crossovers for alveolar-initial stimuli ($M = 4.4$, $SD = .58$) compared to labials ($M = 3.6$, $SD = .61$; see Figure 3.5.D). This is predicted, as later VOT boundaries are generally observed for alveolars (and was also observed with the VAS data). Stimulus type was also significant, $F(2, 122) = 18.91$, $p < .001$, suggesting differences in multiple cue use for different stimulus types. Similarly, to the VAS results, post-hoc pairwise (Bonferroni adjusted) comparisons revealed that CV stimuli ($M = 4.3$, $SD = .72$) differed significantly from both NW ($M = 3.9$, $SD = .71$, $p < .001$) and RW stimuli ($M = 3.8$, $SD = .63$, $p < .001$), but RW did not differ from NW stimuli.

The stimulus type × PoA interaction was also significant, $F(2,122) = 5.00$, $p < .01$. To investigate these interaction terms, we split the data by PoA. The effect of stimulus type on crossover was significant for both alveolars, $F(2,122) = 19.1$, $p < .001$, and labials, $F(2,122) = 7.79$, $p < .001$. Post-hoc pairwise (Bonferroni adjusted) comparisons revealed that for alveolar-initial stimuli, CV stimuli ($M = 4.77$, $SD = .53$) differed significantly from both NW ($M = 4.24$, $SD = .63$; $p < .001$) and RW stimuli ($M = 4.19$, $SD = .38$; $p < .001$), whereas for labial-initial stimuli, CV ($M = 3.80$, $SD = .54$) differed significantly from RW ($M = 3.46$, $SD = .62$; $p < .001$), but not NW stimuli ($M = 3.61$, $SD = .64$; $p = .142$).

Finally, our ANOVA also found a significant effect of F_0, $F(1,122) = 159.5$, $p < .001$, with lower crossovers for high-pitch stimuli ($M = 3.8$, $SD = .76$) compared to low-pitch ($M = 4.2$, $SD = .59$), suggesting that as a whole this task was sensitive to both VOT

and F_0. However, this main effect was moderated by a significant F_0 × PoA interaction, $F(2,122) = 121.95$, $p < .001$, as well as a three-way interaction, $F(2,122) = 31.89$, $p < .001$. Again, we split the data by PoA and found that the effect of F_0 on crossover was significant for both labials, $F(1,122) = 217.85$, $p < .001$, and alveolars, $F(1,122) = 15.8$, $p < .001$, though it was clearly a great deal smaller for the latter.

The effect of F_0 on crossover reflects the degree to which F_0 can shift the boundary (along the VOT dimension) and should be analogous to the θ angle computed from the VAS task. Therefore, we decided to explore the three-way interaction by extracting this measure of secondary cue use (i.e. the difference in crossovers between F_0 values) and looking directly at the effects of PoA and stimulus type on secondary cue use.

For this measure, the main effect of PoA was significant, $F(1,122) = 70.21$, $p < .001$, with labial-initial stimuli showing overall greater difference in crossovers between F_0 values ($M = .74$, $SD = .51$) compared to alveolar-initial stimuli ($M = .36$, $SD = .29$; see Figure 3.5.F). The stimulus type effect was not significant, $F < 1$, but the PoA × stimulus type interaction was significant, $F(2,122) = 13.16$, $p < .001$. To investigate the interaction, we split the data by PoA. The effect of stimulus type on secondary cue use was significant for labial-initial stimuli, $F(2,123) = 7.69$, $p < .001$, and marginally significant for alveolars, $F(2,123) = 3.02$, $p = .053$. Post-hoc pairwise (Bonferroni adjusted) comparisons revealed that for labial-initial stimuli, CV stimuli showed significantly lower differences between crossovers ($M = .51$, $SD = .40$) compared to NW stimuli ($M = .86$, $SD = .58$; $p < .01$) and RW stimuli ($M = .88$, $SD = .47$; $p < .01$), but they did not differ between RW and NW stimuli. The opposite pattern was observed for alveolar-initial stimuli, where CV stimuli ($M = .44$, $SD = .29$) had significantly higher

crossover differences compared to both NW stimuli (M = .34, SD = .29, $p < .001$) and RW stimuli (M = .30, SD = .28, $p < .001$), while RW did not significantly differ from NW stimuli.

Our analyses showed that there were no significant effects of stimulus type or place of articulation (PoA) on 2AFC slope. There were some simple (but no main) stimulus-driven interactions on slope (e.g. steeper 2AFC slopes for labial low-pitch compared to labial high-pitch stimuli). Crossover values were significantly higher in alveolars compared to labials, in CV stimuli compared to the other two stimulus types, and (as expected) in low-pitch compared to high-pitch stimuli. Lastly, F_0 use (quantified as the difference in crossovers between F_0 values) was greater for labial-initial stimuli (compared to alveolars). There were also some simple effects of stimulus type, but they were inconsistent across PoA.

3.4 Discussion

The primary goal of Experiment 1a was to test whether the individual differences in gradiency reported by Kong and Edwards (submitted) can be observed across a variety of stimuli. Our results showed that even though stimuli characteristics should not be overlooked, we can safely use our methodological paradigm (described in Chapter 2) to extract measures of speech perception from different individuals.

Moreover, we conducted some preliminary analyses using our novel VAS-based paradigm in order to evaluate its validity and stability as a measure. When comparing the patterns of responses (e.g., the effect of stimulus type and place of articulation) obtained in the VAS to that of the 2AFC task, we found that the same stimulus-driven effects

appeared in both tasks (see Figure 3.5), and that different aspects of categorization (like categorization boundary and multiple cue use) extracted from both were robustly correlated. That is, there were no main effects of stimulus-type or PoA on either one of the slopes, crossover values were significantly higher for alveolar-initial stimuli and for CV stimuli (compared to the other two stimulus types) for both tasks, and participants showed greater use of F_0 when categorizing labials in both tasks (in the form of smaller theta angle for the VAS task, and greater difference in crossovers between F_0 values for the 2AFC task). In addition, in regards to F_0 use, we even found the same pattern of PoA × stimulus type interaction in both tasks (i.e. greater F_0 use for alveolar CVs compared to alveolar-initial NW stimuli, and the opposite pattern for labial-initial − smaller use of F_0 for labial-initial CVs compared to labial-initial NW stimuli). These analyses provide strong support for the use of VAS measures to assess speech categorization.

This close match between the pattern of VAS and 2AFC results is quite reassuring and provides strong support for the VAS task as an accurate and precise measure of phoneme categorization. As a whole, this suggested that our speech perception measures can reveal aspects of speech perception that are somewhat fixed, thus validating an individual differences approach, as the one taken here.

CHAPTER 4: THE ROLE OF GRADIENCY IN SPEECH PERCEPTION (EXPERIMENT 1B)

4.1 Introduction

Experiment 1a provided strong evidence that our VAS-based measure is a precise and valid measure of phoneme categorization gradiency independently of multiple cue integration. Thus, we moved on to address some of our primary questions regarding the relationship between categorization gradiency and others aspects of speech perception, such as multiple cue and noise, its sources, and its role in speech processing.

To do so, we started by relating our gradiency measure to a more standard measure of categorization extracted from the 2AFC phoneme identification task. While the 2AFC slope can reflect both (a) the gradiency of categorization and (b) the noise in cue-to-category mapping, an explicit comparison between the ratings we collect from the two tasks can help disentangle what the 2AFC task is primarily measuring (gradiency or internal noise). Since both tasks are thought to reflect, at least to some degree, categorization gradiency, we expected to find a positive correlation between the VAS and the 2AFC slopes. However, it was not clear how strong a correlation should be expected, given the ambiguity as to what factors affect performance in the 2AFC task.

More importantly, we also related the gradiency measure (from the VAS task) to multiple cue integration (from the 2AFC task), indexed by the influence of the secondary cue on categorization responses. As we describe above, we predicted that gradient categorizers would be more sensitive to fine-grained information and should, therefore, be better at taking advantage of subtle acoustic differences across multiple cues (this would also be in accordance with the Kong and Edwards results).

Next, we extended earlier investigations by addressing whether these speech measures (gradiency and multiple cue integration) were related to more general (i.e. not language-specific) cognitive abilities. To do so, we collected a series of individual differences measures tapping different aspects of executive function to evaluate these higher cognitive processes as possible (direct or indirect) sources of gradiency. Our hypothesis was that, to the extent that speech perception may draw on domain-general skills, like executive function or working memory, individual differences in these skills may be reflected in speech perception tasks. An investigation of individual differences may thus allow us to identify the constellation of skills that are assembled for perception of speech.

Finally, we performed a preliminary assessment of the functional role of gradiency (i.e. whether it is beneficial for speech perception) using a speech-in-noise recognition task and correlating participants' performance in this task to our measure of gradiency.

4.2 Methods

4.2.1 Participants

(see Section 3.2.1 Participants)

4.2.2 Design and tasks

Immediately after the hearing screening, participants performed a series of six tasks that measured different aspects of speech perception and executive function (see Table 4.1). In addition to the VAS and the 2AFC tasks described in Chapter 3,

participants also performed three tasks measuring different cognitive functions, all roughly linked to different aspects of executive function (EF). We used the Flanker task to assess the inhibitory component of EF, the N-Back task, which taps primarily working memory, and the Trail Making task for a more general measure of planning, cognitive flexibility, and executive performance. Finally, to extract a measure of speech perception accuracy, we administered a computerized version of the AzBio sentences (Spahr et al., 2012).

4.2.3 VAS task

(see Section 3.2.3)

4.2.4 2AFC task

(see Section 3.2.4)

Table 4.1 Order and description of tasks in Experiment 1 (Experiments 1a and 1b)

Order	Task	Domain	Primarily measure of…
1	VAS	Speech categorization	phoneme categorization gradiency
2	Flanker	Cognitive	executive function: inhibitory control
3	N-Back	Cognitive	executive function: working memory
4	2AFC	Speech categorization	secondary cue use
5	Trail Making	Cognitive	executive Function: general
6	AzBio	Speech perception in noise	speech perception accuracy

4.2.5 Measures of executive function

4.2.5.1 The Flanker task (inhibitory control). The Flanker task is commonly considered a measure of inhibitory control (Eriksen & Eriksen, 1974). During this task, participants saw five arrows at the center of a computer screen and reported the direction of the middle arrow by pressing one of two keys. Crucially, the direction of the other four arrows (flankers) was either consistent or inconsistent with that of the middle (target) arrow. On inconsistent trials, the degree to which participants can inhibit the irrelevant flanking stimulus predicts their speed of responding. The Flanker task had 20 trials and it took approximately 3 mins to administer. Performance in the task was calculated based on the NIH Toolbox instructions, creating a measure that is a composite of both speed and accuracy[6].

4.2.5.2 The N-Back task (working memory). The N-back task is commonly used as a measure of complex working memory (Kirchner, 1958). In this task, participants viewed a series of numbers (one at a time for 2000 ms each) on a computer screen and indicated whether the currently presented number was the same with or different than the number presented immediately previously (1-back), two numbers previously (2-back), or three numbers previously (3-back). The three levels of difficulty were presented in this order (easiest to hardest) for all participants. There were 41, 42, and 43 trials for each of the three levels of difficulty respectively, thus resulting in 40 responses to be scored in each level. The N-Back task took approximately 9 mins. Overall accuracy across the three levels of difficulty was taken as an indicator of working memory capacity.

[6] Flanker task accuracy score = 0.125 * Number of Correct Responses; Reaction Time (RT) Score = 5-(5*((log(RT)-log(500))/(log(3000-log(500)); If accuracy levels are < = 80%, the final "total" computed score is the accuracy score. If accuracy levels are > 80%, reaction time score and accuracy score are combined.

4.2.5.3 The Trail Making task (cognitive control/flexibility). Part B of the Trail Making task is a common neuropsychological assessment of cognitive control (Tombaugh, 2004). During this task, participants are presented with a sheet of paper depicting circles containing numbers 1 through 16 and letters A through P. Their task is to use a pencil to draw a line connecting the circles in order, alternating between numbers and letters, starting at number 1 and ending at letter P. The total time that a participant needs to complete this task is recorded by a trained examiner and is used as a measure of cognitive control. On average, the Trail Making task took 2.5 mins to administer.

4.2.6 Speech recognition in noise: The AzBio sentences

In order to measure how well participants perceive speech in noise we administered the AzBio sentences (Spahr et al., 2012), which consists of ten sentences masked with noise (0 dB SNR). The sentences were delivered over high-quality headphones and the participants were given unlimited time to repeat each sentence. An examiner was present in the room and reported the number of correctly identified words on a computer display by clicking on each word of the sentence that was correctly produced. The computer monitor was turned away so that the participant could not see the sentences. The AzBio task took approximately 7 mins to administer. The logit-transformed percentage of correctly identified words across the 10 sentences was used as a measure of overall performance.

4.3 Results

4.3.1 Response consistency/noise and gradiency in phoneme categorization

We first examined the relationship between participants' categorization gradiency (reflected by the VAS slope) and the 2AFC slope, which may reflect both the noise in encoding cues like VOT, and/or the consistency with which they assign stimuli to categories. By averaging the slope values across the two places of articulation, there were no longer any repeated measurements. This enabled us to use hierarchical regression to evaluate VAS slope as a predictor of 2AFC slope (see Table 4.2).

Table 4.2 Hierarchical regression steps: predicting 2AFC slope from VAS slope

		B	*SE*	*β*	R^2
Step 1					0.018
	RW vs others	0.051	0.036	0.150	
	CV vs others	0.035	0.036	0.105	
Step 2					0.020
	VAS slope	0.084	0.163	0.047	
Step 3					0.068
	VAS slope × RW vs others	-0.226	0.123	-1.168^{+}	
	VAS slope × CV vs others	0.088	0.138	0.464	

$^{+}$p<.1, *p<.05, **p<.001, ***p<.001.

On the first level of the model, stimulus type was included as a predictor to account for any between-subject variance in stimulus type. This was contrast-coded into two variables, one comparing CVs to the other two (CV = 2; RW = -.1; NW = -1), and the other comparing RWs to the other two (RW = 2; NW = -1; CV = -1). This explained

1.78% of the variance, which was not significant, $F(2,117) = 1.06$, $p = .35$, as would have been predicted by the prior analyses.

On the second step, VAS slope was added to the model, which did not account for a significantly larger portion of the variance ($R^2_{change} = .002$, $F_{change} < 1$). On the last step, we included the VAS slope × stimulus type interaction, which accounted for a marginally significantly larger portion of the variance ($R^2_{change} = .048$, $F_{change}(5,114) = 2.96$, $p = .056$). To examine this interaction, we split the data by stimulus type. VAS slope did not account for a significant portion of the 2AFC slope variance in any of the subsets.

The lack of a significant relationship between the slopes for the two tasks was initially cause for alarm, that perhaps the VAS task is not related to more standard speech categorization measures. Thus, to confirm that the VAS task could in fact provide good measures of basic aspects of speech perception (such as category boundary and secondary cue use), we also examined correlations between the crossover and F_0 use extracted from the two tasks. Indeed, for both PoA, we found significant positive correlations between the category boundaries (i.e. crossover) extracted from the VAS task and those extracted from the 2AFC task, $r(122) = .375$, $p < .001$; $r(122) = .581$, $p < .001$, (see Fig. 4.1.A). Similarly, as expected, we found a significant negative correlation between secondary cue use[7] extracted from the VAS task (i.e. theta angle) and the one extracted from the 2AFC task (i.e. difference between crossovers), $r(122) = -.454$, $p < .001$, (see Fig. 4.1.B). These results confirm our assumption that there is a robust relationship between these two measures, which supports the validity of the VAS task as a measure of speech perception.

[7] Since preliminary analyses (see above) showed that participants used pitch information more robustly for labials, we only computed the correlation between the two measures of secondary cue sue for those stimuli.

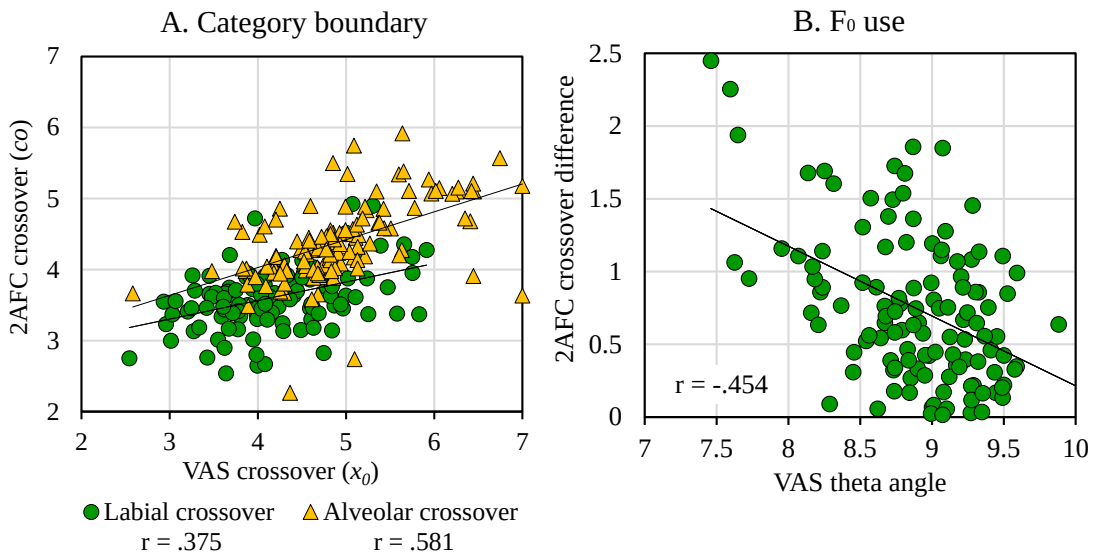


Figure 4.1 Correlations between VAS and 2AFC parameters

Returning to the slope, this lack of correlation between 2AFC and VAS slope is intriguing and implies that the two measures may reflect different aspects of speech categorization. In line with the understanding of the two tasks laid out in Chapter 1, this suggests that the 2AFC task may be more sensitive to noise in the encoding of continuous phonetic cues (and/or their mapping to categories), while the VAS reflects phoneme categorization gradiency. Indeed this is in line with Figure 3.3 from the previous chapter, which suggests that two participants may have similar mean slopes in the VAS task despite large differences in the trial-to-trial noise around that mean. While the 2AFC task cannot assess this, the VAS task may be able to.

To test this hypothesis, we extracted a measure of noise in cue encoding from the VAS task based on the residuals of the function. We first computed the difference between each VAS rating (on a trial-by-trial basis) and the predicted value based on the parameters we estimated for that participant using the rotated logistic (i.e. the residual

variance). We then computed the deviation of those residuals from the mean as the square root of the average of the squared differences (i.e. the standard deviation of the residuals). This was done separately for each PoA. The average SD across PoA was used as an estimate of noise for each participant. Finally, this measure of noise was correlated with the 2AFC and VAS slopes. If we found a negative correlation between the steepness of the 2AFC slope and the noise estimate extracted from the VAS task, this would confirm that 2AFC slopes may derive more from noise in the encoding than from gradiency in the mapping to categories.

The SD of the residuals in the VAS task was marginally significantly correlated with 2AFC slope in the expected direction (negatively), $r = -.168$, $p = .063$, suggesting that (as expected) listeners with steeper 2AFC slopes showed lower levels of noise in the VAS task. Interestingly, this measure was weakly positively even though not significantly correlated with VAS slope, $r = .120$, $p = .185$, suggesting that, if anything, listeners with higher gradiency (i.e. shallower VAS slope) may be *less* noisy in their VAS ratings. That would also be consistent with the sample results presented in Figure 3.3 (in previous chapter), in the sense that more gradient listener seem to give ratings that more systematically reflect the stimulus characteristics.

4.3.2 Secondary cue use as a predictor of gradiency

Next we examined whether gradiency in phoneme categorization was linked to multiple cue integration. Similarly to above, we tested this using hierarchical (multi-level) regression with VAS slope as the dependent variable. The independent variables were stimulus type (coded as before) and F_0 use, measured as the difference in crossover

points (in the 2AFC task) between the low and high F_0. Only labial-initial stimuli were included in this analysis, because participants showed significantly higher overall use of pitch for those stimuli. In the first level of the model, stimulus type was entered as a predictor and non-significantly accounted for 1.4% of the variance, $F < 1$. In the second level, F_0 use was added as a predictor. This explained a significant portion of the variance, $\beta = -.296$; $R^2_{change} = .077$, $F_{change}(1,116) = 11.23$, $p < .01$. On the last level, we included the F_0 use × stimulus type interaction, which did not significantly account for any additional variance ($R^2_{change} = .024$, $F_{change}(2,114) = 1.53$, $p = .22$).

Table 4.3 Hierarchical regression steps: predicting VAS slope from F_0 use

		B	*SE*	*β*	R^2
Step 1					0.014
	RW vs others	0.012	0.027	0.047	
	CV vs others	0.034	0.027	0.133	
Step 2					0.091
	F_0 use	-0.341	0.108	-0.296**	
Step 3					0.115
	F_0 use × RW vs others	0.077	0.090	0.098	
	F_0 use × CV vs others	-0.092	0.085	-0.105	

*p<.05, **p<.001, ***p<.001.

These results corroborate the findings of Kong and Edwards (submitted): listeners who show higher phoneme categorization gradiency (shallower VAS slope) also showed greater use of F_0, thus suggesting a potential link between these two aspects of speech perception.

4.3.3 Executive function and gradiency

Next we explored the relationship between general cognitive functions and phoneme categorization gradiency. Because the distribution of raw N-Back scores reflected percent accuracy, while the distribution of the Trail Making task was moderately positively skewed, we transformed scores using empirical logit (N-Back) and square-root (Trail Making) functions in all analyses (no transformation was necessary for the Flanker scores).

We first estimated the correlations between the different executive function measures. Flanker (inhibition) was not significantly correlated with either N-Back (working memory; $r = .01$) or Trail Making (executive function; $r = .12$). However, N-Back performance was weakly, but significantly, correlated with Trail Making ($r = .19$, $p < .05$).

We then conducted a series of multiple regression analyses to explore the relationship between phoneme categorization gradiency and different aspects of executive function. Three regression models were fitted−one for each executive function measure—using VAS slope (averaged across PoA) as the dependent variable. In the first level of each model we entered stimulus type as a predictor. In the second level, each of the three executive function measures was added.

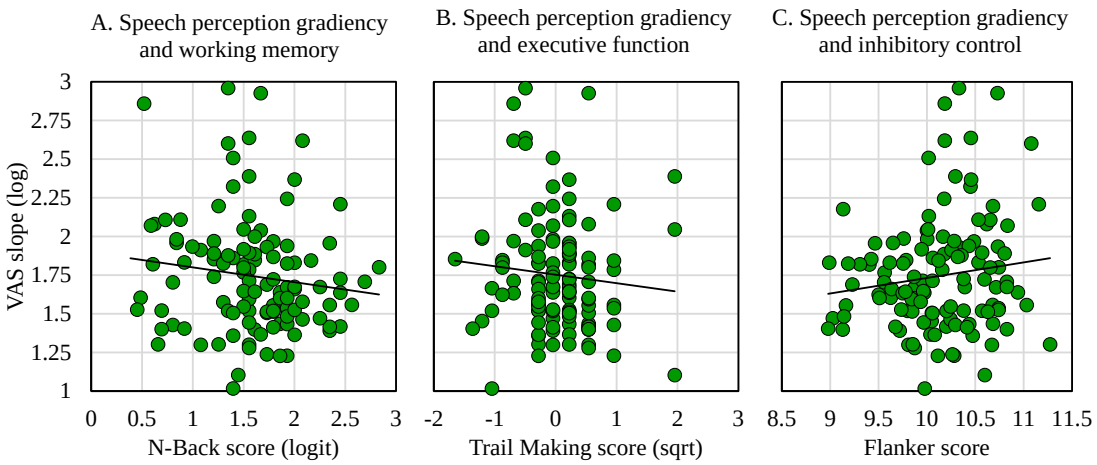


Figure 4.2 VAS slope by executive function measures scatterplots

As shown by prior analyses (see Section 3.3.1), stimulus type did not have a significant effect on VAS slope (see Step 1 in Table 4.4). When N-Back score was entered as a predictor, it explained a significant portion of the VAS slope variance, with higher N-Back scores marginally significantly predicting shallower VAS slopes, $\beta = -.171$; $R^2_{change} = .029$, $F_{change}(1,110) = 3.24$, $p = .075$ (Figure 4.2.A). Trail Making score did not predict VAS slope, $R^2_{change} = .011$, $F_{change}(1,116) = 1.92$, $p = .28$ (Figure 4.2.B), nor did Flanker score, $R^2_{change} = .011$, $F_{change}(1,118) = 1.29$, $p = .26$ (Figure 4.2.C).

Table 4.4 Hierarchical regression steps: predicting VAS slope from executive function measures

		B	*SE*	*β*	R^2
Step 1					0.002
	RW vs others	0.09	0.023	0.043	
	CV vs others	0.001	0.022	-0.003	
Step 2a					0.030
	N-Back	-0.095	0.053	-0.171^{+}	
Step 2b					0.011
	Trail Making	-0.047	0.043	-0.101	
Step 2c					0.011
	Flanker	0.059	0.052	0.104	

$^{+}$p<.05, *p<.05, **p<.001, ***p<.001.

4.3.4 Executive function and multiple cue integration

Given the results reported in the two previous sections, showing a relationship (1) between gradiency and multiple cue integration and (2) between gradiency and N-Back performance (i.e. working memory), we wanted to test the possibility that the first two (gradiency and multiple cue integration) may be driven by a third factor, possibly related to executive function. For example, it could be that greater working memory span allows listeners to better maintain within-category information *and* better combine primary and secondary cues. We addressed this possibility using hierarchical regression with secondary cue use for labials as the dependent variable. As above, three regression models were fitted−one for each executive function measure. In the first level of each model we entered stimulus type as a predictor. In the second level, each of the three measures was added.

Table 4.5 Hierarchical regression steps: predicting secondary cue use from executive function measures

		B	*SE*	*β*	R^2
Step 1					0.120
	RW vs others	0.004	0.022	0.017	
	CV vs others	-0.071	0.022	-0.337**	
Step 2a					0.123
	N-Back	0.036	0.053	0.062	
Step 2b					0.126
	Trail Making	-0.043	0.047	-0.083	
Step 2c					0.126
	Flanker	-0.051	0.055	-0.084	

*p<.05, **p<.001, ***p<.001.

As it is also reported in Section 3.3, stimulus type had a significant effect on secondary cue use (see Step 1 in Table 4.5), with significantly higher crossovers observed for CV stimuli. On the second level of the analysis, none of the EF measures were correlated with secondary cue use (N back: $R^2_{change} = .003$, $F_{change} < 1$; Trail Making: $R^2_{change} = .006$, $F_{change} < 1$; Flanker: $R^2_{change} = .006$, $F_{change} < 1$).

These results seem to suggest that whatever the nature of the relationship is between gradiency and multiple cue integration, it is unlikely to be driven by a third factor related to higher cognitive functions, such as executive function, at least insofar as it was assessed by the measures used in this study.

4.3.5 Perception of speech in noise

Finally, we turned to the hypothesis that maintaining within-category information may be beneficial for speech perception more generally. In a preliminary examination of

the data, we found that our measure of speech recognition in noise (AzBio) was weakly negatively correlated with VAS slope (r = -.143), though this was not significant (p = .116). However, we also noticed that perception of speech in noise was significantly correlated with both N-Back performance (r = .29, p < .01) and Trail Making (r = .29, p < .01), and marginally correlated with Flanker performance (r = .18, p = .055). Since AzBio scores were correlated with the executive function measures, we decided to assess the relationship between gradiency and perception of speech in noise after adjusting for executive function parameters.

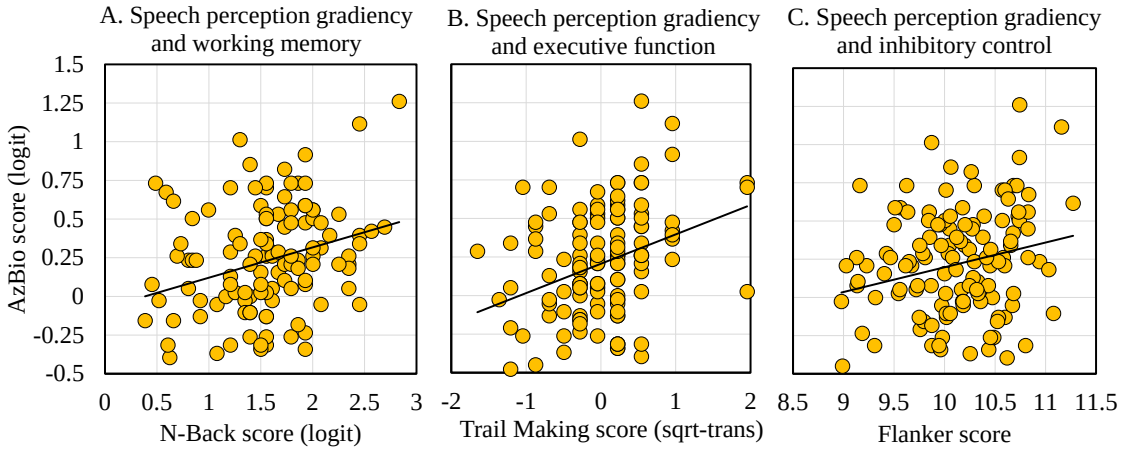


Figure 4.3 AzBio score by executive function measures scatterplots.
Note: AzBio logit score of 0 (zero) corresponds to 50% accuracy

We fitted and compared a hierarchical linear regression with three levels and AzBio score as the dependent variable. In the first level, our three executive function measures were entered as predictors (see Step 1a in Table 4.6), which significantly predicted AzBio score, F(3,108) = 6.76, p < .001, explaining 15.8% of the variance. Within this level, N-Back score was a significant predictor, β = .24, p < .01, as was Trail Making, β = .22, p < .05, while Flanker score was marginally significant, β = .15, p =

.085. As indicated by the direction of the beta coefficients (and the plots in Figure 4.3), higher scores in each of the executive function measure predicted better performance in the AzBio task.

Table 4.6 Hierarchical regression steps: predicting AzBio score from VAS slope

		B	*SE*	*β*	R^2
Step 1a					0.158
	N-Back	0.164	0.062	0.238*	
	Trail Making	0.138	0.056	0.223*	
	Flanker	0.113	0.065	0.155^{+}	
Step 1b					0.025
	VAS slope	-0.159	0.109	-0.127	
Step 2					0.165
	N-Back	0.156	0.063	0.226*	
	Trail Making	0.132	0.056	0.213*	
	Flanker	0.121	0.065	0.167^{+}	
	VAS slope	-0.108	0.113	-0.086	
Step 3					0.172
	VAS slope × N-Back	0.147	0.208	0.070	
	VAS slope × Trail Making	0.011	0.205	0.005	
	VAS slope × Flanker	-0.124	0.317	-0.039	

$^{+}$p<1, *p<.05, **p<.001, ***p<.001.

In the second level, we added VAS slope as a predictor, which did not account for a significantly greater portion of the variance over and above the cognitive measures, $R^2_{change} = .007$, $F_{change} < 1$. Finally, in the third level, we added the VAS slope × N-Back score, VAS slope × Trail Making score, and VAS slope × Flanker score interactions as

predictors. None of the interaction terms accounted for a significant portion of the variance over and above that accounted for by the main effects, $R^2_{change} = .007$, $F_{change} < 1$, $p = .84$. In other words, even though there was a hint of a positive correlation between gradiency and AzBio performance (as reported earlier), when the three executive function measures were added in the model, this relationship disappeared.

Next, we followed the reverse procedure. In this model, we entered VAS slope in the first step (see Step 1b in Table 4.6). This was not significant, $\beta = -0.127$; $F(1,110) = 1.81$, $p = .181$, explaining 1.6% of the variance. In the second level, we added the cognitive measures as predictors, which accounted for a significantly greater portion of the variance over and above that of the VAS slope, $R^2_{change} = .149$, $F_{change}(3,107) = 6.37$, $p < .001$. This suggests that the relationship between gradiency and speech perception in noise may be largely due to individual differences in our measures of executive function, with little unique variance that can be attributed to gradiency.

4.4 Discussion

Experiment 1b aimed at using the VAS task to test a set of specific predictions regarding the role of phoneme categorization gradiency in speech perception. Specifically, we examined whether and how phoneme categorization gradiency may be linked to 1) internal noise in the mapping of cues to phonemes, 2) multiple cue integration, 3) different aspects of executive function, and 4) perception of speech in noise.

While our most important finding was the correlation between VAS slope (phoneme categorization gradiency) and multiple cue integration, our correlational

approach offers a number of additional insights that are worth discussing before we turn to the implications of our primary finding.

4.4.1 VAS slope and 2AFC slope

One of the most striking results of the present study was the lack of correlation between the VAS slope and the 2AFC slope. Indeed, we expected to find some correlation between the two, since both are thought to reflect, at least partly, the degree of gradiency with which an individual categorizes speech sounds. Here, however, we found that the 2AFC slope did not predict VAS slope. This finding could mean that these particular aspects of these tasks assess quite different aspects of speech perception, perhaps more so than what we initially thought. That is, the 2AFC slope may largely reflect noise in encoding, rather than the gradiency of the response function (as does the VAS slope).

There are a number of arguments that support this claim. For example, even if listeners make underlying probabilistic judgements about phonemes, when it comes to mapping this judgement to a response, the optimal strategy is to always choose the most likely response (as opposed to attempting to match the distribution of responses to the internal probability structure; Nearey & Hogan, 1986). Though it is unclear if some (or all) listeners do this, it suggests that the 2AFC slope may not perfectly reflect the underlying probabilistic mapping from cues to categories. Thus, in contrast to the 2AFC task, the VAS task may offer a unique window into this mapping, allowing us to extract information that is not accessible with other tasks, such as the 2AFC task. Indeed, this is supported by our own analyses that demonstrate trial-by-trial variation showing a

markedly different relationship with 2AFC than with VAS slope. While these results should be interpreted cautiously (as the 2AFC × noise correlation was only marginally significant), they seem to suggest that variation in 2AFC slope may be most closely tied to noise in the system (higher noise → shallower slope), whereas VAS slope may much more directly reflect the gradiency of speech categories.

This has a number of major implications when we consider the use of phoneme categorization measures to assess speech processing in populations with communication impairments. First, our findings seem to explain why gradient 2AFC responding is often associated with SLI and dyslexia, even as theoretical models and work with typical populations suggest a more gradient mode of responding is beneficial. In the former case, the 2AFC task is not tapping mode of responding at all, but rather is tapping internal noise (of which impaired listeners are likely to have more). Second, as we have shown here, measures like the VAS may be able to simultaneously tap both, with the slope of the average responding reflecting categorization gradiency and the SD of the residuals reflecting noise. The combination of these measures may thus offer far more insight into the locus of phonological or perceptual impairments than a traditional 2AFC measure, particularly when combined with sensitive online measures like eye-tracking (c.f., McMurray et al., 2014) that overcome other limitations of phoneme judgment tasks.

4.4.2 Phoneme categorization gradiency and multiple cue integration

Another critical result was that phoneme categorization gradiency seems to be linked to multiple cue integration, such that higher use of pitch-related information predicts greater gradiency (see also Kong & Edwards; submitted). Even though this

finding was predicted, it is also correlational, and therefore consistent with a number of possible causal accounts. First, multiple cue integration may allow for higher degree of categorization gradiency. Under this view, the ability to integrate multiple cues may help listeners form a more precise graded estimate of the speech categories. Alternatively, as we proposed in Chapter 1, the causality may operate in the reverse, with more gradiency allowing listeners to be more sensitive to small differences in each cue, permitting better integration. Third, operating in a similar direction, a gradient representation could help listeners avoid making a strong commitment on the basis of a single cue, allowing them to use both cues more effectively. Lastly, there could be a third factor that links the two. In this regard, we examined executive function measures and found a relationship with gradiency for only the N-back task, but no relationship between any EF measures and multiple cue integration. However, a variety of other such factors need to be considered in future research. For example, it could be that listeners with greater auditory acuity are more sensitive to fine-grained differences across all speech cues, which also allows them to show higher gradiency and to integrate cues better. Even though our study was not designed to distinguish between these mechanisms, it offers strong evidence for a link between these two aspects of speech perception, which remains to be clarified.

4.4.3 Links between phoneme categorization gradiency and broader cognitive processes

Our findings show a potential link between working memory (the N-Back task) and participants' response pattern in the VAS task. One possibility for this correlation is that working memory mediates the relationship between gradiency in the system and individuals' responses; there may be individuals who do show gradiency at the cue and

phoneme level, but this gradient activation is not maintained all the way to their response due to working memory limitations. In other words, the degree to which gradiency at the level of activations (at the cue of phoneme level) is reflected in an individual's response pattern may be dependent on their working memory span. It may be that measures that tap earlier stages of processing (e.g., ERPs, see Toscano et al., 2010), or earlier times in processing (e.g., eye-movements in the visual world paradigm: McMurray, et al., 2002) may be less susceptible to working memory constraints, possibly explaining why these measures offer some of the strongest evidence for gradiency as a characterization of the modal listener.

4.4.4 Phoneme categorization gradiency and perception of speech in noise

Participants performed a speech-in-noise task (AzBio sentences) as part of a preliminary exploration of the functional value of gradiency in speech perception. Our hypothesis was that higher gradiency may allow listeners to be more flexible in their interpretation of the signal and, thus, outperform listeners with lower levels of gradiency in our speech-in-noise task. In contrast to our prediction, gradiency was not a significant predictor of performance in the AzBio task, which was, however, significantly predicted by our three executive function measures (N-Back, Trail Making, and Flanker task).

This lack of correlation between our gradiency measure and performance in the AzBio task may reflect difficulties in linking laboratory measures of underlying speech perception processes (and cognitive processes more generally) to simple outcome measures. Such difficulty could arise from at least two sources. First, speech-in-noise perception may be more dependent on participants' level of motivation and effort than

the laboratory measures. This is supported by recent work on listening effort (Wu, Stangl, Zhang, Perkins, & Eilers, in press.; Zekveld & Kramer, 2014), which suggests that listeners put forth very low effort at low signal-to-noise ratios – in fact, they appear to just give up. Even though it is highly unlikely that in our study participants gave up in the AzBio task, the point being made here is that variation in motivation may be a significant source of unwanted variability in these measures. Indeed, it is possible that the significant correlations between our speech-in-noise measure and scores on the three executive function tasks, may derive from a similar source. If so, this correlation may have little to do with speech perception processes.

Furthermore, while speech-in-noise perception is a standard assessment of speech perception accuracy, performance in such tasks may not be strongly affected by differences in categorization gradiency. As we describe in Chapter 1, theoretical arguments for gradiency are not typically framed in terms of speech-in-noise perception; rather the motivation seems to derive from the demands of interpreting ambiguous acoustic cues, such as those related to anticipatory coarticulation, speaking rate, or speaker differences. Noise does not necessarily alter the cue values; rather it masks the listeners' ability to detect them. Thus, this task may not properly target the functional problems that categorization gradiency is attempting to solve.

In a related vein, it may be the case that both gradient and categorical modes of responding are equally adaptive for solving the problem of speech perception in noise. That is, to the extent that differences in listeners' mode of categorization reflects a different weighting of different sorts of information (e.g., between acoustic or phonological representations in the Pisoni & Tash, 1974, model; or between dorsal and

ventral stream processing in the Hickock & Poeppel, 2007, model), both sources of information may be equally useful for solving this problem (even as there are advantages of gradiency for other problems).

Gradiency and non-gradiency in the categorization of speech sounds can both be advantageous in different ways. Therefore, in order to find the link that connects the underlying cognitive processes to a performance estimate, we need to use different measures of performance that are more closely tied to the theoretical view of speech perception that is being evaluated. Similar concerns may suggest the need to reconsider the way we evaluate speech perception tests used in a variety of different settings, including for clinical evaluations, so that they tap more into the underlying processes linked to our predictions.

The key results of Experiment 1b can be outlined as followed: First, we showed that differences in phoneme categorization gradiency seem to be theoretically independent from differences in the degree of internal noise in the encoding of acoustic cues and/or cue-to-phoneme mappings. Thus, our results question the traditional interpretation of shallow slopes as indicating noisier categorization of phonemes. Both categorization gradiency and such forms of noise contribute to speech perception, but may be tapped by different tasks. Second, differences in categorization gradiency seem to matter for speech perception, as they appear to be linked to differences in multiple cue integration. Third, we found only limited relationship between executive function and gradiency, suggesting differences in categorization sharpness may derive from lower-level sources. Lastly, gradiency may be weakly related to speech perception in noise, but this seems to be modulated by executive function-related processes.

These results provide useful insights as to the mechanisms that subserve speech perception. Most importantly, they seem to stand in opposition to the commonly held assumption that a sharp category boundary, along with poor within-category discrimination, is the optimal approach for categorizing speech sounds. Overall, speech sound categorization is gradient, although to different degrees among listeners, and further work is necessary to reveal the sources of these differences and the consequences they may have for spoken language comprehension.

CHAPTER 5: THE SOURCES OF GRADIENCY IN EARLY AUDITORY PROCESSING (EXPERIMENT 2)

5.1 Introduction

Experiment 1 examined individual differences in speech perception gradiency by (1) linking this gradiency to a different aspect of speech perception, multiple cue integration; (2) exploring possible links between patterns of speech perception and general cognitive processes, such as executive function (EF); and (3) investigating the functional role of gradiency in speech perception accuracy more generally. Our findings suggested a weak link between gradiency and higher cognitive processes (specifically, working memory), however effects were not robust and may reflect other factors like the degree to which within-category information can be maintained to the response stage. Moreover, neither our measure of inhibitory control (the Flanker task), nor our more general EF measure (the Trail Making task) were correlated with gradiency. Thus, there is little evidence for a strong causal relationship between speech perception gradiency and executive function. This leaves open the issue of what are the sources of phoneme categorization gradiency.

The primary goal of this experiment was to test alternative hypotheses as to the sources of individual differences in phoneme categorization gradiency, focusing this time on differences *within* the language/speech perception system (rather than outside it). In particular, we examine 1) whether such differences stem from differences in how listeners encode acoustic cues during early processing stages; and 2) whether they are linked to individual differences in lateral inhibition between words (i.e. local inhibition

within the word recognition system). Lastly, the weak but moderate effects of EF led us to continue that investigation with a new EF measure (spatial Stroop).

A secondary goal was to examine the relationship between our VAS measure of phoneme categorization gradiency and other existing measures of gradiency at different levels of language processing. To do this, we correlated our results with a standard eye-tracking paradigm used to assess gradiency at the lexical level. In particular, this paradigm allows us to evaluate the degree to which listeners' gradient sensitivity to acoustic cues affects the strength of lexical activations independently of their responses. In that way, it provides a measure of how strongly do listeners maintain a lexical representation partially active (i.e. in a gradient manner), even when they commit to another word. By examining this more focused (within-category) form of gradiency, we can ask whether the gradiency observed in the VAS task reflects how continuous cues are mapped onto categories more generally (i.e. both within and across categories).

In the remainder of this introduction we talk about the theoretical motivation for each of these goals separately.

5.1.1 Sensory-level processes

The primary question addressed by Experiment 2 was whether differences in phoneme categorization gradiency observed in the VAS task are driven by differences at earlier stages of processing, and specifically at the level of cue encoding. In examining this, we assume for simplicity a two-stage process; first listeners encode continuous cues (such as VOT and F_0), and then they map them onto phoneme categories. If listeners encode cues in a graded way, this should allow for either a graded or a categorical

activation of phoneme categories depending on how cues are mapped to categories (see Figure 5.1.A). However, if listeners encode cues categorically, this should in turn limit their sensitivity to within-category phoneme differences, which would be reflected in a more categorical/step-like pattern of phoneme categorization (see Figure 5.1.B).

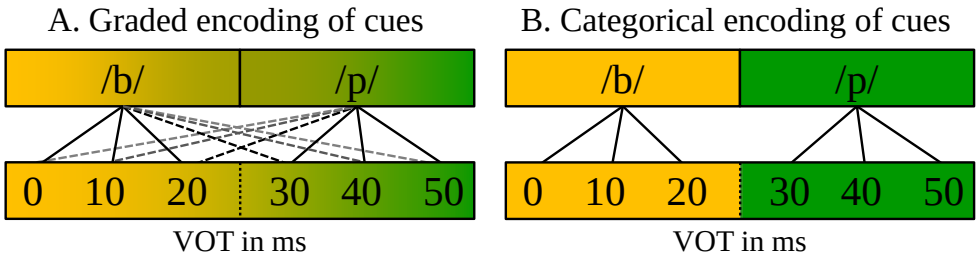


Figure 5.1 Examples of graded and categorical mapping of speech cues to phoneme categories

To examine cue encoding more directly, we used an event-related potential (ERP) paradigm based on a study by Toscano, McMurray, Dennhardt, and Luck (2010). This study used an ERP paradigm to test (among others) whether continuous information in the speech signal is reflected by neural markers of early perceptual processing. The key measure was the amplitude of the fronto-central auditory N1, a negative ERP component that is thought to be generated in Heschl's gyrus and is considered a marker of the perceptual encoding of auditory information. Specifically, this component appears ~100 ms post stimulus onset and is thought to reflect early neural encoding of VOT (Sharma & Dorman, 1999; Sharma, Marsh, & Dorman, 2000).

Toscano et al examined whether the early perceptual encoding of speech cues is affected by category-related information. Previous studies using the N1 as a marker of cue encoding have reported that stimuli with short VOTs elicit a single N1 peak, while

long VOTs elicit a double peak (presumably one peak for the release, and a second one for the onset of voicing). This morphology shows a qualitative shift as VOT increases (from a single-peak to a double-peak morphology), which has been used to argue in favor of a category effect on cue encoding. However, as pointed out by Toscano et al, this discontinuity could be an artifact of the high-amplitude burst of the stimuli; in the case of long VOTs, both the release burst and the voicing onset may elicit a separate N1 (thus the double peak), while in the case of short VOTs the two merge together. Toscano et al avoided this issue by using stimuli with low-amplitude bursts.

Specifically, they presented stimuli varying continuously in VOT (e.g. nine steps from *beach* to *peach*) and measured the auditory N1 amplitude. Their hypothesis was that, if listeners encode fine-grained, within-category differences in a veridical way, they would observe a linear relationship between VOT and N1 amplitude, whereas a more categorical approach should show a discontinuity near the boundary. Results strongly favored a linear model (see also Frye et al., 2007, for analogous findings showing linear encoding of VOT in the M100, which is the megnetoencephalographic equivalent of N1). Furthermore, to rule out the possibility that this pattern of results was an artifact of averaging across participants with different category boundaries, they also fitted and compared two mixed effects models: a linear and a categorical one, which took into account any differences between individual participants' category boundaries. In line with their prediction, they found that the linear model was a better fit of the data.

In addition to the N1, Toscano et al also looked at a different ERP component, the P3, which appears later (~300 ms – 800 ms post stimulus onset) and is thought to reflect categorization of speech stimuli (Maiste, Wiens, Hunt, Scherg, & Picton, 1995). In

accordance with their prediction, Toscano et al found evidence for gradiency in the P3 amplitude, which supports the view that within-category information is maintained in post-perceptual stages of processing.

This study offers a useful tool for studying whether individual differences in speech categorization tasks reflect differences in the early perception of acoustic cues. For example, if we found that individuals with higher levels of gradiency show a different relationship between VOT and N1 (or P3) amplitude, that would suggest that the sources of gradiency could be traced down to differences in the perceptual (or post-perceptual for the P3) encoding of acoustic information.

5.1.2 Lexical inhibition

Experiment 2 also examined an alternative hypothesis as to the potential locus of the individual differences we observed. That is, there is a possibility that gradiency at the level of phoneme categorization is in some way linked to (or maybe driven by) higher levels of spoken language processing, such as inter-lexical inhibition, which has been shown to occur during spoken word recognition.

There is now strong evidence that active words *suppress* their competitors during spoken word recognition (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Luce & Pisoni, 1998). In a sense, this helps the system "sharpen" decisions between words, committing more strongly to the target word over competing candidates. In addition, there is also evidence for *feedback* from the level of word forms to that of sublexical representations, such that information travels from higher to lower levels of processing (Elman & McClelland, 1988; Magnuson, McMurray, Tanenhaus, & Aslin, 2003). This top-down

flow of information has been shown to influence word recognition in real time (Magnuson et al., 2003; McClelland, Mirman, & Holt, 2006), but also drive perceptual learning (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Kraljic & Samuel, 2006; Leach & Samuel, 2007; Samuel, 2001; but see Norris et al., 2000). This could enable processes operating at the lexical level to influence lower level speech categorization.

Putting together these two ideas raises the possibility that sharpening at the lexical level (via local inhibition) may cascade to sharpen up categorization at lower (phonological) levels. Based on this rationale, our hypothesis was that stronger inter-lexical inhibition may lead to greater and/or faster suppression of competing lexical candidates, which may in turn lead to the target word exerting stronger feedback to the phoneme layer, and thus leading to greater and/or faster de-activation of competing phonological representations. For example, (adopting a localist framework for ease of description) if an ambiguous (þeach) item is heard, in a system with strong inter-lexical inhibition, the more active word (e.g., *beach*) would exert stronger inhibition on the less active item (*peach*), which would lead to the faster suppression of /p/. In contrast, a system with weaker inter-lexical inhibition dynamics would take more time to settle, thus allowing for longer-lasting parallel and somewhat gradient activation of more than one phoneme categories. This could be exclusively the result of real-time dynamics[8], or a combination of real-time dynamics and long-term learning in the system to allow for more or less gradient phoneme activations (i.e. a system with stronger inter-lexical

[8] Note that this could in principle be possible both for real word and nonword stimuli that partially overlap with real words (McClelland & Elman, 1986).

competition dynamics may with time be shaped to also have strong inter-phoneme competition).

To test this, we administered a task specifically designed to assess the degree of real-time interference between competing lexical items (Dahan et al., 2001; Kapnoula & McMurray, 2016a). This paradigm relies on an auditory stimulus manipulation in which two lexical items (e.g. *net* and *neck*) are cross-spliced such that the beginning of *net* is spliced onto the offset of *neck* to make *ne$_t$ck*. As a result, the coarticulatory information in (what is commonly described as) the vowel boosts activation of the competitor item (*net*), which, in turn, suppresses activation of the ultimate target (*neck*). Then later, when the final phoneme (/k/) is heard, the target (*neck*) may have a hard time being fully recognized (Dahan et al., 2001; Marslen-Wilson & Warren, 1994).

In the present study, this task was used to extract a measure of the overall strength of inter-lexical inhibition within a given individual. Even though this task has not been previously used as an individual differences measure, it has been used to compare between participants at the group level (Kapnoula & McMurray, 2016a). In addition, reliability testing of the VWP measures has shown moderate to high within-subject reliability for looks to the target (which is the measure used in this paradigm) with $R \approx .6$, (Farris-Trimble & McMurray, 2013). If individual differences in the degree of lexical inhibition, measured by this task, are found to be correlated with differences in phoneme categorization gradiency, measured by the VAS task, this would suggest a possible link between the two.

5.1.3 Alternative measures of inhibitory control

It is yet unclear what to make of the role of EF in phoneme categorization. N-Back was weakly but significantly correlated with Trail Making, but none of the other correlations among EF measures were significant, suggesting they may not form a homogenous constellation of skills. Furthermore, while we did observe a marginally significant correlation between N-Back (a measure of working memory) and gradiency, Kong and Edwards (submitted) did not find a correlation between their VAS-based measure and N-Back. Moreover, the opposite pattern was observed for the Trail Making task: Kong and Edwards found a significant correlation, but we did not. This suggests that correlations with EF ability, if present, are perhaps small and variable.

As a result it seemed prudent to continue our investigation of EF as a potential moderator. In this regard, inhibitory control seemed like the most important factor to consider. In part this is because if we found a correlation between inter-lexical inhibition and gradiency, it would be important to address whether this is attributed to broader inhibition-related mechanisms, or whether it is specific to lexical inhibition (but see dissociation between "automatic"/"obligatory" inhibition and attention-based inhibition in Burke and Shafto, 2008).

Thus, we also included a spatial Stroop task assessing top-down inhibitory control as an aspect of executive function. This form of inhibition is seen as theoretically distinct from lexical inhibition (as defined, for example, in TRACE; McClelland & Elman, 1986), though this has not been explicitly tested. Moreover, since we did not find a correlation between speech perception gradiency and inhibition measured by the Flanker task in

Experiment 1, we hypothesized that individuals' score in the Stroop task would not correlate with any of the other two tasks.

5.1.4 Within-category lexical level gradiency

Lastly, we were interested in testing whether gradiency at the phoneme level (as assessed by the VAS task) maps onto specifically within-category gradiency at the lexical level. One of the more robust demonstrations of this lexical level effect comes from McMurray, Tanenhaus, and Aslin (2002). They used the visual world paradigm (VWP) and showed that continuous differences in VOT lead to gradient activation of competing lexical items. To test this, they presented listeners with auditory items varying in equidistant steps between two endpoints (e.g., between *beach* and *peach*). Participants saw four pictures and clicked on the correct picture while their eye-movements were recorded. McMurray et al found that even when participants clicked on the target picture, they often looked to the competitor. Crucially, the likelihood of any participant fixating the competitor picture (when clicking on the target) was predicted by the VOT – when the VOT approached the category boundary (i.e. when the stimulus was more ambiguous), participants had a higher probability of looks to the competitor. This was observed even when the analysis was restricted to trials for which the auditory stimuli were all assigned to the same category, and when the VOT was treated as relative to the participants' own boundary.

This was interpreted as evidence that even as the system settles on a decision, competing representations remain active in a way that is consistent with a gradient pattern of lexical activation that reflects the probability of an auditory item being the

target (see also McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008). That is, even as listeners are likely to respond /b/ (for example), they keep /p/ active to varying degrees depending on the continuous cue value in the input (e.g., the VOT). Crucially this approach to analysis, by focusing on VOT relative to each participants' own boundary, focuses exclusively on *within-category* differences (ignoring differences in VOT that map to a different categories).

While this approach has not been examined in the context of individual differences, there has been work on populations with communication impairments such as adolescents with language impairment (McMurray et al., 2014) and cochlear implant users (Farris-Trimble, McMurray, Cigrand, & Tomblin, 2014). Thus, in Experiment 2, we included a task very similar to that used by McMurray et al. (2002) in order to examine whether speech perception gradiency, as assessed by a VAS task, is correlated with within-category gradient lexical activation. The intuitive prediction is that the two measures would be positively correlated, since phonological gradiency should be a prerequisite of lexical gradiency. However, it is also conceivable that we will not find a correlation, because the VWP task is assessing lexical gradiency in real time, whereas the VAS task gives us an estimate of phonological gradiency at the end of a categorization process. In other words, there is a possibility that all listeners are somewhat gradient (both at the phoneme and the lexical level), but some "lose" access to fine-grained, within-category information by the time they make a response.

We also need to consider that the VWP task measures lexical gradiency *within* each category, while our VAS-based measure reflects gradiency both within and across categories. Therefore, another possibility (not necessarily mutually exclusive with the

previous one) is that all listeners are to some degree sensitive to within-category differences, but for some listeners sensitivity to between-category differences may be disproportionally amplified. This could also explain why we find individual differences with the VAS task, while studies like that of McMurray et al. (2002) find robust evidence for graded lexical activation. In that case, we may find clear evidence for gradient activation of lexical representations across individuals using the VWP task, but no correlation with any individual differences found in regard to phoneme-level gradiency measured by the VAS task.

In sum, Experiment 2 aimed to investigate possible causal pathways linking phoneme categorization gradiency to 1) early perceptual encoding of speech cues (the ERP tasks), 2) top-down effects driven by inter-lexical inhibition (subphonemic mismatch task), and 3) broader inhibitory control mechanisms (spatial Stroop). In addition, we sought to examine the relationship between categorization gradiency at the phonological level with the within-category gradiency observed in the activation of lexical representations.

In addition to addressing these questions, Experiment 2 also included a visual version of the VAS task. The goal of this task was more methodological, to validate the VAS slope as a measure of specifically phoneme categorization gradiency. Since our VAS measure is extracted from participants' overt responses, it is conceivable that any differences are due to individuals' bias in how they use the continuous scale in the VAS task, and not due to underlying differences in regard to phoneme categorization. By including an analogous task with visual stimuli, we could measure this bias and partial it out of our measure of interest.

5.2 Methods

5.2.1 Participants

Seventy-one (71) monolingual American English speakers participated in Experiment 2. Participants were pre-screened to exclude individuals over 40 y.o., with any neurological disorders, and non-typical hearing, or vision. The age range of participants was 19-39 y.o. (M = 25.4, SD = 4.7) and thirty-three of them were male. Participants received monetary compensation for their participation in the study, and underwent informed consent in accord with University of Iowa IRB policies. Technical problems and experimental errors led to the results of different tasks not being available for one or more participants. As a result, between 2 and 8 participants were excluded from the analyses of the specific tasks for which there were missing data.

5.2.2 Design and tasks

Participants performed five tasks, four of which were designed to measure different aspects of spoken language processing (see Table 5.1 for the tasks and order). The visual analogue scaling task (VAS; Kong & Edwards, 2011; Munson & Carlson, in press; Schellinger, Edwards, Munson, & Beckman, 2008) measured speech categorization gradiency (see Section 2.1) using a similar VOT × F_0 continuum as in the prior experiments. Similarly, we developed a visual version of the VAS task (see Section 2.2) using a visual apple/pear continuum as a way of extracting a baseline of each participant's overall tendency to use the endpoints versus the whole range of the line (independently of phoneme categorization processes).

As a measure of inhibitory control independently of language, we administered a spatial version of the Stroop task (Jensen & Rohwer, 1966; Stroop, 1935; Wühr, 2007).

We also included two VWP tasks designed to tap different aspects of lexical processing. To estimate participants' degree of inter-lexical inhibition, we used a cross-splicing stimulus manipulation (the subphonemic mismatch paradigm) coupled with a VWP task specifically designed to tap this (Dahan et al., 2001; Kapnoula & McMurray, 2016a, 2016b; Kapnoula, Packard, Gupta, & McMurray, 2015). To assess gradiency at the lexical level, we administered a VWP task that has been shown a number of times to estimate the gradient relationship between within-category differences in cues like VOT and lexical activation (McMurray et al., 2002).

Lastly, to estimate individual differences in early auditory processing of speech sounds, we collected electrophysiological measures of participants' brain responses to stimuli with different VOTs using an ERP paradigm developed by Toscano et al. (2010).

Table 5.1 Order and description of tasks

Order	Task	Domain	Primarily measure of…
1	Phoneme VAS	Speech categorization	phoneme categorization gradiency
	Visual VAS	Visual categorization	task gradiency
2	Stroop task	Cognitive	executive function: inhibitory control
3	Lexical interference task (VWP)	Language processing	lexical interference
4	Within-category lexical gradiency task (VWP)	Language processing	lexical activation gradiency
5	Early auditory processing task (ERP)	Speech categorization	early and late encoding of speech cues

5.2.3 Phoneme and visual VAS tasks

5.2.3.1 Phoneme VAS design and materials. Similarly to Experiment 1, we used the VAS task to measure individual differences in gradiency. We used two places of articulation (labial: *bill-pill*, and alveolar: *den-ten*) and for each of the two sets we constructed a 7 VOT × 5 F_0 continuum. All stimuli were based on natural speech recordings spoken by a male monolingual speaker of American English.

We started by manipulating pitch. For each of these four recordings, we extracted the pitch contour and replaced it with a pitch contour of identical shape, but shifted either upwards or downwards so that the mean pitch would be either 95 Hz (for the low pitch condition) or 145 Hz (for the high pitch condition). This led to the construction of 8 new items (4 words × 2 F_0). Next, we modified the pitch contours using the pitch-synchronous overlap-add (PSOLA) algorithm in Praat. We first synced the recordings so that the pitch contours started at the same time, then extracted them into txt format, and used them to create three intermediate pitch contour steps. The five resulting pitch contours were approximately 12.5 Hz apart. These were used to replace the original contours of the eight items, giving us 20 new items (4 words × 5 F_0).

For the voicing manipulation, we used the progressive cross-splicing method described by Andruski, Blumstein, and Burton, (1994) and McMurray, Aslin, Tanenhaus, Spivey, and Subik, (2008), as described in Section 2.1.1. VOT steps varied from 7 ms to 43 ms and were 6 ms apart.

5.2.3.2 Phoneme VAS procedure. As in Experiment 1, participants saw a line at the two ends of which were two words. However, in contrast to Experiment 1, there was no rectangular bar in the middle of the line. This was changed to minimize participants'

possible inclination to drag the bar away from the center (or to leave it there). Instead, participants were asked to listen to each stimulus and then click on the line to indicate where they thought the stimulus they heard falls on the line. As soon as they clicked, the rectangular bar appeared at the point where they clicked and then they could either change their response (by clicking elsewhere on the line) or press the space bar to verify it. Unless the participant had clarifying questions, no further instructions were given. The task took approximately 15 mins to complete.

5.2.3.3 Visual VAS design and materials. The task and materials are described in detail in Section 2.2. The VAS procedure was similar to the phoneme VAS procedure (with no rectangular bar in the middle of the line). The task took approximately 10 mins to complete.

5.2.3.4 Order of VAS sets. The VAS tasks were conducted first on the first day of the experiment. The order of the three individual sets (labial, alveolar, or visual) was counterbalanced between participants with six different possible orders.

5.2.4 Spatial Stroop task

5.2.4.1 Spatial Stroop task design and materials. Participants performed the spatial Stroop task immediately after the VAS tasks. This task was based on the Stroop task (Jensen & Rohwer, 1966; Stroop, 1935), which has been broadly used as a neuropsychological assessment of executive function and inhibitory control (Shum, McFarland, & Bain, 1990).

In this experiment, the Stroop task was meant to serve as a measurement of executive function/inhibitory control outside the language domain. For this reason we

used a spatial variant of the Stroop task (Wühr, 2007) rather than the more common color/word version. In this task, participants saw an arrow, located on the left or right side of the screen and pointing to the left or right. They then responded as to which direction the arrow points to (suppressing the irrelevant cue, the side of the screen). It has been found that individuals respond faster and more accurately when the direction of the arrow is congruent with its location on the screen (i.e. congruent condition; Wühr, 2007). Therefore, performance in this task serves here as assessment of individual's ability to suppress irrelevant information (i.e. the location of the arrow).

To intensify the effect of incongruence, we used 64 congruent and 32 incongruent trials. Our materials consisted of two arrows 300 × 150 pixels in size, which were presented on a 19" monitor operating at 1280 × 1024 resolution, centered vertically, 100 pixels away from the corresponding edge of the display.

5.2.4.2 Spatial Stroop task procedure. At the beginning of each trial, a fixation point was presented at the center of the screen for 200 ms. Then the fixation disappeared and one arrow was presented on one side (left/right). The arrow stayed on the screen until the participant responded by pressing one of two keyboard keys (left/right) to report which direction the arrow was pointing to. After the response, there was a 1000 ms pause (blank screen), at the end of which the next trial began.

5.2.5 Lexical inhibition task

5.2.5.1 Overview and design. This task was designed to give us an estimate of the strength of lexical inhibition for each participant. Following previous experiments (Dahan et al., 2001; Kapnoula & McMurray, 2016a; Marslen-Wilson & Warren, 1994;

Streeter & Nigro, 1979; Whalen, 1984), we created auditory stimuli which were manipulated so that the onset of each word was either consistent (matching) or temporarily boosted activation for a competitor (to observe the inhibition effect on the target).

Each target word (e.g., *net*) appeared in three experimental conditions, corresponding to the three types of splicing. In the *matching-splice* condition, both the onset and release burst of the auditory item came from the same word (ne_tt), though from different recordings. This would be expected to lead to rapid activation of the correct word. In the *word-splice* condition, the onset of a competing word (e.g. *ne-* from *neck*) was spliced onto the release burst of the target word (*net*). The result of this is that the competitor word (*neck*) would be briefly over-activated, and temporarily inhibit the target (*net*); then, when the release burst arrived, it would be much more difficult to fully activate the target (due to its prior inhibition). To ensure that any inhibition effects in the *word-splice* condition were not simply due to the fact that the cross-spliced stimuli are poorer exemplars of the target word, we also included a control condition, the *nonword-splice* condition (ne_pt), in which the onset of the stimulus was taken from a nonword. In this case, a bottom-up mismatch is still present, but the onset does not activate a competing word.

We used the visual world paradigm (VWP) to measure the activation of the target word at each point in time for each of the three splicing conditions. In this task, participants saw four pictures, a picture of the target (*net*) along with three semantically unrelated words, and heard a word that was the label of one of the four pictures. During the task, participants heard each of the four words of a set in all three splice conditions.

Therefore, each set of four pictures was presented 12 times (4 pictures × 3 splice conditions), and each picture in a set had an equal probability of being the referent of the auditory stimulus.

Previous studies have found that listeners look less to the picture of the target word in the mismatching (*ne$_{ck}$t*) compared to the matching (*ne$_{t}$t*) condition, (Dahan et al., 2001; Kapnoula & McMurray, 2016a, 2016b; Kapnoula et al., 2015). We expected to replicate this finding. Crucially, we also hypothesized that the strength and/or timing of this effect may be linked to our measure of gradiency (i.e. VAS slope). Specifically, as discussed above, we predicted that individuals with steeper VAS slope (i.e. showing less gradiency) would show a stronger lexical interference effect. The rationale behind this prediction was that stronger lateral inhibition at the lexical level may lead to sharper lexical activations (i.e. stronger activation of the target word and stronger suppression of the competitor), which in turn may lead to sharper activation of phonemes.

5.2.5.2 Stimuli. Our stimuli consisted of 28 pairs of words: a target (e.g. *net*) and its competitor (e.g. *neck*; for a list of the word pairs, see Table A.1 in Appendix). Only the target was ever heard during the experiment, while its competitor was never played and no picture of it was ever shown. For each target, we chose three semantically unrelated words. One of these words had an initial-phoneme overlap with the target word.

A total of 112 pictures (28 target words × 4 pictures in each set; for a list of the visual stimuli, see Table A.2 in Appendix) were used, all of which were developed using a standard lab procedure (Apfelbaum, Blumstein, & McMurray, 2011; McMurray et al., 2010). For each word, we downloaded 5-10 candidate images from a commercial clipart database, which were viewed by a group of 3-5 undergraduate and graduate lab members.

One image was selected and was subsequently edited to ensure a prototypical depiction of the target word. The final images were approved by a lab member with extensive experience using this paradigm.

To construct the auditory stimuli, we spliced the offset of each target word onto three different onsets. We took the release burst from each target word (e.g. *net*), starting at the onset of the release burst and until the end of the word, and spliced it onto the onset portion[9] of 1) another recording of the target (*ne$_t$t*), 2) its competitor (*ne$_{ck}$t*), and 3) the nonword (*ne$_p$t*; see Table A.3 in Appendix for a full list of the spliced stimuli). Raw stimuli were recorded by a male native speaker of American English in a sound attenuated room at 44,100 Hz. The splice point was always at the zero crossing closest to the onset of the release. We also created three spliced versions of each of the filler words with each target word spliced with itself and one of two nonwords.

5.2.5.3 Procedure. Participants were first familiarized with the 112 pictures by seeing each picture along with its orthographic label. Then they were fitted with an SR Research Eyelink II head mounted eye-tracker. After calibration, participants were given instructions for the task.

At the beginning of each trial, participants saw four pictures in the four corners of a 19" monitor operating at 1280 × 1204 resolution. In addition, a red circle was present at the center of the screen. After 500 ms, the circle turned blue, which prompted the participant to click on the circle to start the trial. This delay allowed participants to take a brief look at the pictures before hearing anything, thus minimizing eye movements due to

[9] This onset portion was taken from the beginning of each recording and included everything up to the onset of the release, including the closure.

visual search (rather than lexical processing). Once participants clicked on the blue circle, it disappeared and an auditory stimulus was played. Participants then clicked on the picture that matched the word they heard, and the trial ended. There was no time limit and participants were encouraged to take their time and perform the task as naturally as possible. They typically responded in less than 2 sec (M = 1216 ms, SD = 109.37 ms).

5.2.5.4 Eye-tracking recording and analysis. We recorded eye-movements at 250 Hz using an SR Research Eyelink II head-mounted eye-tracker. Both corneal reflection and pupil were used whenever possible. Participants were calibrated using the standard 9-point Eyelink procedure. The Eyelink II yields a real-time record of gaze in screen coordinates while compensating for head-movements. This 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., 2010; McMurray, Tanenhaus, & Aslin, 2002).

Eye-movements were recorded from the onset of the trial (the blue circle) to the participants' response (mouse click). This variable-time offset makes it difficult to analyze results late in the time course. To address this issue, we adopted the approach of many prior studies (Allopenna, Magnuson, & Tanenhaus, 1998; McMurray et al., 2002) by setting a fixed trial duration of 2000 ms (relative to stimulus onset). For trials that ended before this point, we extended the last eye-movement; trials which were longer than 2000 ms 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 an approximation of the final state of the system and not necessarily what the participant

was fixating at a particular point in time. The coordinates of each look were used to obtain information about which object was being fixated.

For assigning fixations to objects, boundaries around the objects were extended by 100 pixels in order to account for noise and/or head-drift in the eye-track record. However, this did not result in any overlap between the objects; the neutral space between pictures was 124 pixels vertically and 380 pixels horizontally.

5.2.6 Within-category lexical gradiency task

5.2.6.1 Overview and design. This task was designed to measure the degree to which participants use fine-grained acoustic information to activate lexical representations in a gradient manner. It specifically targets within-category sensitivity – that is sensitivity to gradient changes in the acoustic input which do not affect the final response. The design and stimulus manipulation of this task were based on the design of McMurray et al. (2002). Specifically, we manipulated the VOT of the initial consonant in minimal pairs of words, such as *bear-pear*, to create a continuum between them. Fixations to each picture were computed as a function of VOT. Critically, these data were split by participants' identification responses (e.g. whether they clicked on the picture of the *bear* or the *pear*). This allows us to compute a measure of how strongly *within-category* changes in VOT are reflected in lexical activation (since the analysis is predicated only on trials where the participant chose the same picture while examining their fixations to both pictures). In line with the findings of McMurray et al. (2002), we expected to find that, even when participants clicked on the picture of the voiced item (*bear*), their looks to the competitor (*pear*) would increase as the distance from the target

(in VOT steps) increases. Lastly, in addition to the VOT manipulation, as was in the original McMurray et al. (2002) design, we also manipulated F_0.

Our hypothesis was that gradient activation at the phoneme level (measured by the phoneme VAS task) should be related to gradient activation of lexical representations. Therefore, we expected to find a positive correlation between VAS slope and a lexical gradiency measure extracted from the lexical gradiency task.

5.2.6.2 Stimuli. Stimuli consisted of 10 monosyllable CVC word pairs beginning with a stop consonant; 5 with a labial and 5 with an alveolar onset. The two words in each pair were identical except for the voicing of the initial stop consonant (e.g. *bear-pear*). Each labial-initial pair was paired with an alveolar-initial pair, making a quadruplet (e.g. *bath-path, deer-tear*; see Table 5.2 for a list of the stimuli pairings). The four images corresponding to the each of the items in a quadruplet were presented together throughout the task and across participants.

Table 5.2 List of stimuli presented in the within-category lexical gradiency task

Set no	Labials		Alveolars	
	Word 1	Word 2	Word 1	Word 2
1	bath	path	deer	tear
2	beach	peach	drain	train
3	bear	pear	dot	tot
4	bees	peas	dent	tent
5	bowl	pole	dart	tart

For each of the 10 minimal pairs, we constructed our ten 7 VOT × 2 F_0 continua from natural speech. First, we manipulated pitch. For each recording, we extracted the

pitch contour and replaced it with a pitch contour of identical shape, but shifted either upwards or downwards so that the mean pitch would now be either 95 Hz or 145 Hz. This led to the construction of 40 endpoints (10 pairs × 2 endpoints × 2 F_0 values). Then, for the voicing manipulation, we used the two items in each pair with the same mean pitch (one with a voiced and one with a voiceless initial consonant) and followed the progressive cross-splicing method described by Andruski, Blumstein, and Burton (1994) and McMurray, Aslin, Tanenhaus, Spivey, and Subik (2008). That is, as in the stimuli used in the VAS tasks, progressively longer portions of the onset of a voiced sound (e.g., /b/) were replaced with analogous amounts taken from the aspirated period of the corresponding voiceless sound (e.g., /p/). This procedure resulted in the construction of seven equidistant (6 ms apart; 7-43 ms) VOT steps for each of the 20 continua (10 pairs × 2 F_0), resulting in 140 auditory stimuli (10 pairs × 2 F_0 × 7 VOT).

For each of the 20 words, visual stimuli (referents) were developed using a standard lab procedure (Apfelbaum et al., 2011 and McMurray et al., 2010). For each word, a set of 5–10 candidate images were downloaded from a commercial clipart database and viewed by a small focus group of 3–5 undergraduate and graduate students. Then one image was selected as the most prototypical exemplar of that word. These images were edited to remove extraneous elements, adjust colors, and ensure a clear and prototypical depiction of the intended word. The final images were approved by a lab member with extensive experience using the VWP.

5.2.6.3 Procedure. Participants were first familiarized with the pictures by seeing each picture along with its orthographic label. Then they were fitted with an SR Research

Eyelink II head mounted eye-tracker. After calibration, participants were given instructions for the task.

At the beginning of each trial, four pictures (corresponding to a quadruplet set) were presented in the four corners of a 19" monitor operating at 1280 × 1204 resolution. At the same time, a small red circle appeared at the center of the screen. After 500 ms, the circle turned blue, cueing the participant to click on it to start the trial. This allowed the participants to briefly look at the pictures before hearing anything, thus minimizing eye-movements due to visual search (rather than lexical processing). When participants clicked on the circle, it disappeared and an auditory stimulus corresponding to one of the four words was played. Participants clicked on the picture corresponding to the word and the trial ended. There was no time limit on the trials, and participants were not encouraged to respond quickly. Participants typically responded in less than 2 s (*M* = 1038.11 ms, *SD* = 104.92 ms).

5.2.6.4 Eye-tracking recording and analysis. The eye-tracking recording and analyses procedures were identical to those described earlier for the lexical interference task (see Section *5.2.5.4*).

5.2.7 Early auditory processing (ERP) task

The purpose of this task was to evaluate whether there were any significant differences between participants' early brain responses to continuous acoustic differences in the speech signal (specifically, differences in VOT). In order to assess participants' early perceptual encoding of acoustic information, we used an ERP paradigm, which has been shown to be sensitive to fine-grained manipulations of acoustic cues such as VOT

(Toscano et al., 2010). In this paradigm, early encoding of VOT is thought to be reflected in the amplitude of the N1 ERP component, which is triggered by the onset of a sound.

Following Toscano et al, we were also interested in a second ERP component, the P3, which appears later than the N1 and is thought to be more susceptible to category-related information. Traditionally, the P3 is elicited in an "oddball" task, in which participants respond to an infrequent target (Polich & Criado, 2006). Thus, we designed our ERP task so that participants would have to respond as to whether they heard a specific target word (e.g., *bill*) or any of the other words (*pill*, *den*, *ten*). Consequently, participants were expected to make a "target" response (e.g. *bill*) approximately 25% of the time, and a "non-target"/"other" (*pill, den, ten*) response about 75% of the time. This, thus, fulfills the requirement of having trials with infrequent targets. Each of the four words served as the target on different blocks of trials.

5.2.7.1 Design and stimuli. The auditory stimuli came from the same natural recordings as those used in the VAS task; the endpoints were those used in the VAS task (VOT of 7 and 43 ms), however, for the ERP task, we constructed 9 (instead of 7) VOT steps, 4.5 ms apart, and only used the two extreme F_0 values.

Each stimulus was presented in one of four conditions, in which one of the four words (*bill*, *pill*, *den*, or *ten*) was the target. This was important for the elicitation of the P3, which, as we described above, is contingent upon the presence of infrequent targets. In this case, we expected that about half of the 9 steps in each continuum would be classified as one word in the pair, and half as the other. This means that when, for example, *bill* was the target, ~50% of the labial-initial stimuli would be classified as targets, while none of the alveolars would. As a result, across stimuli, participants were

expected to make a "target" response about 25% of the time. Lastly, we also manipulated the location of the target button (left or right; see Figure 5.2).

Each stimulus × target word × target location combination was repeated 7 times. Therefore, each of the 36 (2 PoA × 9 VOT steps × 2 F_0 steps) auditory stimuli were presented 56 times (4 target words × 2 target locations × 7 repetitions), giving us a total of 2016 trials. The trials were split into 8 blocks of 252 trials each, one block for each of the 4 target × 2 target location conditions, and the order of each block was pseudorandomized and kept the same for all participants.

5.2.7.2 Procedure. Participants performed the ERP task on the second day of the experiment. First, the EEG recording equipment was set-up and participants were seated inside a grounded and electrically-shielded booth. Next, electrode impedances were minimized, and the earphones were inserted. Preparation took approximately 30 minutes.

At the beginning of the task, participants read the instructions and performed a few trials to familiarize themselves with the task, while the experimenter remained outside the booth and monitored their responses to ensure they performed the task as instructed. After practice, if they had no questions, they started the task.

Auditory stimuli were presented over earphones (ER3-14 by Etymotic Research) connected to an amplifier located outside the booth. Instructions and visual stimuli were presented on a computer monitor located approximately 75 cm in front of the participant. Instructions, stimulus presentation, and sending of event codes to the EEG amplifier were handled by Presentation (by Neurobehavioral Systems). Participants responded using one of two buttons on a Play Station gamepad (L1 and R1).

At the beginning of each trial, participants saw a black fixation cross at the center of the screen and heard a word over the earphones. After the word was played, the cross was replaced with a green circle and two words, one each side of the circle, indicating which button corresponded to which response. One word was always the target for that block (e.g. *bill*) and the other was the word *other*. The participant had 2000 ms to make a response (by pressing one of the two buttons) and the trial ended. Detailed information about the timing of the events within a trial is shown in Figure 5.2.

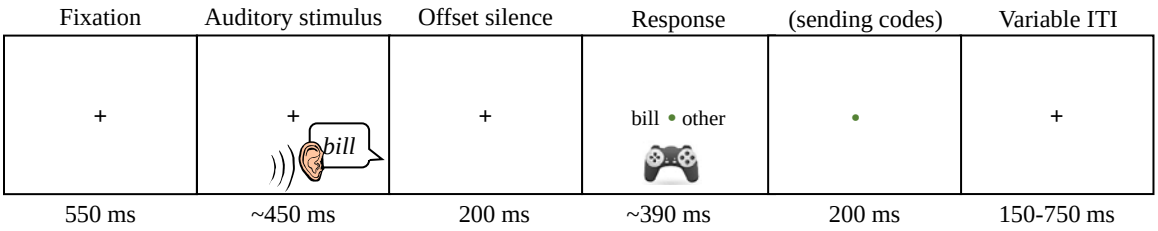


Figure 5.2 Structure of single trial of the ERP task

As shown in Figure 5.2, average total trial duration (including RT) was ~2240 ms. With 2016 total trials, the task took approximately 75 minutes. Participants were given an opportunity for a break every 36 trials and were encouraged to take a break and ask for water half-way through the experiment. They usually completed the task within 90 minutes.

5.2.7.2 EEG recording. ERPs were recorded from 32 electrode sites (International 10-20 System sites Fp1, Fz, F3, F7, T7, C3, Cz, P3, P7, O1, Fp2, F4, F8, T8, C4, Pz, P4, P8, Oz, O2, FT9, FC5, FC1, T9, CP5, CP1, TP10, CP6, CP2, FT10, FC6, FC2). EEG channels were collected using the reference-free acquisition provided by Brain Products actiCHamp and were referenced to the average of the two mastoids after recording. Horizontal electrooculogram (EOG) recordings were collected via two electrodes located

approximately 1 cm lateral to the external canthus of each eye. Vertical EOG recordings were made using an electrode located approximately 1 cm below the lower eyelid of the left eye. Recordings were made with the Brain Products actiCHamp amplifier system at a sampling rate of 500 Hz. Reception and storing of the recordings, as well as linking them to the event codes sent by Presentation were handled by Brain Vision PyCorder. No filter was applied during recording.

5.2.7.3 EEG data pre-processing. Data were analyzed using Brain Vision Analyzer 2. A 1 Hz 48 dB/octave low cut-off filter, a 30 Hz 48 dB/octave high cut-off filter, and a 60 Hz notch filter were applied to the data prior to processing. We evaluated different artifact rejection procedures to remove eye blinks and the most efficient one was the following: segments 400 ms before and 900 ms after the stimulus onset were checked. For each one, if voltage shifted by more than 50 μV/ms (in either direction), or if voltage shifted by more than 75 μV (in either direction) within any 100 ms part of that segment, then that part (as well as 100 ms before and 100 ms after) was marked as bad. This was applied to the three EOG channels (VEOG, REOG, and LEOG) and any segments containing marked portions were excluded from further processing.

Next, we evaluated different artifact rejection procedures to remove other artifacts (e.g. due to movement, muscle tension, or sweat) and the most efficient one was the following: segments 300 ms before and 800 ms after the stimulus onset were checked. For each one, if voltage shifted by more than 50 μV/ms (in either direction), then a marker was placed at the time of the voltage shift and a portion of the segment (200 ms before the marker to 200 ms after that marker) was marked as bad. If voltage shifted by more than 75 μV (in either direction) within any 100 ms portion of that segment, then

that part (as well as 100 ms before and 100 ms after) was marked as bad. Lastly, if amplitude was higher than 150 μV or lower than -150 μV, then a marker was placed at the time of the voltage divergence and a portion of the segment (200 ms before the marker to 200 ms after that marker) was marked as bad. This was applied to all remaining channels. In addition, we used the "individual channel mode" option, which allows us to exclude segments for specific channels. On average 7.3% of the trials (i.e. 15 trials) were rejected for each participant (3.9% were blink removals and 3.4% other artifacts).

Each trial was baselined using as a baseline the average voltage within a time window starting 100 ms before the onset of the auditory stimulus up until its onset.

5.3 Results

We first report our findings relating to the nature of categorization gradiency at the phoneme level (VAS task) and at the lexical level (within-category lexical gradiency task). Then we move on the possible sources of gradiency (1) at higher-level cognitive functions, that are not language-specific (i.e. inhibitory control assessed via the spatial Stroop task), (2) at higher-level processes, but within the language system (i.e. lexical inhibition assessed via the lexical inhibition task), and (3) at low-level perceptual processes (i.e. early perceptual encoding of acoustic cue information assessed by the early auditory processing task).

5.3.1 Phoneme categorization gradiency and secondary cue use

As in the prior experiments, participants' responses in the phoneme and visual VAS tasks were fitted with the rotated logistic function (see Section 2.1). Fits were good

(R^2 = .96 and R^2 = .97 respectively[10]). Based on these fits, we extracted the estimated degree of phoneme and visual gradiency separately for each participant.

5.3.1.1 Phoneme and visual categorization gradiency. We started by examining the relationship between different measures of phonological and visual categorization gradiency. Specifically, we were interested in whether and to what degree our measures of phoneme categorization gradiency extracted from two different types of phoneme contrasts (labial and alveolar) may be related to each other and to the corresponding measure extracted from the newly-added visual VAS task.

In line with the results of Experiment 1, we found that VAS slope for labial stimuli was significantly correlated with that of alveolar stimuli (r = .407, p = .001). Visual VAS slope, on the other hand, was marginally significantly correlated with labial VAS slope (r = .208, p = .089) and was not significantly correlated with alveolar VAS slope, (r = .164, p = .182). These results suggest that participants' bias towards using the whole range of the VAS line (versus using mainly the endpoints) is not a major factor behind our VAS-based measure of phoneme categorization gradiency. However, given the one marginal correlation, we decided to compute residualized phoneme VAS slopes (by extracting the standardized residual of the phoneme VAS slope variance after partialing out the variance explained by the visual VAS slope) and include them in our analyses of the VAS slope as well.

5.3.1.2 Phoneme categorization gradiency and use of secondary cues. Next we moved on to assessing the relationship between phoneme categorization gradiency and multiple cue integration. Based on previous findings (Kong & Edwards, submitted;

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

Experiment 1 in this study), we expected to find a significant correlation between the two. While this experiment did not include an independent measure of secondary cue (pitch) use, we were able to extract a measure of this from the VAS task, the θ angle. As demonstrated by our Monte Carlo analyses in Section 2.1.4 (see Figure 2.3), this parameter does not inherently correlate with the VAS slope (when there is no underlying correlation), and in Chapter 3 we showed that it is well correlated with 2AFC measures. Thus, it offers a measure of multiple cue integration that is independent of slope permitting a partial replication of Experiment 1b.

To assess the degree to which gradiency is linked to multiple cue integration (as reported in Experiment 1), we followed a hierarchical regression approach with VAS slope and residualized VAS slope as the dependent variables. The independent variables were place of articulation (PoA; effect-coded) and secondary cue use (θ angle). In the first level of the model, with VAS slope as the dependent variable, PoA was entered as a predictor and significantly accounted for 10.7% of the variance, $\beta = -.328$, $F(1,131) = 15.76$, $p < .001$) with higher VAS slopes (lower gradiency) observed for labial-initial stimuli. However, PoA was not a significant predictor of residualized VAS slope, accounting for less than 1% of the variance, $\beta = -.023$, $F < 1$. In the second step, secondary cue use (i.e. theta angle) was added as a predictor. This explained a significant portion of the VAS slope variance, $\beta = .273$; $R^2_{change} = .045$, $F_{change}(1,130) = 6.92$, $p < .01$, and the same was found for the residualized VAS slope, $\beta = .314$; $R^2_{change} = .060$, $F_{change}(1,130) = 8.25$, $p < .01$. This suggests that higher theta angle (i.e. weaker use of pitch) predicts steeper VAS slope (i.e. less gradiency), in line with the results of Experiment 1.

The results from the VAS task are overall in line with both the Kong and Edwards (submitted) findings and our results from Experiment 1, showing that individuals with greater categorization gradiency seem to make greater use of a secondary cue. This strengthens the hypothesis that there is a link between the two that needs to be described in mechanistic terms.

5.3.2 Phoneme categorization gradiency and lexical gradiency

5.3.2.1 Analyses of responses. Participants performed this task without difficulties and responded in a prompt manner (*Mean RT* = 1038.11 ms, *SD* = 104.92 ms). Due to problems with the eye-tracking, eight participants were excluded from the analyses of fixations (but were included in the analyses of responses).

We first looked at the likelihood of participants clicking on the picture corresponding to the unvoiced word in the pair (henceforth: likelihood of unvoiced response or LUR).

We fitted a logistic mixed effects model implemented using the *glmer* command in the *lme4* package (Kuznetsova, Brockhoff, & Christensen, 2013), with VOT, F_0, and PoA, as well as their interactions as fixed effects. VOT and F_0 were centered and PoA was effect-coded (alveolar = 1; labial = -1). The random effects included a random slope of VOT, PoA, and their interaction for subjects.

We found a significant main effect of VOT, $z = 35.16$, $p < .001$, with greater VOT predicting higher LUR (as expected). In addition, there was a main effect of F_0, $z = 4.32$, $p < .001$, in the expected direction; stimuli with higher pitch were more likely to be classified as unvoiced. There was also a main effect of PoA, $z = -15.93$, $p < .001$,

showing that participants were more likely to give an unvoiced response for labial, compared to alveolar, stimuli. Finally, the VOT × F_0 interaction was also significant, z = -4.74, p < .001, as was the three-way interaction, z = 3.69, p < .001, which is consistent with the stronger effect of F_0 seen for labial stimuli with low VOTs (see Figure 5.3.A).

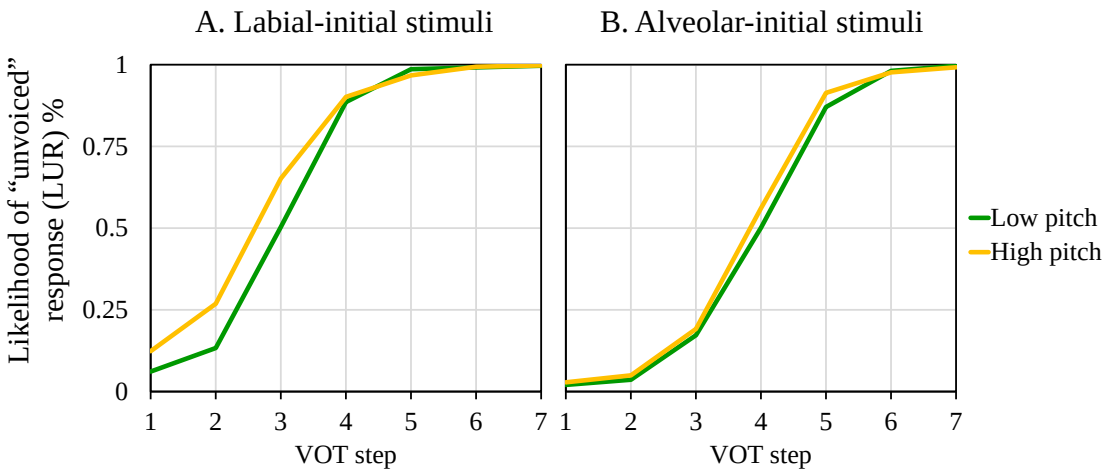


Figure 5.3 Likelihood of "unvoiced" response per VOT/F_0 step

Overall, these analyses revealed significant effects of our secondary variables (PoA and F_0) and for this reason we decided to not collapse across them in our main analyses.

5.3.2.2 Analyses of fixations. Figure 5.4 shows the likelihood of fixating the target object, the competitor, and the unrelated fillers as a function of time in two cases: when the stimulus was unambiguous (i.e. VOT step = 1/7; Figure 5.4.A) and when it was ambiguous (i.e. VOT step = 4; Figure 5.4.B). What can be seen is that in both cases participants overall looked substantially more to the target, but they also showed more fixations to the picture of the competitor compared to that of the filler items. In addition, when the auditory stimulus was ambiguous, participants looked overall less to the target

and more to the competitor. This suggests that the acoustic difference between the stimulus and the target may have had a significant effect on participants' fixations.

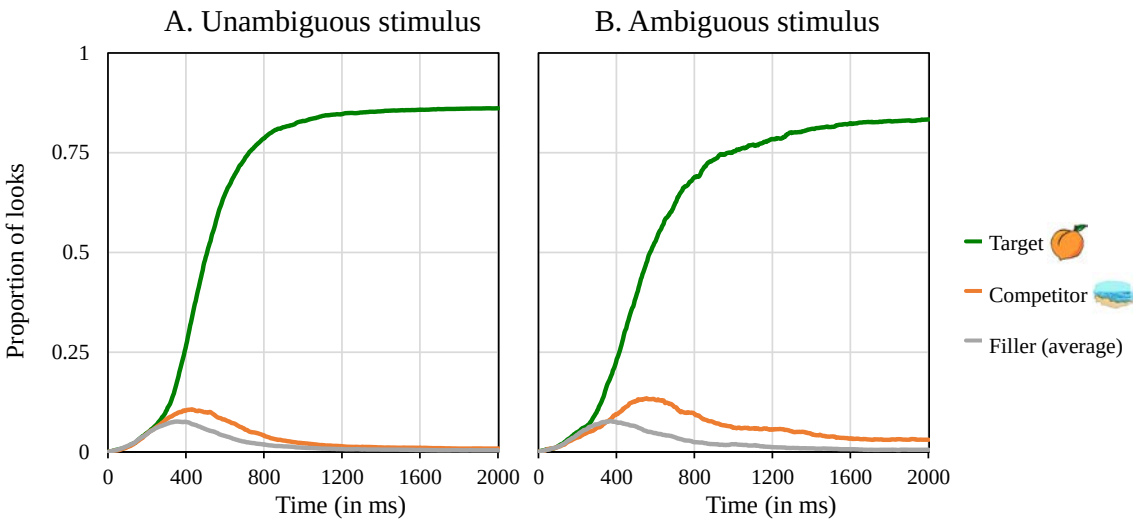


Figure 5.4 Proportions of looks to the picture of the target, the competitor, and the filler when participants clicked on the target

Note: target is defined as the object that the participant clicked on

We then examined the effect of our stimulus manipulations on participants' looks to the competitor. Because we are interested in individual differences, we were concerned that a pure use of raw competitor fixations may confound differences in lexical / phonological processes with differences in overall looking. Thus, we calculated the proportion of looks to the competitor item (i.e. the picture of *peach* when participants clicked on that of *beach* and vice versa), as well as to the filler items, and used their difference as our dependent variable (henceforth *Comp-Filler*). This difference-based measure allows us to evaluate the degree to which participants kept the competitor activated independently of any individual differences in the overall number of eye-movements to the pictures.

Crucially, in this task we were interested in how *within-category* differences affect lexical activation. However, variation between listeners' category boundaries could complicate things, because a given difference between two adjacent VOT steps may be a within-category difference for one participant, but a between-category difference for another. To avoid this problem, we first computed the category boundary (crossover) separately for each participant. To do so, we first fitted each participant's response function using a four-parameter logistic curve-fitting procedure (see Eq.4) and used the crossover parameter (*co*) as an estimate of category boundary. We did this for each place of articulation and pitch value separately, which yielded four different crossovers for each participant. There were insufficient data in this shortened version of the paradigm to do this within each continuum within each subject (e.g., per condition / participant). Thus, we did the same for each continuum (collapsed over subject), and adjusted the participant's crossovers by subtracting the deviation of each stimulus crossover from the average stimulus crossover. In other words, for each participant we computed a VOT category boundary adjusted for the effects of place of articulation, F_0, and item (see McMurray, Farris-Trimble, Seedorff, & Rigler, 2016, for a similar procedure).

Having established an estimate of the category boundary for each participant × stimulus combination, we then calculated the distance between this crossover and the actual VOT step, (henceforth, *relative VOT* or *rVOT*; see also McMurray et al., 2008). For example, in a case where the crossover was 4.3, a stimulus with a VOT of 6 would be considered "unvoiced" with an rVOT of 1.7, whereas a stimulus with a VOT of 2 would be classified as "voiced" and have an rVOT of -2.3.

For our final analysis, we considered only trials in which the participant's response matched the side of their boundary (e.g., for an rVOT of -1, they had to have selected the voiced response; for an rVOT of +2, the unvoiced). The result of this procedure meant that differences in eye-movements reflected truly within-category sensitivity, since they came from trials corresponding to the same phoneme category, and from stimuli which were precisely located on one side of the participants' observed category boundary.

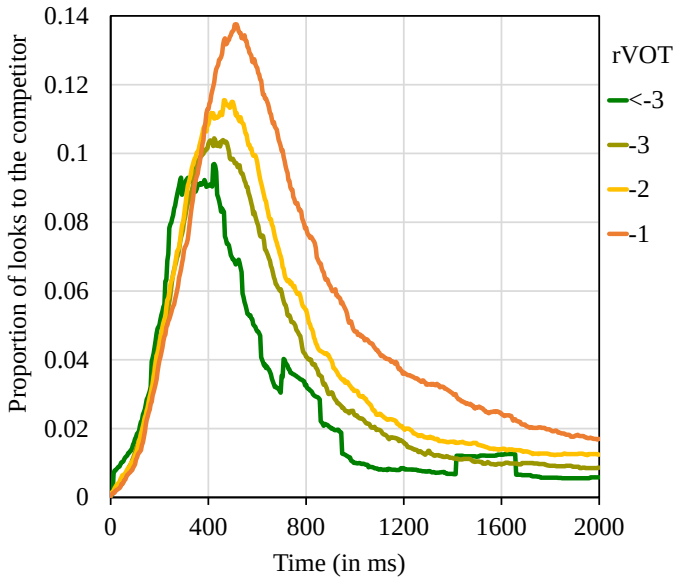


Figure 5.5 Looks to competitor as a function of distance from the crossover

As seen in Figures 5.5 and 5.6, even when participants clicked on one picture (e.g. the picture of *beach*), they still looked to the picture of its competitor (*peach*) and the proportion of trials seems to increase as the relative VOT approached the crossover. In addition, participants seem to look more to the competitor at the beginning of the trial, but the effect of rVOT seems to hold across the duration of the trial. Thus, for our analyses, we split the data in two separate time windows: an early one (300 ms – 1000

ms) and a late one (1000 – 2000 ms). The onset of the early time window was 300 ms to adjust for the 100 ms of silence at the beginning of the auditory stimulus plus the 200 ms oculomotor delay needed to plan an eye-movement. The end of this time window was chosen based on the average RT, which was 1038 ms. In that sense, the late time window was chosen to reflect the end-state of the system (i.e. the status of lexical activations after the system had settled).

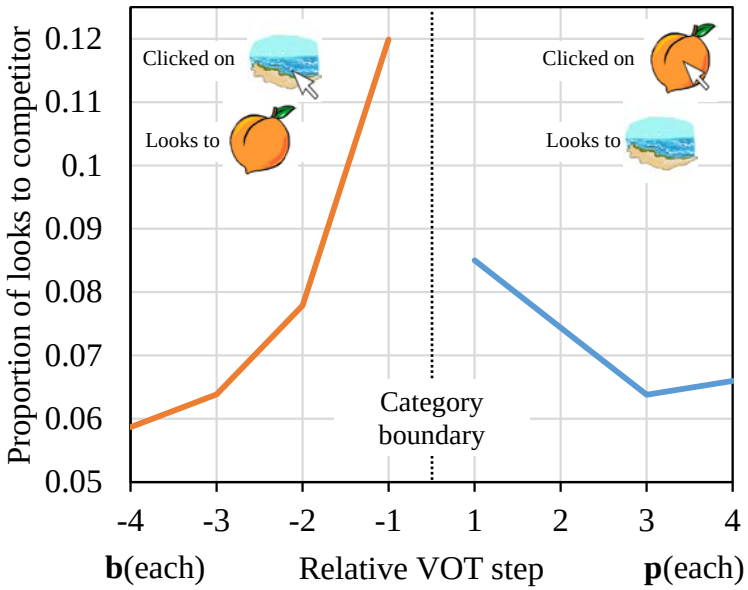


Figure 5.6 Proportions of looks to the competitor when participants clicked on the target per rounded relative VOT step (time window: 300-1000 ms)

To assess statistically the effect of rVOT on our Comp-Filler measure, we fitted four linear mixed effects models implemented with *lme4* (version 1.1-6; Bates, Maechler, & Dai, 2009), and *lmerTest* (version 2.0-6; Kuznetsova, Brockhoff, & Christensen, 2013) packages in R (R Development Core Team, 2009). A separate model was fitted for each combination of time window (early versus late) × side of the continuum (voiced versus

unvoiced). The only fixed effect in the models was that of rVOT (the effects of PoA and F_0 were already incorporated in the computation of rVOT, so we collapsed over these factors to simplify the analysis). The random effects structure included random intercept and random rVOT slopes for subjects and items.

In the early time window, we found a significant main effect of rVOT for both voiced-onset, B = .019, t(12.4) = 3.77, p < .01, and unvoiced-onset stimuli, B = -.008, t(9.82) = -3.49, p < .01. In the late time window, rVOT was marginally significant for labial-onset stimuli, B = .005, t(10.83) = 1.87, p = .089, and significant for unvoiced-onset stimuli, B = -.004, t(10.83) = -3.07, p < .05. As expected, these results show an overall strong effect of rVOT, which is more robust for the early time window (i.e. before the system settles).

Next we turned to our primary question; whether gradiency at the phoneme level is linked to gradiency at the lexical level. To evaluate this, we added VAS slope (or residualized VAS slope, in a separate set of four models), as well as their interactions with rVOT to the fixed effects in the aforementioned models. All continuous measures were centered.

In the early time window, for the voiced-onset stimuli, we found no main effect of VAS slope, B = -.026, t(60.72) = -1.61, p = .114, or residualized VAS slope, B = -.006, t(60.21) = -1.18, p = .24. However, there was a marginally significant rVOT × VAS slope interaction, B = -.02, t(75.69) = -1.76, p = .083, consistent with a stronger rVOT effect for shallower categorizers (i.e. participants with higher gradiency). In contrast, for unvoiced-onset stimuli in the same time window, neither VAS slope, t < 1, nor residualized VAS slope, t < 1, had a significant effect and none of the interactions were

significant. Lastly, in the late time window, none of the main effects of VAS slope or residualized VAS slope or their interactions with rVOT were significant for either the voiced or unvoiced models.

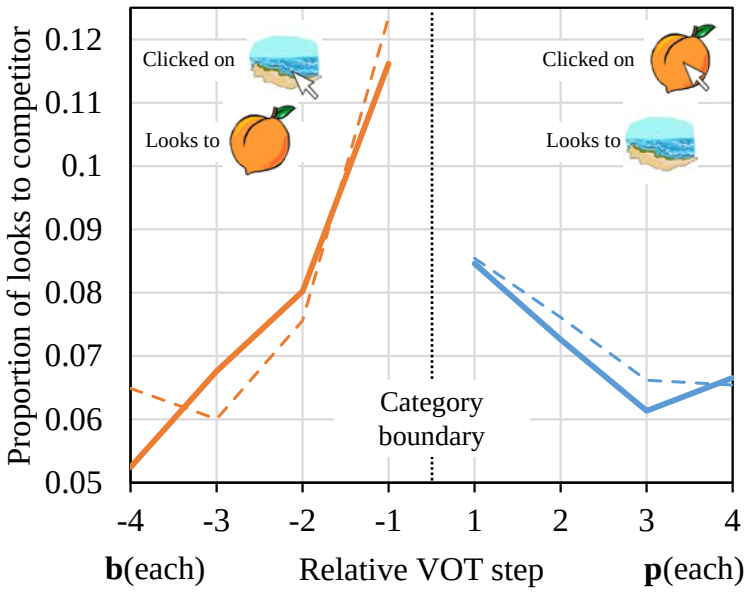


Figure 5.7 Proportions of looks to the competitor for high gradiency group (dotted lines) and low gradiency group (solid lines) when participants clicked on the target per rounded relative VOT step (time window: 300-1000 ms)

In sum, these analyses show that there was a strong effect of rVOT across participants, responses, and time windows, meaning that, as expected, greater distance from the target led to more looks to the picture of the competitor. As to our main question, none of the models showed a robust main effect of any measure of gradiency. That said, we did find a marginally significant VAS slope × rVOT interaction in the early time window for voiced-onset stimuli, suggesting that participants with higher levels of gradiency may use cue-level information to a higher degree when it comes to activating lexical items. However, this was not a particularly robust effect.

5.3.2.5 Phoneme categorization gradiency and lexical gradiency: Summary. Overall, our results seem to suggest that gradient activation of lexical representations is a fundamental aspect of spoken word recognition that is not strongly modulated by differences in sublexical categorization. However, there seems to be a weak link in the expected direction, showing a somewhat stronger effect of rVOT for individuals with higher gradiency VAS scores.

5.3.3 Phoneme categorization gradiency and inhibitory control (spatial Stroop task)

We next investigated the relationship between phoneme categorization gradiency and inhibitory control measured with the spatial Stroop task.

5.3.3.1 Accuracy and response times: Establishing the congruency effect. Participants performed this task without problems. Overall accuracy was good (*M* = 96.4%, *SD* = 4%) and participants' responses were prompt (*M* = 441 ms, *SD* = 67 ms). Prior to the analyses, accuracy percentages were logit-transformed.

To assess the congruency effect, we ran paired-samples t-tests with RT and accuracy as dependent measures. Three trials with RT > 2000 ms were excluded from both analyses, and when RT was the dependent measure, only correct trials when included. As expected, participants responded to congruent trials significantly faster (*M* = 419 ms, *SD* = 63 ms) compared to incongruent trials (*M* = 492 ms, *SD* = 77 ms), $t(70) = 15.52$, $p < .001$. They were also significantly more accurate when responding to congruent (*M* = 99%, *SD* = 2%) than incongruent trials (*M* = 91%, *SD* = 10%), $t(70) = 8.28$, $p < .001$.

5.3.3.2 Phoneme categorization gradiency and inhibitory control. Next, we addressed our primary question as to whether phoneme categorization gradiency is correlated to inhibitory control. To test this, we followed a hierarchical regression approach with separate analyses for VAS slope and residualized VAS slope as the dependent variables. The independent variables were overall spatial Stroop accuracy, overall spatial Stroop RT (across conditions), and the difference in RT between congruent and incongruent conditions (i.e. spatial Stroop / congruency effect). These first two variables were added to account for any effects of overall speed and/or accuracy independently of inhibition.

In the first level of the model, overall accuracy and overall RT were entered as predictors and non-significantly accounted for < 1% of the variance for both residualized VAS slope and VAS slope, $F < 1$, $F < 1$. In the second step, the spatial Stroop effect (i.e. differences in RTs between conditions) was entered and significantly accounted for 11.8% of VAS slope variance and 13.6% of res. VAS slope variance, $\beta = .367$, $F_{change}(1,67) = 8.65$, $p < .01$, $\beta = .395$, $F_{change}(1,67) = 10.24$, $p < .01$.

The results of this analysis suggest that in contrast to our prediction, there may be a link between the way in which individuals deal with top-down inhibition and how they perform the VAS task. Different possibilities as to the nature of this link are discussed in the General Discussion.

5.3.4 Phoneme categorization gradiency and lexical inhibition

We next examined the degree to which inhibition within the lexical system may drive differences in phoneme categorization gradiency. While the primary measure in this task was the fixation record, we started by examining accuracy and reaction times.

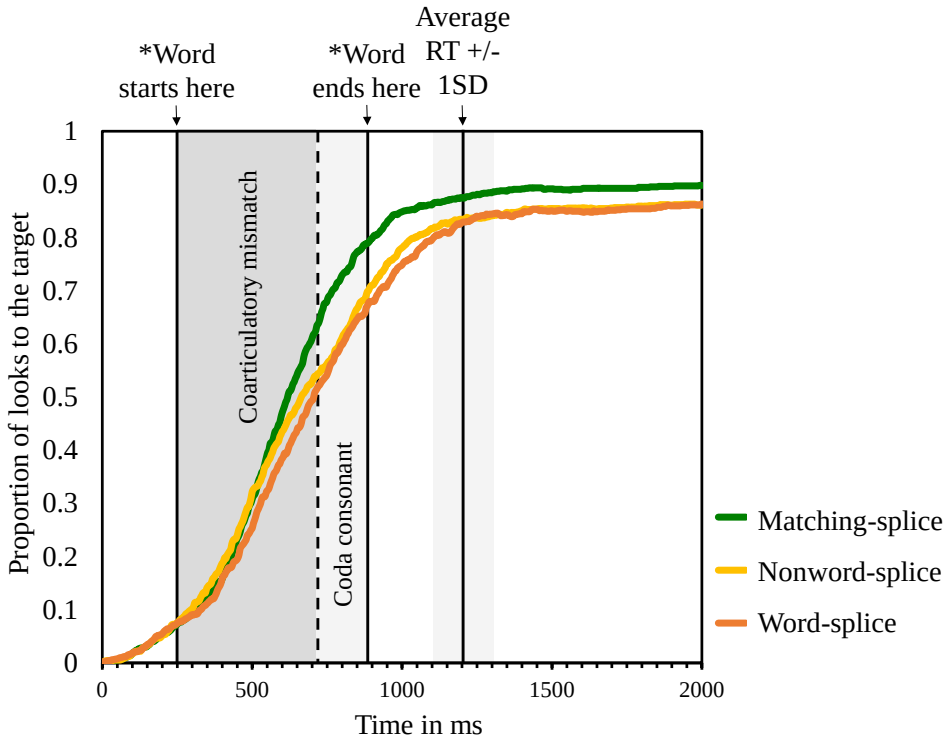
5.3.4.1 Accuracy and response times. Participants performed the word recognition task without difficulties. Overall accuracy was good (*M* = 99.6%, *SD* = 1%) and participants responded in a prompt manner (*M* = 1216 ms, *SD* = 109 ms). Prior to the analyses, accuracy percentages were logit-transformed and RT values were log-transformed to normalize the positive-skewed distribution of the raw data.

A one-way repeated measures analysis of variance showed that splicing condition did not have a significant effect on accuracy $F < 1$. This was not unexpected given that participants' responses were at ceiling (matching-splice: *M* = 99.6%, *SD* = 1.4%; nonword-splice: *M* = 99.6%, *SD* = 1.3%; word-splice: *M* = 99.6%, *SD* = 1.3%).

In contrast , a similar analysis of variance on RT showed that RTs differed significantly between splicing conditions, $F(2,140) = 63.67$, $\eta^2 = .476$, $p<.001$. Post-hoc comparisons revealed that participants were significantly slower in the word-splice (*M* = 1271.8 ms, *SD* = 128.3 ms) compared to the matching-splice (*M* = 1145.0 ms, *SD* = 115.4 ms), $F(1,70) = 123.55$, $\eta^2 = .638$, $p < .001$, and the nonword-splice (*M* = 1230.9 ms, *SD* = 124.7) conditions, $F(1,70) = 11.12$, $\eta^2 = .137$, $p < .001$. This offers preliminary evidence supporting an inhibitory effect as the word splice slowed recognition of the target.

5.3.4.2 Analysis of fixations: *Establishing the splicing effect.* We next analyzed the fixation data to verify the presence of lexical inhibition via the well-replicated splicing effect documented in previous studies (Dahan et al., 2001; Kapnoula &

McMurray, 2016a, 2016b; Kapnoula et al., 2015). To do this, we computed the proportion of trials on which participants fixated the target for each of the three splicing conditions at each point in time (Figure 5.8). This shows an effect of splicing condition with the matching-splice showing earliest rise, followed by the nonword-splice, and word-splice conditions.



*time stamp adjusted for 200 ms oculomotor delay

Figure 5.8 Looks to the target per splice condition in the lexical inhibition task

We tested this statistically by computing the average proportion of fixations to the target between 600 ms and 1600[11] ms post stimulus onset (logit-transformed). The

[11] As in previous studies using this paradigm (Kapnoula & McMurray, 2016a, 2016b; Kapnoula et al., 2015) we chose to analyze fixations starting at 600 ms, because the stem duration (i.e., pre-splice sequence) is about 400 ms long (plus the 200 ms needed to plan an eye movement). The 1600 ms offset was chosen based on the range of participants' reaction times in this kind of task (about 1000 –1600 ms); the broader time window ensured we captured differences in both fast and slow participants.

resulting value was compared between splicing conditions with a series of linear mixed effects models implemented in R and utilizing the *lme4* (version 1.1-6; Bates, Maechler, & Dai, 2009), and *lmerTest* (version 2.0-6; Kuznetsova, Brockhoff, & Christensen, 2013) R packages (R Development Core Team, 2009). We coded the three splice conditions with two contrast codes; 1) matching- versus nonword-splice (-.5/+.5) and 2) nonword- versus word-splice (-.5/+.5). Next, we evaluated models with various random effect structures and found that the most complex model supported by the data was the one with random intercepts for subjects and items.

We found that looks to the target in the matching-splice differed significantly from the nonword-splice condition, $B = -1.055$, $t(4687) = -7.78$, $p < .001$, which suggests that participants were sensitive to the sub-phonemic mismatch. In addition, we found a significant difference between the word- and nonword-splice conditions, $B = -.678$, $t(4687) = -5.00$, $p < .001$, with fewer looks to the target in the former case. Therefore, we replicated previously reported findings showing an effect of splicing manipulation, which is interpreted as evidence for active interference between words (in this case the target word and its competitor).

5.3.4.3 Analysis of fixations: *Top-down inhibition and lexical inhibition.* As our first individual differences analysis with this task, we asked whether lexical inhibition is correlated with top-down inhibition. To examine this, we added the spatial Stroop congruency effect (the RT difference between congruent and incongruent trials) to the model along with the two splice-condition contrasts and their interactions. Neither the spatial Stroop score, $t < 1$, nor any of the interactions, $t < 1$, $t < 1$, had a significant effect

on the proportion of looks to the target, suggesting that the two kinds of inhibition rely on different mechanisms.

5.3.4.4 Analysis of fixations: *Phoneme categorization gradiency and lexical inhibition.* Lastly, we turned to the primary question for the lexical inhibition task: whether gradiency in speech perception may be related to the degree to which lexical representations interfere with each other. We tested this by adding raw and residualized VAS slopes and their interactions with the two contrasts in the fixed effects (after removing the Stroop effect from the model, which was found non-significant). Neither VAS slope, B = .857, t(55) = 1.54, p = .128, nor residualized VAS slope, B = .182, t(55) = 1.33, p = .189, were found to have a significant effect on the proportion of looks to the target and none of the interactions were significant, all $t < 1$.

These results suggest that differences in phoneme categorization gradiency are likely not due to differences in lexical-level inhibition. The alternative hypothesis that the sources of gradiency lie in the early encoding of speech cues is addressed in Section 5.3.5.

5.3.5 Perceptual encoding differences and phoneme categorization gradiency (ERP task)

Lastly, we examined the ERP data, focusing on the N1 as a measure of early cue encoding (Toscano et al., 2010), and the P3 as potentially tapping later categorization processes.

Participants performed the task without problems, with the exception of two participants, who did not come back for the second day of the experiment. In addition, due to a programming error, the first participant was not exposed to all conditions

(specifically, stimuli were blocked by place of articulation and, due to this, the participant was exposed only to auditory stimuli of the same place of articulation as the target) and for this reason was excluded from analyses. Lastly, one participant felt some discomfort after ~10 mins in the ERP booth and was let go. Thus, we excluded 4 participants for this analysis, leaving valid ERP data from 67 participants.

5.3.5.1 Behavioral results. To evaluate participants' ability to perform the task, we computed the number of trials in which the participant responded appropriately. For many VOTs there is no "right" answer (e.g., if the participant was detecting /b/, and they heard a VOT of 25 ms, then a target or non-target response may have been appropriate depending on their own boundary). Thus, we simply used place of articulation as our criterion for accuracy – if they were monitoring for a /b/ and responded "target" for anything from the alveolar continuum, that was considered incorrect. By this measure, average accuracy was good (M = 98.6%, SD = .59%) and participants responded promptly (M = 387.4[12] ms, SD = 345.9 ms).

We next examined the proportion of target responses as a function of VOT and F_0 (Figure 5.9). As expected, participants used both VOT and F_0 information to respond; stimuli with lower VOT/F_0 values were categorized as voiced more frequently compared to stimuli with higher VOT/F_0 values (and vice versa).

[12] Remember that RT was calculated starting at the onset of the prompt (i.e. they were not allowed to respond before). The prompt appeared ~200 ms after the end of the word.

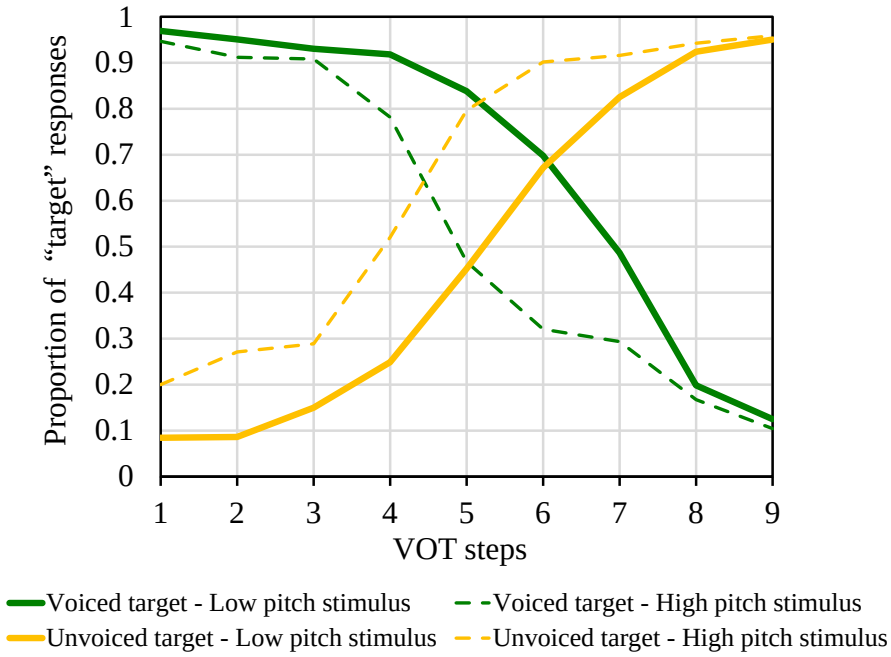


Figure 5.9 Proportion of "target" responses per VOT × F_0 step

To test this statistically, we re-coded the VOT and F_0 in terms of their distance from / compatibility with the target (e.g. the 7th VOT step was recoded as 6 steps distance from the target, when the target was voiced, and as 2 steps distance, when the target was unvoiced, whereas F_0 was recoded as matching/mismatching with the target). We entered these recoded VOT and F_0 variables and their interaction as fixed effects in a logistic mixed effects model implemented using the *glmer* command in the *lme4* package (Kuznetsova et al., 2013). In addition, place of articulation (PoA; effect-coded: alveolar =1; labial = -1), voicing of the target (effect-coded: "voiced" = -1; "unvoiced" = 1) and their interaction were entered as covariates. The dependent variable was whether the participant made a "target" (coded as 1) or "other" response (coded as 0). The random effects structure included random intercepts and random VOT and F_0 slopes for subjects. Continuous predictors were centered.

The results showed significant effects of VOT distance, B = -1.03, z= -22.26, p < .001, and F_0 compatibility, B = -.33, z = -16.09, p < .001. The direction of these effects show, as expected, that greater VOT/F_0 distance from the target predicted lower probability of a "target" response. The VOT × F_0 interaction was not significant. Thus, overall, participants appeared to have performed the ERP task as expected.

5.3.5.2 Electrophysiological results: Establishing the VOT effect on N1. We next examined the electrophysiological data. Figure 5.10 shows the voltage over time as a function of VOT step. A clear negative deflection is observed at around 150 ms, with the characteristic morphology of the N1. In addition, in line with previous findings (Toscano et al., 2010), there seems to be a clear effect of VOT step on the amplitude of N1, with smaller VOTs (i.e. more voiced stimuli) leading to stronger N1.

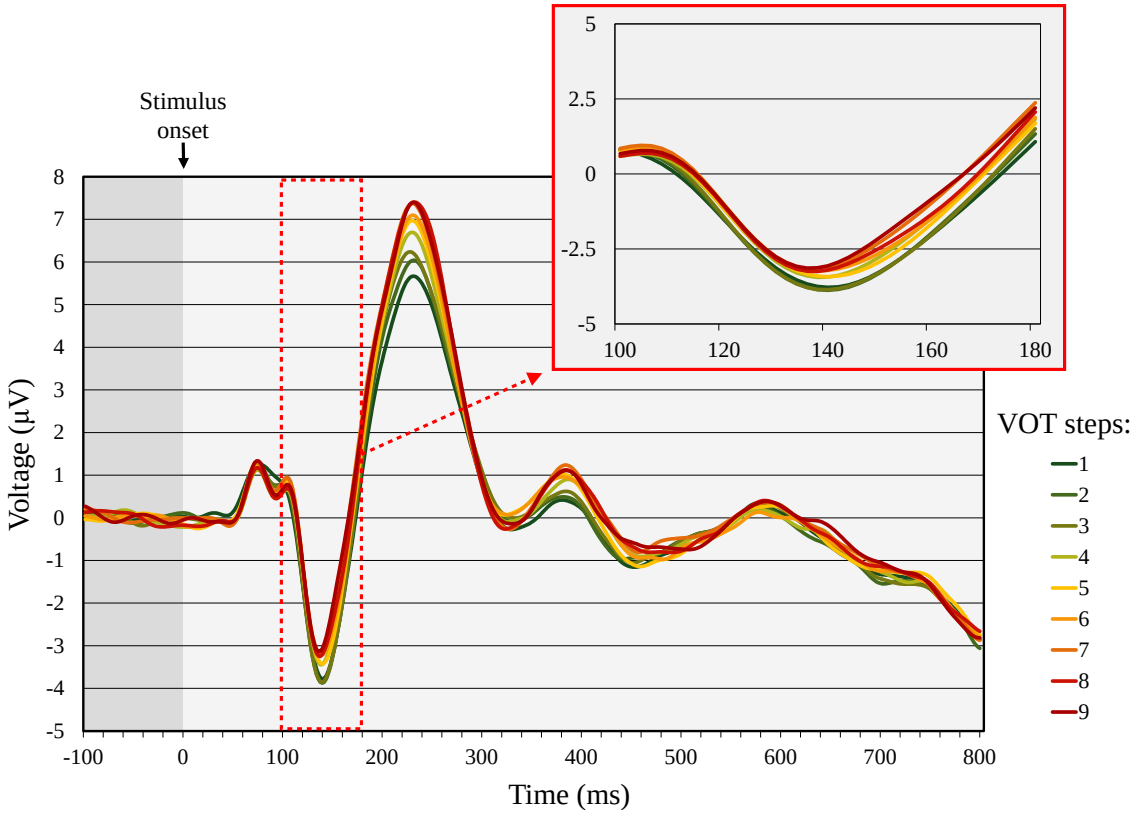


Figure 5.10 Voltage in time per VOT step

To evaluate this statistically, we first calculated the average voltage within a time-window 115 to 170 ms post stimulus onset (henceforth N1 time window) across frontal and central sites. Next, in order to identify the channels that were most sensitive to the N1 component, we computed the average voltage within the N1 time window for each channel. We found that for 12 out of the 20 channels this number was negative (Cz, CP1, CP2, C3, C4, FC2, FC1, CP5, CP6, Fz, FC5, and FC6; see Figure 5.11.A), and thus we included only these channels in the analyses (see voltage fluctuations in time per electrode site in Figure 5.11.B).

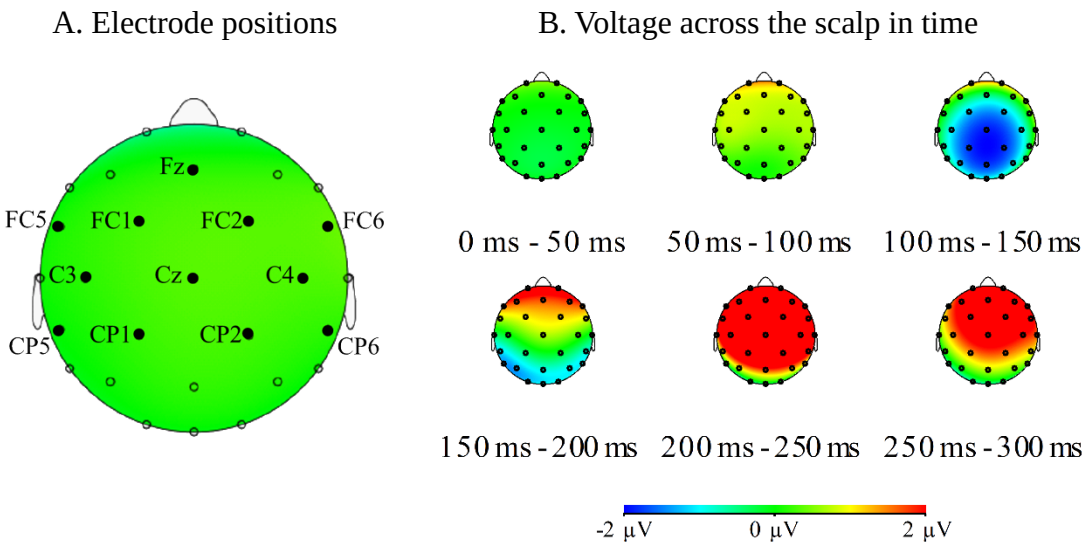


Figure 5.11 Voltage fluctuations in time per electrode site

We then computed the average voltage within the N1 time window separately for each experimental cell. A single cell corresponded to data from 1 to 14 trials (7 stimulus repetitions × 2 target locations) depending on how many trials were excluded during artifact rejection. To eliminate any noise due to the low number of contributing trials, we

excluded from the analyses any cells reflecting 6 or fewer trials. This was then used as the dependent variable in a linear mixed effects model with VOT step, F_0, and their interaction as fixed effects, while PoA, response (target versus other; effect-coded), target stimulus (*bill*, *pill*, *den*, or *ten*; effect-coded), and their interactions were added as covariates. The random effects included both subject and channel. In this regard, we used random slopes of VOT and F_0 slopes for subject (as well as their interaction) and a random slope of VOT for channel.

The results showed a significant main effect of VOT step, $B = .114$, $t(39) = 8.03$, $p < .001$, and F_0, $B = .310$, $t(62) = 5.52$, $p < .001$, with higher VOT and F_0 values predicting higher average voltage (i.e. smaller N1). There was also a significant VOT × F_0, interaction, $B = -.040$, $t(67) = -2.19$, $p < .005$, consistent with a stronger effect of VOT for stimuli with low F_0 (see Figure 5.12). Overall, these results are consistent with previous findings showing that word-initial speech sounds with lower VOTs (i.e. more voiced speech sounds) elicit stronger N1 components (Toscano et al., 2010); the F_0 effect lines up quite clearly with that as well, as lower F_0 (more voiced) show stronger N1s.

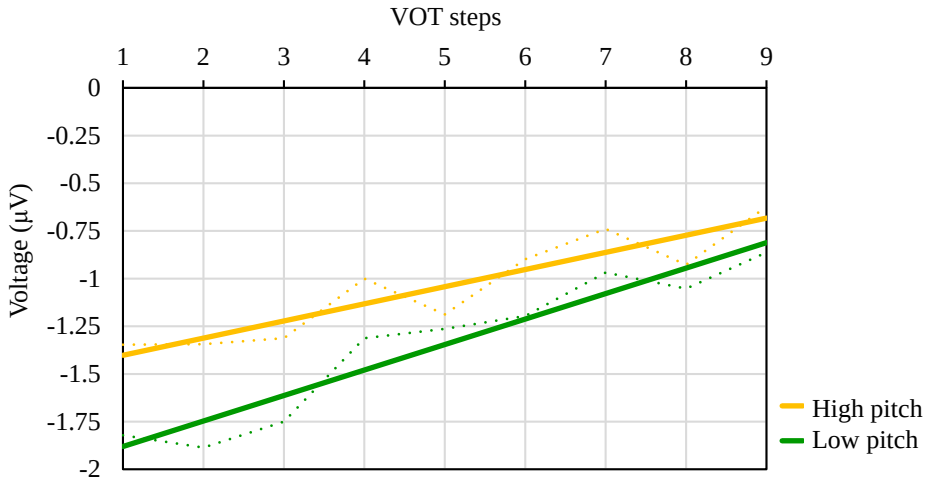


Figure 5.12 N1 amplitude per VOT × F_0 step

5.3.5.3 Electrophysiological results: Evaluating the effect of gradiency on the early encoding of VOT. Next, we added VAS slope / res. VAS slope and their interactions with VOT in the fixed effects. The effect of VAS slope was significant, B = -.159, t(107200) = -3.28, p < .001, as was the effect of res. VAS slope, B = -.059, t(107700) = -4.02, p < .001, with higher gradiency (i.e. shallower VAS slope) predicting smaller N1 amplitude. Crucially, the VOT × VAS slope interaction was also significant, B = -.066, t(3478) = -4.49, p < .001, while the VOT × res. VAS slope was marginally significant, B = -.010, t(2218) = -1.88, p = .06, with the direction of the interactions showing a stronger effect of VOT on N1 amplitude for individuals with higher gradiency scores (i.e. shallower VAS slopes).

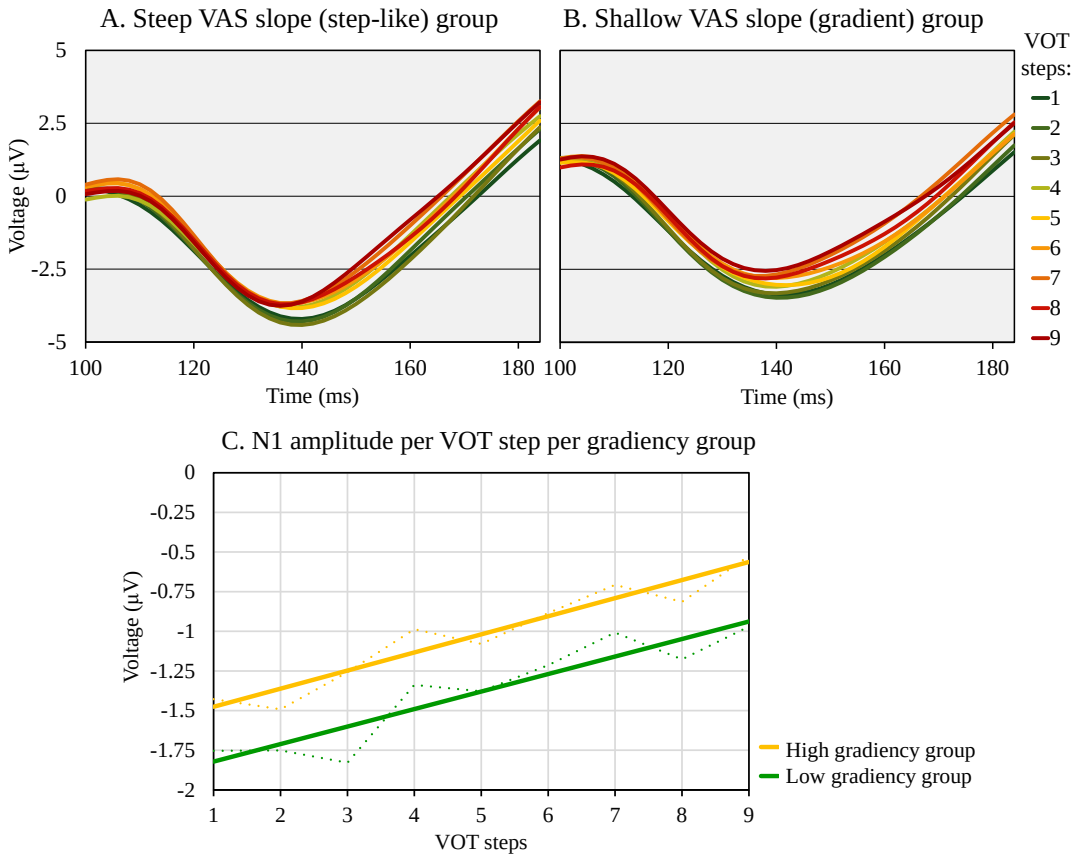


Figure 5.13 N1 amplitude per VOT step per gradiency group

This interaction seems to be in line with the top two panels of Figure 5.13; for gradient listeners (Figure 5.13.B), there seems to be a clearer linear relationship between VOT step and the amplitude of the N1. In contrast, for the categorical group (Figure 5.13.A), we see what looks like a more step-like function, especially around the N1 peak (~140 ms post-stimulus onset), with a large gap between VOT step 3 and 4.

Given this pattern of results, we proceeded to test whether the linearity of the relationship between N1 amplitude and VOT step is linked to phoneme categorization gradiency. To do this, we computed a *binary* variable reflecting whether a specific stimulus belongs to the voiced or unvoiced category for a given subject. This variable was plotted across VOTs for a given participant and, since it is essentially a step function

(pure categorical), we termed it *step*VOT. To compute this, we used each participant's crossover parameters (estimated via the logistic fitting procedure described in Section 2.3.3 from their behavioral responses in the ERP task) for each of the two PoA. We then defined a new variable which was coded as -1 or 1, depending on whether a given VOT was higher or lower than the boundary for that participant-stimulus combination. We added this *step*VOT variable, as well as its interaction with VAS slope / res. VAS slope in the fixed effects. Both the main effect of *step*VOT, B = .053, t(109600) = 3.77, p < .001, B = .049, t(109800) = 3.52, p < .001, and the *step*VOT × VAS slope / res. VAS slope interactions were significant, B = .366, t(110200) = 8.53, p < .001, B =.106, t(110200) = 7.50, p < .001. The direction of the interaction effect suggested a stronger effect of *step*VOT on N1 amplitude for participants with steeper VAS slopes (i.e. less gradient; see Figure 5.14).

Because *step*VOT effect was strongly collinear with VOT, we also conducted a set of log-likelihood model comparisons, comparing each of the two models that included *step*VOT in the fixed effects to the corresponding model without the *step*VOT factor. This ignores the collinearity in each term to ask if *step*VOT accounts for unique variance over and above the VOT model. Both comparisons revealed that the models that included the *step*VOT in the fixed effects were significantly better, $\chi^2(2) = 84.62$, $p < .001$, $\chi^2(2) = 66.59$, $p < .001$.

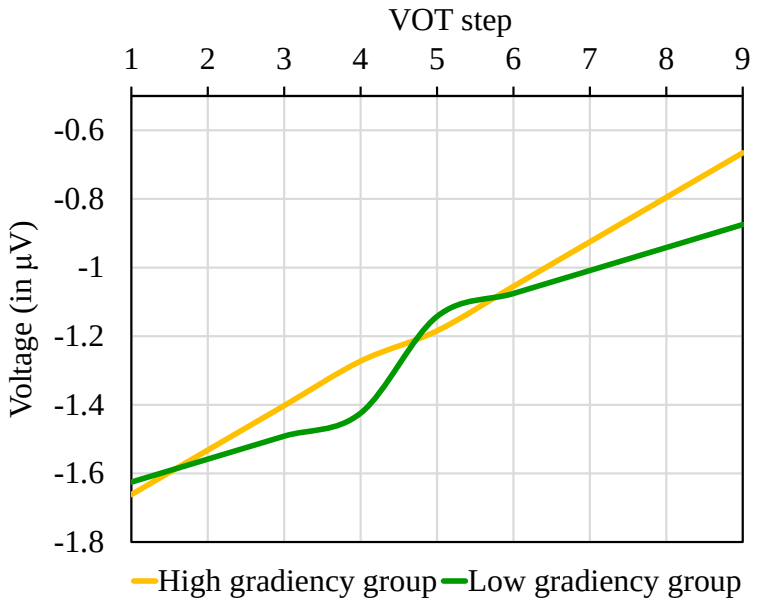


Figure 5.14 Model-estimated effect of VOT step on N1 amplitude when *step*VOT variable is included

To investigate closer the *step*VOT × VAS slope / res. VAS slope interactions, we split the participants by VAS slope (i.e. high and low gradiency group; the 33 participants with the lower gradiency scores were classified as "categorical", while the other 34 were included in the "gradient" group). The same fixed and random effects structures were used as in the main model described in the previous section (*5.3.5.2*). A significant linear effect of raw VOT step on N1 amplitude was found for both the low and the high gradiency group, B = .073, t(45) = 4.24, p < .001, B = .126, t(48) = 6.29, p < .001, respectively. However, the low gradiency group also showed a significant main effect of *step*VOT, B = .117, t(56540) = 6.22, p < .001, whereas the high gradiency group did not, B = -.020, t(57540) = -1.04, p = .30.

Lastly, we tested whether there was a significant linear effect of raw VOT over and above that of *step*VOT for steep-slope categorizers. To get at this, we conducted the reverse analysis (including only data from participants classified as steep categorizers); in

the first model the same fixed and random effects[13] structures were used as above, but instead of raw VOT, we now included *step*VOT in the fixed effects. Then raw VOT was added in the fixed effects of the second model and the two models (with and without raw VOT) were compared in terms of their log-likelihood. As expected, the model that included the raw VOT was a significantly better fit for the data, $\chi^2(1) = 14.73$, $p < .001$, meaning that even for participants that showed evidence for a more categorically-driven perceptual pattern, there was still a linear effect of VOT over and above its categorical/step-like effect.

Overall, the results from our analyses of the N1 suggested that listeners' sensitivity to within-category differences (as assessed in the VAS task) is reflected in their early brain responses to such differences. Specifically, we found evidence for a linear relationship between N1 amplitude and VOT across listeners (replicating Toscano et al., 2010). Crucially, however, we also found evidence to support that this relationship may be better described by a *combination of a linear effect and a step-like function* when looking only at listeners who also show a more categorical / step-like categorization pattern of responses when performing the VAS task. Furthermore, this finding is compatible with the idea that differences in the degree to which listeners exhibit phoneme categorization gradiency behaviorally stem from differences in the perceptual encoding of acoustic cues.

5.3.5.4 Electrophysiological results: Establishing the VOT distance effect on P3. Next we evaluated the effect of stimulus distance from the target on P3 amplitude, which

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

is thought to be a marker of post-perceptual categorization. As pointed out in the description of the rationale behind the ERP task design (see Section 5.2.7), a P3 is usually elicited when a participant responds to an infrequent "target" trial. Therefore we expected to find a significant voltage positivity around the P3 window for trials with a "target" response. In addition, according to previous results (Toscano et al., 2010), we expected to find that for the "target" trials, acoustic distance from the target would be linearly related to P3 amplitude in a negative way (i.e. greater distance → smaller P3).

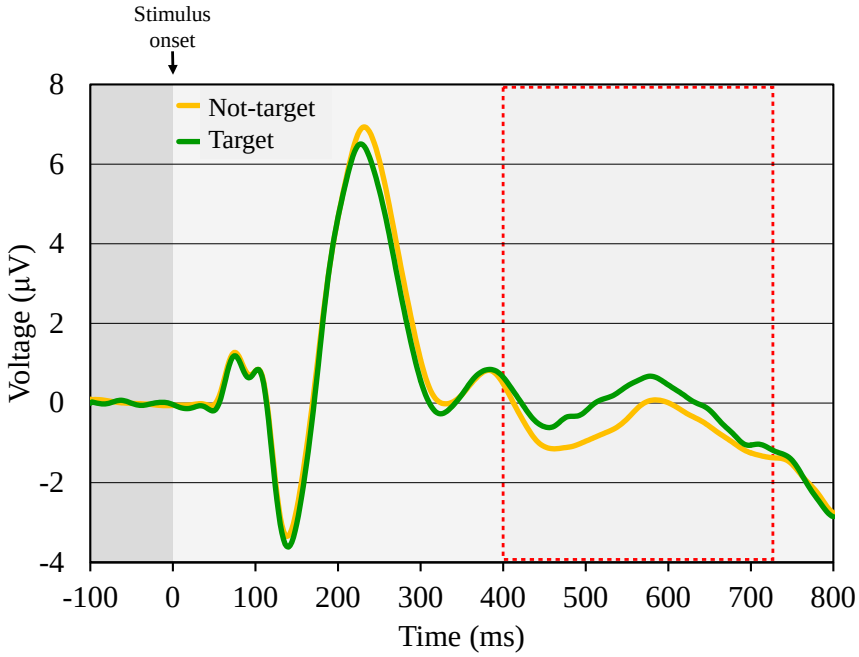


Figure 5.15 Voltage in time by response

Even though we did not observe a robust P3 around the same time window as Toscano et al (~300 – 800 ms), we did observe a difference between "target" and "other" response trials in the expected direction (see Figure 5.15). Based on this difference, and after visual inspection of the data, we decided to calculate the average voltage within a time-window 400 to 730 ms post stimulus onset (henceforth P3 time window) across

central and parietal sites (i.e. locations that are associated with the P3; Nasman & Rosenfeld, 1990). Next, in order to identify the channels that were most sensitive to the P3 component, we computed the average voltage within the P3 time window for each channel, including only trials with a "target" response. We found that for 5 adjacent channels out of the 12 channels this number was positive (Pz, P7, P3, CP2, and CP1; see Figure 5.16.A) and these were the channels that were included in the P3 analyses.

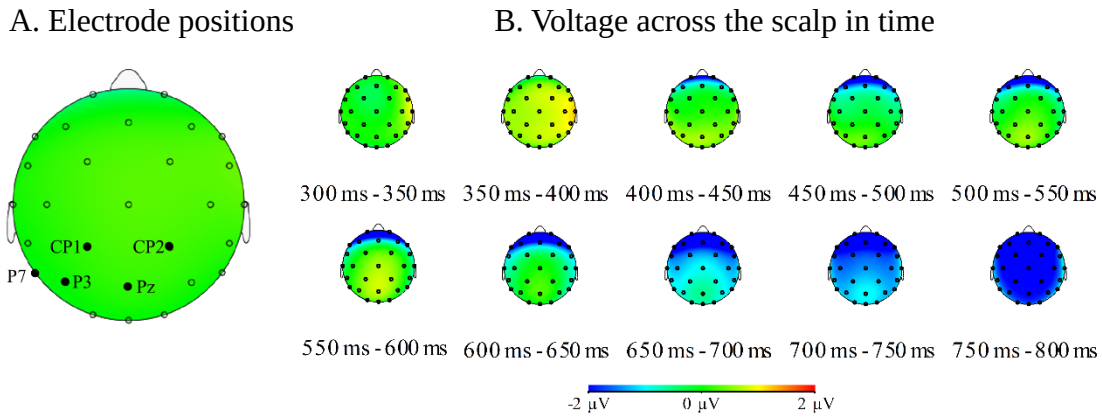


Figure 5.16 Voltage fluctuations in time per electrode site (only "target" trials)

The average voltage in the P3 time window was our dependent variable in a linear mixed effects model with VOT step, F_0, and their interaction, as well as response (coded as 1 for "target" and -1 for "other") and its interaction with VOT as fixed effects, while PoA, target block (*bill*, *pill*, *den*, or *ten*; effect-coded), and their interaction were added as covariates. The random effects included random VOT and F_0 slopes (as well as their interaction) for subjects and random VOT slopes for channels. Given that we aimed at evaluating the effect of the VOT distance from the target (and not VOT per se), we split the data by voicing and fitted two models; one for trials in which the target was a word with a voiced initial (*bill* or *den*) and one for trials in which the target was a word with an

unvoiced initial (*pill* or *ten*). In all analyses, we only included trials in which the PoA of the stimulus matched that of the target (for example, we excluded trials where *bill* was the target and the stimulus was *den* or *ten*). Lastly, similarly to the N1 analyses, we excluded from these analyses any cells reflecting 6 or fewer trials in order to eliminate any noise due to the low number of contributing trials.

For voiced-initial targets, the effect of VOT step was not significant, B = -.018, t(94) = -1.66, p = .101; see Figure 5.17.A); for unvoiced-initial targets, it was significant, t(36) = 4.50, p < .001, in the expected positive direction (see Figure 5.17.B). Despite the absence of a robust effect, the overall pattern was in the expected direction, meaning that, in both cases, higher VOT distance from the target predicted smaller P3.

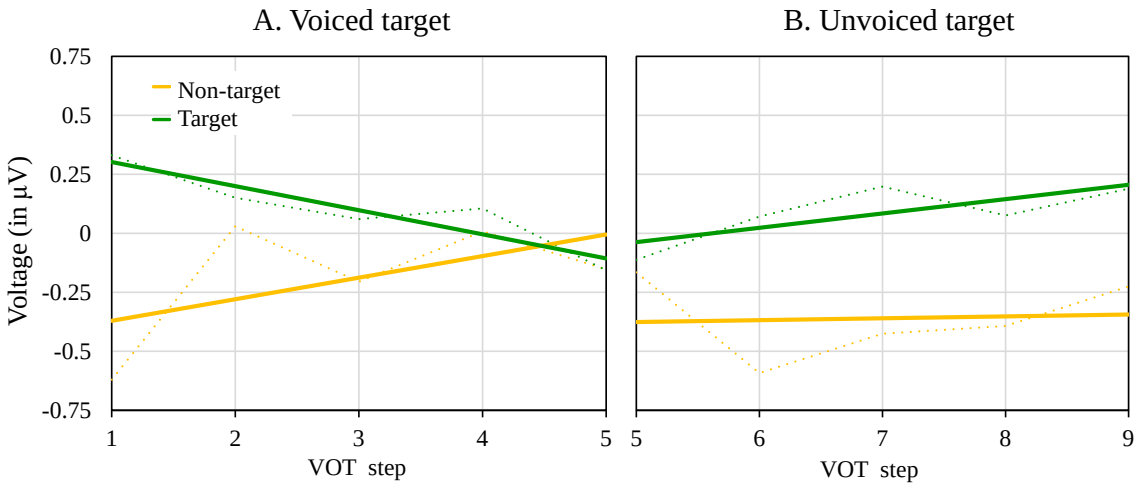


Figure 5.17 Effect of VOT on P3 amplitude per response

Also, as expected, we found a significant positive effect of response (i.e. higher P3 amplitude when the response was "target"), for both voiced-initial, B = .063, t(23370) = 3.29, p < .001, and unvoiced-initial targets, B = .086, t(22980) = 4.77, p < .001 (see vertical difference between lines in Figure 5.17). We also found a significant positive

effect of F_0 (i.e. higher P3 when the pitch was consistent with the target), but only for unvoiced-initial targets, B = .106, t(66) = 2.73, p < .01. Lastly, the VOT × response interaction was significant for both voiced-initial and unvoiced-initial stimuli, B = -.029, t(23430) = -4.90, p < .001, B = .028, t(23050) = 5.04, p < .001, with the direction of the effects suggesting that the VOT effect was stronger for trials with a "target" response (see Figure 5.17).

These results are consistent with previous findings showing that distance from the target (in VOT steps) is negatively correlated with the amplitude of the P3 component. In addition, our results expand previous findings by showing that this can also hold when distance from the target is defined in terms of pitch information; we found that when the pitch of the auditory stimulus is more compatible with the pitch that is characteristic of the target (e.g., utterances of voiced stop consonants tend to also have lower pitch), then a larger P3 is observed compared to when the pitch does not match that of the target. This suggests that the relationship between distance from the target and P3 amplitude is not tied to a specific cue, but may hold true independently of the exact way in which acoustic distance is measured.

5.3.5.5 Electrophysiological results: Evaluating the effect of gradiency on the late encoding of VOT. Next, we added VAS slope / res. VAS slope and their (three-way) interactions with VOT and response in the fixed effects. The question addressed by this analysis was whether the main effects of VOT and response, or their interaction, were modulated by the degree of gradiency participants exhibited in the VAS task.

The main effect of VAS slope was significant for both voiced-initial targets, B = -.243, t(22070) = -3.04, p < .001, and unvoiced-initial targets, B = -.189, t(21550) = -2.66,

p < .01. Similarly, the main effect of res. VAS slope was significant for both voiced-initial targets, B = -.057, t(22230) = -2.31, p < .05, and unvoiced-initial targets, B = -.076, t(21720) = -3.41, p < .001. The direction of the effects suggests that higher gradiency (i.e. shallower VAS slope) predicts higher P3 amplitude across responses.

In addition, the response × VAS slope and response × res. VAS slope interactions were significant, for both voiced-onset targets, B = -.302, t(22360) = -5.28, p < .001, B = -.074, t(22370) = -4.05, p < .001, and unvoiced-initial targets, B = -.151, B = -.048, t(22220) = -3.31, p < .001, t(22190) = -3.18, p < .01. The direction of the interactions suggests a stronger effect of response on P3 amplitude for individuals with higher gradiency scores (i.e. shallower VAS slopes). Lastly, the three-way (VOT × response × (res.) VAS slope) interactions were significant, for the unvoiced-onset targets, B = .063, t(22190) = 3.54, p < .001, t(22170) = 3.76, p < .001, but not for the voiced-onset targets, t < 1, t < 1.

To explore the three-way interaction, we split the data into two gradiency groups based on each participant's average VAS slope (similarly to the post-hoc tests performed for the N1 analyses, the 33 participants with the lower gradiency scores were classified as "categorical", while the other 34 were included in the "gradient" group). For the gradient group, the effect of response on P3 amplitude was significant for both voiced-onset, B = .082, t(11740) = 3.23, p < .001, and unvoiced-onset targets, B = .142, t(11640) = 5.89, p < .001, whereas for the categorical group, it was not for neither the voiced-onset, B = .036, t(11500) = 1.25, p = .21, nor the unvoiced-onset targets, t < 1. Furthermore, the VOT × response interaction was significant for the categorical group for both voiced-onset targets, B = -.034, t(11580) = -3.79, p < .001, and unvoiced-onset targets, B = .030,

t(10910) = 3.72, p < .001; and the same was true for the gradient group, B = -.026, t(11730) = -3.28, p < .05, B = .024, t(11710) = 3.28, p < .001.

To summarize, we found that individuals with shallower VAS slopes (i.e. higher gradiency) had stronger P3s, and the expected effect of response on P3 was robust only for gradient categorizers. The three-way interaction showed that the VOT × response interaction was more robust for participants with steeper VAS slopes (i.e. more categorical). Crucially, this last finding, coupled with the finding that gradient listeners exhibit a more robust effect of response on P3 amplitude, seems to suggest that for the more categorical participants, the effect of response depended highly on VOT step (i.e. distance from the target), whereas for gradient participants, the effect of response is robust independently of the VOT step (see Figure 5.18).

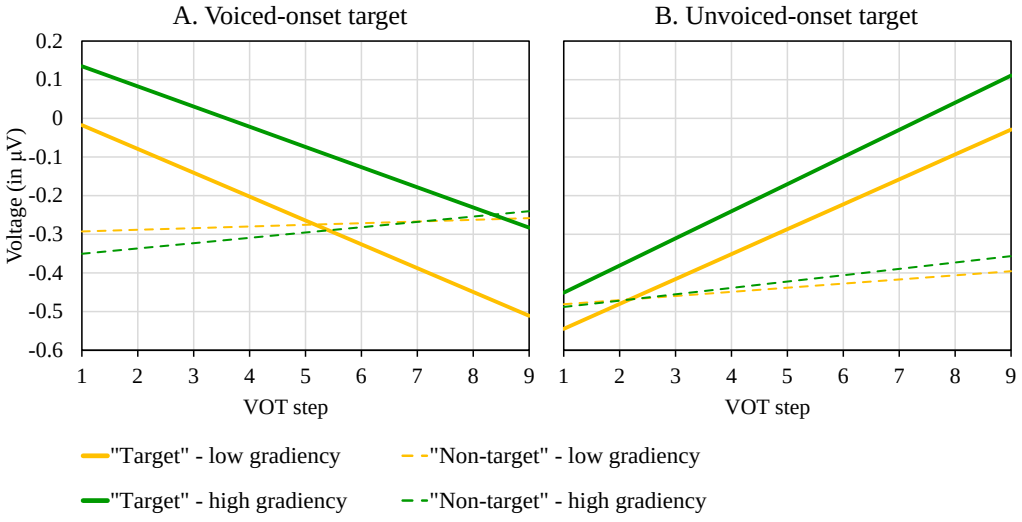


Figure 5.18 Model-estimated effect of VOT and response on P3 amplitude per gradiency group

5.3.5.6 Electrophysiological results: Summary. First, the results from our baseline analyses of the electrophysiological data (see Sections *5.3.5.2* and *5.3.5.4*) are consistent

with previous studies (Toscano et al., 2010) showing: 1) a linear main effect of VOT on N1 amplitude, 2) an effect of response ("target" versus" "other") on P3 amplitude, and 3) an effect of VOT distance from the target on P3 amplitude. Second, we expanded these findings by including an experimental manipulation of F_0 and were able to show that 1) the N1 amplitude is modulated by pitch, with greater N1 amplitude for low-pitch stimuli, which is consistent with previous research showing larger N1 amplitude for low tones (Näätänen, Teder, Alho, & Lavikainen, 1992), and 2) the effect of distance from the target on P3 amplitude is not specific to VOT, but applies to other speech cues as well.

Now, we turn to our primary questions: the role of gradiency in the early and late encoding of speech cues. Our analyses of the N1 component suggest that the effect of speech cues, such as VOT, on auditory ERP components is more robust for individuals that exhibit higher levels of phoneme categorization gradiency. This could mean that gradiency affects the encoding of acoustic cues at a pre-perceptual stage (or the reverse, such that more precise perceptual encoding of acoustic cues allows some listeners to be more gradient when categorizing speech sounds). Crucially, despite the robust linear effect of VOT on N1 across participants, we also found a significant main effect of a binary VOT variable (*step*VOT), but only for participants with steeper VAS slopes (i.e. less gradiency). This may point to some form of early perceptual warping of the acoustic space around the category boundary, which we come back to in the Discussion.

Lastly, our P3 results seem to suggest that even though (as expected) P3 was strongly affected by the response ("target" versus "other"), this effect was modulated by VOT (i.e. larger response effect when the acoustic distance from the target was small) for both groups (see Figures 5.17 and 5.18), which replicates the results reported by Toscano

et al. (2010). Interestingly, this modulation was stronger for listeners with steeper VAS slopes (i.e. less gradient). The interpretation for this finding is not clear. One possibility, however, is that, if there is some form of warping of the acoustic space in the case of categorical listeners (as suggested by the effect of step VOT on N1), this may lead to a clearer distinction between target and non-target stimuli. We elaborate on this possibility in the Discussion.

5.4 Discussion

Our discussion starts with the potential sources for phoneme categorization gradiency. We next turn to the relationship between VAS measures of phoneme level gradiency and lexical level gradiency.

5.4.1 Sources of phoneme categorization gradiency

Experiment 1 revealed only weak linkages between individuals' patterns of phoneme categorization and their performance in tasks measuring general cognitive abilities (e.g. working memory). Therefore, the sources of this gradiency are likely rooted elsewhere. The main goal of this study was to examine different possibilities, both within and outside the language domain.

One of these possibilities was domain-general, top-down inhibitory control (as assessed by a spatial Stroop task). Interestingly, our results revealed a positive relationship between VAS slope and the congruency effect in the spatial Stroop task, meaning that participants with more gradient VAS response pattern exhibited a smaller congruency effect. This finding was surprising given the lack of a correlation between

VAS slope and the Flanker score (which is also thought to measure top-down inhibitory control) in Experiment 1. That said, we see a pattern forming across the different measures of executive function tasks used in different experiments consistent with a weak, but positive relationship between executive function and gradiency. Therefore, this discrepancy between the results from the Flanker and the Stroop task may be indicative of the weakness of the underlying effect. We will return to this issue in the General Discussion.

In addition to inhibitory control, we also looked at another possible source of differences in phoneme categorization gradiency, this time within the language system: lexical inhibition. Our hypothesis was based on two aspects of spoken word recognition: 1) words actively inhibit with each other during spoken word recognition (Dahan et al., 2001), and 2) activation at the lexical level flows back to the level of phonemes (Elman & McClelland, 1988; Magnuson et al., 2003). In the present context, this means that individuals with higher degree of inter-lexical inhibition may suppress competitor words faster or more effectively, reducing any sensitivity to subtle activation differences due to differences in fine-grained detail. This rationale is also consistent with recent evidence showing (behaviorally and computationally) that higher degree of lexical inhibition leads to more robust competitor inhibition (Kapnoula & McMurray, 2016a). Then, stronger competitor inhibition may in turn lead to faster decay of competitor phonemes due to the feedback flow of activation (for example when an ambiguous þeach word is heard, stronger lexical inhibition should lead to faster suppression of the slightly less active word – e.g. *peach* – which in turn would eliminate the feedback to the phoneme /p/).

Therefore, our hypothesis was that individuals with steeper VAS slopes would also exhibit higher levels of inter-lexical inhibition.

This hypothesis, however, was not confirmed. Even though the interpretation of a null effect is often tricky, a tentative conclusion is that the sources of differences in speech gradiency are likely rooted elsewhere (e.g. differences in the perceptual processing of speech sounds), and not in differences in top-down feedback.

The last hypothesis tested by Experiment 2 was that differences between listeners in how they categorize speech sounds in a behavioral (VAS) task are due to differences in how they perceive them. To address this, we collected 1) a measure of phoneme categorization gradiency (VAS slope) and 2) measures of pre- and post-perceptual encoding of acoustic differences in speech segments (N1 and P3 ERP components respectively), and examined possible links between them.

Our results provided evidence for the first time that individual differences in phoneme categorization gradiency are linked to differences in how listeners encode speech cues, such as VOT. Specifically, we found an overall higher positivity (i.e. smaller N1s and larger P3s) for participants with shallower VAS slopes (i.e. higher gradiency). Second, in addition to the linear main effect of VOT on N1 amplitude, we also found evidence that, for steep-slope categorizers, the link between VOT and N1 amplitude has a step-like component (see Figure 5.14). Third, we found that the (expected) effect of response (target versus other) on P3 amplitude was overall significant; but, interestingly, for the steep-slope categorizers, it seemed to be strongly modulated by the degree of the distance between the stimulus and the target (i.e. stronger effect of response for stimuli that were acoustically closer to the target). Together these

results provide valuable insight into the sources of the individual differences we often observe in phoneme categorization; they show that these differences are likely due to differences in how fine-grained differences are encoded early on.

The two findings that we believe to be most noteworthy are 1) the step-like effect of VOT on N1 amplitude for steep-slope categorizers, and 2) the dependence of the response effect on the distance between the stimulus and the target, again for the steep-slope categorizers. The former provides strong evidence for a pre-categorical basis for the differences observed in the VAS task; steeper VAS slope is likely due to warping of the acoustic space around the boundary. However, it is also critical to note here that, as seen in Figure 5.14, and as implied by the independent effects of the linear and *step*VOT variables and our follow-up analyses (see Section *5.3.5.3*), gradiency seems to be preserved *within* categories in all types of listeners. This point will be brought up again in the discussion of the results from the within-category lexical gradiency task.

The P3 results are a little more difficult to interpret, and to do so we will assume that the P3 component is dependent on both the response *and* the distance between the stimulus and the target category (which is consistent with Toscano et al., 2010). As briefly mentioned earlier, this dependence of the response effect on acoustic distance may be explained by some form of warping of the acoustic space in the case of steep-slope listeners. Specifically, close acoustic similarity to the target plus a "target" response can together lead to a robust P3 for both groups of listeners. However, when the stimulus is acoustically dissimilar, group differences arise: for the steep-slope categorizers, any similarity between the stimulus and the target is further minimized due to the perceptual warping and the "target" response alone is not strong enough to generate a P3. In

contrast, for gradient listeners, perceptual similarity between the stimulus and the target is better preserved, thus contributing to the generation of a P3. This could also explain why gradient categorizers exhibit overall more robust P3s.

5.4.2 Phoneme categorization gradiency and lexical gradiency

Experiment 2 also examined whether phoneme categorization gradiency (observed in the VAS task) is linked to lexical-level gradiency (assessed via a VWP task) and our findings did not show a strong correlation between the two, which is quite intriguing. However, if we take into account what exactly the two tasks (VAS and VWP) are tapping into, this finding may not come as a surprise.

As mentioned in the Experiment 2 Introduction (Section 5.1), the VAS task gives us an estimate of phonological gradiency across cue values and categories. It is, in that way, a measure of how cues are mapped onto categories. In contrast, the VWP task used here captures gradiency only within-categories. This is because in the VWP task we need to split trials by category, based on participants' response functions, and then analyze their fixations to the target and the competitor item *within* that category. Therefore, it becomes clear that these two (i.e. across the board within-category gradiency and differences in overall gradiency) are not mutually exclusive.

The strong evidence for lexical gradiency across listeners in our VWP task (also McMurray et al., 2002) is also consistent with the electrophysiological results presented in sections *5.3.5.2* through *5.3.5.5* showing that even though for some listeners early perceptual encoding of speech sounds seems to be warped around the boundary, gradiency is still preserved *within* categories. This pattern of results speaks directly to the

discrepancy between VWP studies, showing gradiency as the typical pattern of speech processing, and the idea that listeners differ substantially in how gradiently they categorize phonemes. What we show here is that these seemingly contrastive patterns of findings may contribute different pieces of the puzzle.

5.4.3 Conclusions

In conclusion, the results across the different tasks of Experiment 2 are consistent with the idea that all listeners are sensitive to within-category differences. However, some warping of the acoustic space may occur close to the category boundary, leading to the amplification of between-category differences for some listeners. This, in turn, may lead to differences between listeners in how well they can preserve this gradiency.

CHAPTER 6: THE CONSEQUENCES OF GRADIENCY FOR SPOKEN LANGUAGE COMPREHENSION (EXPERIMENT 3)

6.1 Introduction

Experiments 1 and 2 examined the function and sources of phoneme categorization gradiency by relating categorization slope estimated in the VAS paradigm to a variety of other speech and non-speech processes. While Experiment 2 focused largely on the *causes* of individual differences in phoneme categorization guardiancy, Experiment 3 focuses on the *consequences*. One critical finding of Experiment 1 was a significant correlation between VAS slope and multiple cue use, with more gradient individuals being more likely to use secondary cues. This finding indicates there may be a link between categorization gradiency and at least one aspect of speech perception efficiency: how well listeners combine multiple speech cues. The primary goal of Experiment 3 was to continue this investigation of the role of phoneme categorization gradiency in speech perception, by examining whether and how it affects listeners' ability to cope with temporary ambiguities.

As we discussed earlier (Chapter 1), gradiency may prove to be particularly beneficial in circumstances where upcoming input is inconsistent with listeners' early interpretation, creating a potential garden path situation. Consider a situation in which an unfamiliar speaker utters a word such as *þumpernickel* (where *þ* is a labial stop with a VOT of 10 ms, i.e. a speech sound closer to /b/, but somewhat ambiguous between /b/ and /p/). In this case, the listener may initially activate /b/-initial words like *bumpercar* and *butter*. However, once the listener has heard *-nickel*, they must revise this initial interpretation. If the listener was fully committed to /b/, they have effectively made a

garden path error and they may be quite slow to recover (if they can do so at all). In contrast, if they have kept /p/-initial items partially active, they may be able to reactivate them more quickly.

Crucially, it is when the VOT is ambiguous (e.g. around 10-15 ms) − either because the of noise in the encoding of the VOT, or because the talker produced the wrong phoneme (Goldrick & Blumstein, 2006) − that such a misperception is most likely. It is exactly in these situations that the likelihood of needing to revise is thus the greatest. Consequently, a system that maintains competitor activation to the degree that it might be needed later would be well situated for recovering from ambiguity (see also Clayards et al., 2008; McMurray & Farris-Trimble, 2012). In contrast, a categorical system – which ignores within-category detail – cannot take advantage of this kind of processing. Instead, it may fully commit to the incorrect phoneme, and find itself in a costly garden path situation when the disambiguating information arrives.

McMurray et al. (2009) tested this prediction of the gradient account as a general description of typical listeners' performance. They argued that if the gradient activation of phonemes is reflected in the activation of lexical representations, then the time to recover from such garden paths should be related to within-category differences in VOT. To examine this, they used pairs of words with partially overlapping onsets (e.g., *bumpercar* and *pumpernickel*). They constructed VOT continua from these words that ranged from a well-articulated word (e.g. *bumpercar*) to an overt misarticulation (e.g. *pumpercar*). This manipulation resulted in ambiguous stimuli (e.g. *þumpercar*) the onset of which (*þumper*) was partially consistent with both words. As the VOT of the initial ambiguous consonant approached the misarticulated endpoint, this induced lexical garden

paths, meaning that listeners temporarily activated the competitor word (in this case, *bumpercar*) and only later (at the offset of the word) they received evidence in favor of a different word (the target: *pumpernickel*).

McMurray et al. (2009) examined listeners' response to this manipulation in an eye-tracking paradigm that assessed how strongly participants committed to the competitor, as well as how efficiently they recovered from these initial misinterpretations. In line with their prediction, they found that both the probability of a lexical garden path occurring (initial fixations to the competitor), as well as the time to recover from it (fixate the target) were linearly related on the magnitude of the acoustic discrepancy between the target word and the auditory stimulus. This suggests that an early graded commitment may permit more flexible updating when subsequent information arises.

Even though word pairs such as these constitute a rather unnatural situation, it captures something that is quite common. Most of the utterances heard every day are less than ideal due to speech errors and disfluencies, while speech is often processed in poor listening conditions (e.g., a cellphone on a noisy bus) making misperceptions fairly common. In fact, in a naturalistic corpus, Bard, Shillcock, and Altmann, (1988) found that as many as 21% of words could not be recognized until *after* their offset (see also Connine, Blasko, & Hall, 1991, for evidence that final commitment to a word can be affected by subsequent context). Consequently, maintaining gradiency at the phonemic and lexical levels may allow for greater flexibility and/or faster recovery from incorrect interpretations in everyday language comprehension (see Clayards et al., 2008, for evidence that this gradiency can be tuned to the variability of phonetic cues).

While past work focused on the modal listener, Experiment 3 applies this paradigm to individual differences, asking if listeners who are more gradient are also more flexible in dealing with such garden paths compared to listeners who are more categorical. This question is particularly important given the findings from Experiment 2. This experiment demonstrated that even though both gradient and categorical listeners showed evidence for linear encoding of VOT (see N1 results in Section *5.3.5.3*), steeper categorizers also showed evidence for *warping* of the acoustic cue space around the boundary. This could mean that even though all listeners maintain a gradient representation within the category, some kind distortion near the boundary could also limit listeners' ability to recover from ambiguities.

To address this issue, we used two measures of speech processing. As in previous experiments, we continued to assess individual differences in categorization gradiency using a VAS task with a traditional VOT continuum (*bin-pin*). In addition to this, we employed a variant of the lexical garden path VWP task (McMurray et al., 2009) to directly test our hypothesis about the functional role of gradiency in helping listeners cope with ambiguities.

A secondary goal of Experiment 3 was to re-evaluate the role of gradiency in listeners' ability to perceive speech in noise using a different task than the one used in Experiment 1. This change in task was motivated by the idea that the sentence-level information available in the AzBio sentences (used in Experiment 1) may have helped listeners to figure out the missing information using top-down information. Consequently, the finer grained analysis of the signal tapped by our VAS measures may not have been needed. In contrast, the task used in Experiment 3 did not allow

participants to take advantage of any sentence-level information, thus forcing them to rely more on the input itself. Therefore, if gradiency is important for perception of speech in noise, this task would reflect this better.

Finally, Experiment 3 also aimed to evaluate the degree to which categorization gradiency and secondary cue use are relatively stable aspects of speech perception at the individual level. To address this, we introduced a new continuum (between /s/ and /ʃ/) in the VAS task to determine if VAS gradiency was stable across different acoustic/phonetic cues. We also measured the use of two secondary cues for voicing (F_0 and vowel length), and an additional secondary cue (formant transitions) used for distinguishing between fricatives (e.g., between /s/ and /ʃ/).

6.2 Methods

6.2.1 Participants

Sixty-seven (67) monolingual English speakers participated Experiment 3. Participants received course credit for participation in the study, and underwent informed consent in accord with University of Iowa IRB policies. One participant was excluded from the analyses due to failure to follow the instructions in performing the VAS tasks, leaving us with 66 participants for the VAS analyses. Three participants were excluded from the VWP analyses due to eye-tracking-related problems.

6.2.2 Design

All participants performed four tasks developed to measure different aspects of spoken word recognition (see Table 6.1). We used the visual analogue scaling task (VAS;

Kong & Edwards, 2011; Munson & Carlson, in press; Schellinger, Edwards, Munson, & Beckman, 2008) with two kinds of speech continua (a VOT × F_0 /b/-p/ continuum and a frication × formant transition /s/-/ʃ/ continuum) to measure gradiency of speech categorization (see Section 2.1). As in Experiment 2, we also used a visual version of the VAS task to extract a baseline of each participant's overall tendency to use the endpoints versus the whole range of the line (independently of phoneme categorization processes). As before, this measure was used to compute residualized VAS slope (by extracting the standardized residual of the phoneme VAS slope variance after partialing out the variance explained by the visual VAS slope). Most importantly, these measures (raw and residualized VAS slopes) were used in correlational analyses to examine whether categorization gradiency is related to listeners' flexibility in recovering from lexical garden paths. This was measured with a task similar to that of McMurray et al., 2009 (see *Lexical garden path task*).

As in Experiment 1, we used the 2AFC task to extract an independent measure of secondary cue use (see Section 2.3). Because one of our goals was to test whether the degree to which a listener uses secondary cues is an aspect of speech perception that is relatively stable within an individual, we included more than one set of primary and secondary cues. Specifically, we tested three cue combinations, two associated with voicing (VOT and F_0, and VOT and vowel length), and one associated with fricative place of articulation (frication spectrum and formant transitions).

Finally, as a measure of speech perception accuracy that is more closely aligned to our theoretical questions, we used a subset of the *"Easy-Hard" Word Multi-Talker Speech Database* (Torretta, 1995) (see *Spoken word recognition in noise task*). In

contrast to the AzBio task used in Experiment 1, in this task, listeners only heard one word at a time, spoken by multiple talkers. By eliminating top-down information, and adding bottom-up acoustic variability (talker variation) we hoped to force listeners to rely more on a bottom-up analysis of the speech signal, potentially revealing a role for categorization gradiency.

Participants performed the first three tasks in one day in the following order: 1) VAS tasks (phoneme and visual), 2) 2AFC tasks, 3) Speech-in-noise task. They then returned on a different day to perform the lexical garden path task.

Table 6.1 Order and description of tasks

Order	Task	Domain	Primarily measure of…
1	Phoneme VAS	Speech categorization	phoneme categorization gradiency
	Visual VAS	Visual categorization	task gradiency
2	2AFC	Speech categorization	secondary cue use
3	Speech-in-noise	Speech perception	speech perception in noise
4	Lexical garden path (LGP)	Speech perception	flexibility in spoken word recognition

6.2.3 VAS tasks

6.2.3.1 Phoneme VAS design and materials. Similarly to Experiments 1 and 2, we used the VAS task to measure individual differences in categorization gradiency. To test whether our findings from the previous experiments regarding stop-onset words also apply to other phoneme contrasts, we used two types of stimuli: a labial-onset minimal pair (*bin-pin*) and a fricative-onset minimal pair (*same-shame*).

All four words were recorded by a male monolingual speaker of American English. For the labial set we created a 7 × 5 two-dimensional continuum by orthogonally manipulating VOT and F_0. VOT varied in 7 steps from 0 to 40 ms approximately 6.7 ms apart. For pitch, we used the pitch tiers extracted from the natural recordings of the two endpoints as steps 1 and 5 (with average pitch 138 and 146 Hz respectively) and we then used these as endpoints to construct the intermediate steps similarly to Experiment 2 (see Section 5.2.3).

For the fricative-onset set, we created a 7 × 2 two-dimensional continuum by orthogonally manipulating the frication and the transition/vowel. Note that in this case we could not make a 7 × 5 continuum, like we did for the labials, since transition is not as easily manipulated as F_0. Each participant was presented with all 35 steps from the labial and 14 steps from the fricative stimuli in separate blocks. For labials, each step was presented three times resulting in 105 trials, while for fricatives, each step was presented 7 times, resulting in 98 trials.

6.2.3.2 Phoneme VAS procedure. Similarly to Experiments 1 and 2, participants were presented with a line at the two ends of which were two words. As in Experiment 2, there was no rectangular bar in the middle of the line and participants were asked to listen to each stimulus and then click on the line to indicate where they thought the stimulus they heard falls on the line. As soon as they clicked, the rectangular bar would appear at the point where they clicked and then they could either change their response (by clicking elsewhere on the line) or press the space bar to verify it. Unless the participant had clarifying questions, no further instructions were given. The task took approximately 15 mins to complete.

6.2.3.3 Visual VAS design and materials. The task and materials were identical to those used in Experiment 2 (see Chapter 2).

6.2.3.4 Order of VAS sets. The VAS tasks were conducted first on the first day of the experiment. The order of the three individual sets (voicing, fricative, or visual) was counterbalanced between participants with 6 different possible orders.

6.2.4 2AFC Phoneme identification

6.2.4.1 2AFC task design and materials. Similarly to Experiment 1, we evaluated how much participants use secondary cues using a 2AFC phoneme identification task. For this, we used the same word pairs as for the VAS task: (*bin-pin* and *same-shame*). The labial sets were used to evaluate two secondary cues (F_0 and vowel length [VL]), while the fricative stimulus set was used to evaluate the use of a secondary cue used in the categorization of fricatives (formant transition). In the first labial set, we manipulated VOT and F_0 using 14 of the 35 stimuli from the VAS task, including all seven VOT steps, but only the two extreme pitch values. In the second labial set, we manipulated VOT and vowel length (which has been shown to be a secondary cue for the perception of voicing; (Summerfield, 1981; Toscano & McMurray, 2012). To construct these stimuli we used the VOT continua from the VAS task with the neutral/middle F_0 values. We then kept the portion of the recording up to the burst steady while manipulating the length of the post-burst portion of the recording using the PSOLA algorithm of Praat to construct 5 length steps each of them with a duration of 60%, 80%, 100%, 120%, and 140% that of the original recording. Finally, for the fricative set, we used all seven frication steps

which we spliced onto both of the transition/vowel portions from each of the two endpoints (thus having a 7 × 2 manipulation).

6.2.4.2 Order of 2AFC sets. All three 2AFC sets were presented on the first day of the experiment immediately after the VAS tasks. The fricative set was always presented second, but the order of the two labial sets was counterbalanced between participants.

6.2.4.3 2AFC procedure. As in previous experiments, participants were presented with two rectangular shapes on the two sides of the screen, each one containing one of the two words for that set (*bin*, *pin*, *same*, or *shame*), and heard one word. *Bin* and *same* were always presented on the left side of the screen in their respective blocks. Participants listened to each stimulus and clicked on the box that contained the word they thought best matched what they heard. Once they clicked, the outline of that box would become bold and they could then either change their response or press the space bar to verify it. The task took approximately 14 mins.

6.2.5 Lexical garden path (LGP) task

6.2.5.1 LGP design and materials. To measure how flexibly listeners cope with temporary ambiguities during spoken word recognition, we used a visual world paradigm (VWP) task, originally used by McMurray et al. (2009). In this task, participants were presented with an auditory stimulus that came from pairs like *barricade-parakeet.* For such items, VOT was manipulated along a continuum, such that at some values it could lead participants to temporarily activate the competitor (e.g. *parakeet* with a VOT of 30 ms). However, in this case, the offset of the stimulus (*-cade*) was inconsistent with this initial interpretation, thus forcing them to reactivate the target (*barricade*). Using such

stimuli we were able to measure how well listeners suppressed their initial interpretation (*parakeet*) and activated the target word (*barricade*).

Experimental auditory stimuli consisted of five pairs of phonemically similar (simple or compound) words, each beginning with a labial stop consonant: *bumpercar-pumpernickel*, *barricade-parakeet*, *blanket-plankton*, *beachball-peachpit*, *billboard-pillbox* (see Table 6.2). The words in each pair differed in the voicing of the first consonant, but were identical for the next 2-5 phonemes. For each word pair, we constructed 4 versions of a 7-step continuum between the two words.

Table 6.2 Stimuli used in the LGP task (in International Phonetic Alphabet; IPA)

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

Note: underlining marks the phonemic overlap between the two words in each pair; bolded portions mark the words' offsets

In contrast to McMurray et al., (who used synthesized stimuli), our stimuli were constructed from natural recordings. These were built by splitting a complete recording of each of the words into two parts: an onset (e.g. *bumper–* from *bumpercar*) and an offset (e.g. *–car* from *bumpercar*). Words were split at the point of disambiguation (Table 6.2) and different onsets were spliced together with different offsets to create the stimuli. Specifically, for each word pair, we recorded four types of items, one for each onset ×

offset combination (e.g. *barri-cade*, *para-keet*, *bara-keet*, and *parri-cade*), resulting in a total of 20 items (5 word pairs × 4 types of recordings). For each item, a native speaker of American English recorded multiple tokens in a sound attenuated room at 44,100 Hz. The best two[14] recordings for each of the 20 items were identified and from these we extracted the onset portion (e.g., *parak*-) and offset portion (-*eet*) by cutting at the zero-crossing closest to the point of disambiguation (POD; ~ 384.1 ms). This yielded 20 onsets and 20 offsets that we used to construct our stimuli. Critically, this allowed us to counter-balance any long distance coarticulation in the stimuli as there was a version of each onset (or offset) that coarticulatorily matched both potential offsets (or onsets).

Next, for each of the five word pairs, we constructed eight cross-spliced items following this procedure: Each of the two voiced onsets in a pair (e.g. *bumper~~car~~* and *bumper~~nickel~~*) was spliced onto each of the two offsets that were extracted from different recordings of the same stimuli (e.g. *~~bumper~~car* and *~~bumper~~nickel*). An equivalent procedure was followed for the unvoiced stimuli. This counterbalancing of onsets and offsets ensured that coarticulatory cues in the onsets would not serve as a cue to the offset (e.g., the –er in *bumper* appeared with coarticulation from both the –ar in *car* and with -ɪ from -*nickel*).

Items differing only in the voicing of the onset consonant were next paired together (e.g. *bumpercar* and *pumpercar*) and were used as the endpoints to create 7-step (0-48 ms) VOT continua using progressive cross splicing similar to Experiments 1 and 2. This yielded 140 auditory items (5 pairs × 2 splice conditions × 2 possible words × 7

[14] We needed two recordings because one recording was used to extract the onset and a different one to extract the offset to ensure that even when the onset and offset came from the same word they still had undergone splicing.

VOT steps; see Figure 6.1.B). Each item was presented 3 times resulting in 420 experimental trials.

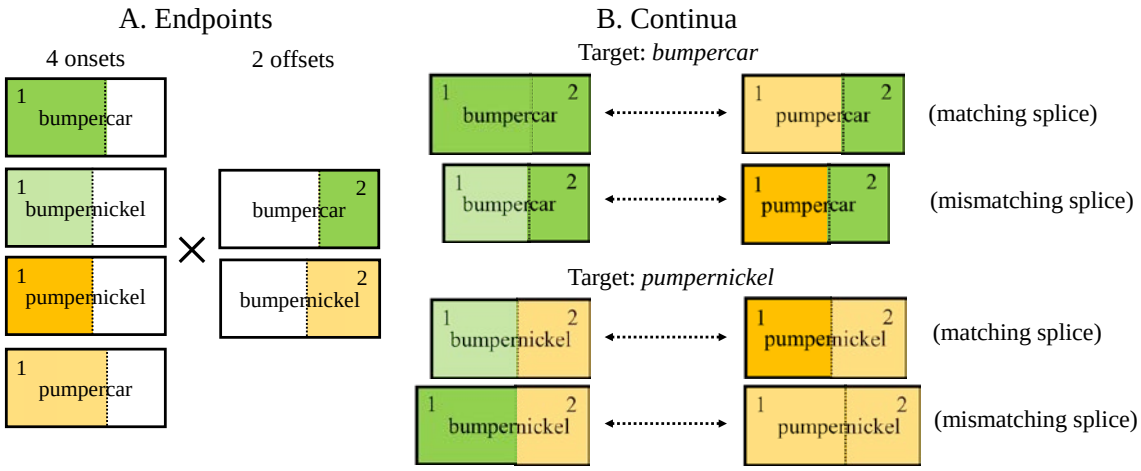


Figure 6.1 Construction of spliced stimuli for continua

Note: the number at the corner of each box indicates the number of the recording token; color saturation indicates whether the recording contained onset and offset that came from a correctly pronounced (dark) or mispronounced

In addition to these 20 VOT continua, 10 pairs of filler items were also used. Filler items began with continuants (half began with an /l/ and half with an /r/); they were not phonetically similar to each other; and they had minimal overlap (e.g. *limousine* and *raspberry*). Similarly to the procedure used for the experimental items, for each filler item we used two types of recordings: one consistent (e.g. *limousine*) and one mispronounced (e.g. *rimousine*). No splicing was performed on the fillers. Each of the 10 filler words was presented an equal number of times (21) in its correct and mispronounced form, yielding a total of 420 (5 pairs × 2 variants × 21 repetitions) filler trials (i.e. as many as the experimental).

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

Visual stimuli consisted of pictures of the referent for each word in a set. For each of these 20 words (five sets of four words), a picture was developed using a standard lab procedure (Apfelbaum et al., 2011; McMurray et al., 2010). For each word, several pictures were downloaded from a commercial clipart database and viewed by a small focus group of undergraduate and graduate students. From this set, one image was selected as the most representative exemplar of that word. These were subsequently edited to remove extraneous elements, adjust colors, and ensure an even clearer depiction of the intended word. The final images were approved by a lab member with extensive experience using the VWP. In addition to the pictures of the stimuli, we also presented a large black X (font: *Trebuchet MS*; size: 85; see Figure 6.2.) at the bottom of the screen. Participants were given the option to click on this X if they felt that none of the four pictures matched what they heard.

6.2.5.2 LGP procedure. Participants were first familiarized with the 20 pictures used in the LGP task by seeing each of the pictures along with its orthographic label. They were then fitted with an SR Research Eyelink II head-mounted eye-tracker. After calibration, participants were given instructions for the task. At the beginning of each trial, the four pictures of a given set (along with the black X) were presented on a 19" monitor operating at 1280 × 1204 resolution. The five visual stimuli were presented in a pentagonal display (see Figure 6.2). The center of each picture was equidistant from the center of the screen (440 pixels) and from each other (517 pixels). Each of the four

pictures were 300 × 300 pixels in size, while the X had 66 pixels width and 80 pixels height.

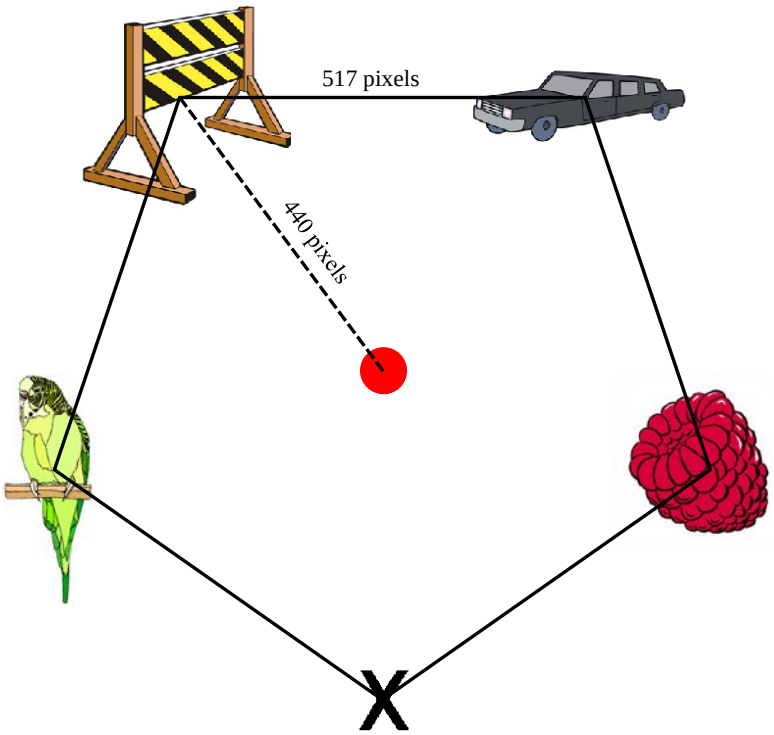


Figure 6.2 Presentation of the LGP visual stimuli in a pentagonal configuration

At the beginning of each trial, along with the five stimuli, a red circle appeared at the center of the screen. This turned blue after 500 ms, cueing the participant to click on it to start the trial. This delay gave time to the participants to briefly look at the pictures before hearing the target word, thus minimizing eye-movements due to visual search (rather than lexical processing). As soon as participants clicked on the circle, it disappeared and the auditory stimulus was played. Participants then clicked on the picture corresponding to the word they heard, and the trial ended. There was no time limit on the

trials, and participants were not encouraged to respond quickly, but they typically responded in less than 2 sec (*M* = 1325.43 ms, *SD* = 200.1 ms).

6.2.5.4 Eye-tracking recording and analysis. Recording and pre-processing of eye-movements was identical to that described in Chapter 5 (Section *5.2.5.4*). The only difference was in assigning fixations to objects; boundaries around the objects here were again extended by 100 pixels or to the end of the screen, whichever was shorter. This did not result in any overlap between the objects.

6.2.6 Spoken word recognition in noise (speech-in-noise) task

In order to measure how well participants cope with noise during spoken word recognition, we presented a sample of 100 words taken from the *"Easy-Hard" Word Multi-Talker Speech Database* (Torretta, 1995). Half of the words were classified as "hard" and half as "easy" (by the developers of the test) based on frequency and neighborhood density measures. Specifically, the "easy" words had high frequency and few neighbors with a lower mean frequency than the target word; while the "hard" words had low frequency and many neighbors with higher mean frequency than the target word. Each word was presented in three different voices (10 different voices were used in the task, five male and five female) and each of the three presentations also varied in terms of the speaking rate in which the word was recorded (a fast, medium, and slow speaking rate condition were used), yielding 300 trials.

Words were masked with white noise at an SNR of 8 dB. For testing, words were presented one at a time over high quality headphones. Participants responded by typing the word they heard and were given unlimited time. Accuracy was computed

automatically and further checked offline by trained research assistants, who corrected any typos.

6.3 Results

Participants performed all tasks successfully with the exception of one participant who failed to follow the instructions for the VAS tasks and was dropped from all analyses of VAS-based measures.

We start by reporting simple correlation and regression analyses that 1) replicate findings from the prior experiments; 2) explore the stability of both VAS slope and cue integration across different types of stimuli (both speech and visual stimuli); and 3) examine the relationship between these measures and the new speech perception in noise task. We next turn to our primary question as to whether measures of categorization gradiency are related to flexibility in recovering from lexical garden paths.

6.3.1 Phoneme categorization gradiency and secondary cue use

We started by fitting participants' responses in the phoneme and visual VAS tasks using the rotated logistic function (see Section 2.1.3). Overall fits were good (R^2 = .97 and R^2 = .95 respectively)[15]. We next fitted participants' 2AFC responses using the logistic function described in Section 2.3.3. Overall fits were good (R^2 = .99).

Then we examined whether and how the three VAS slopes (labial, fricative, and visual) were correlated to each other. VAS slope in the labial-onset VAS task was not significantly correlated with VAS slope extracted from the visual VAS task (r = .205, p =

[15] Six fits (1 from the visual VAS task and five from the labial VAS task) were excluded due to problematic fits.

.116). The same was true for the correlation between fricative VAS slope and visual VAS slope (r = .107, p = .39). However, the two phoneme VAS slopes (for fricative and labial stimuli) were marginally correlated to each other (r = .231, p = .076). This pattern of results overall agrees with the results we reported in Chapter 5 (i.e. stronger correlations between phoneme categorization slopes, than between phoneme and visual slopes), though it also suggests that individual differences in gradiency may derive more from how individual acoustic cues are processed than from an overall inclination to be more or less gradient across cues.

We subsequently looked at whether the estimates of secondary cue use (from the 2AFC task) were correlated to each other across the three sets of cues (F_0, vowel length, and formant transition). We found F_0 use to be positively correlated with use of vowel length (r = .264, p < .05; see Figure 6.3.A), and also with formant transition use in fricative categorization (r = .260, p < .05; see Figure 6.3.B). Formant transition and vowel length were also marginally correlated with each other (r = .210, p = .090; see Figure 6.3.C).

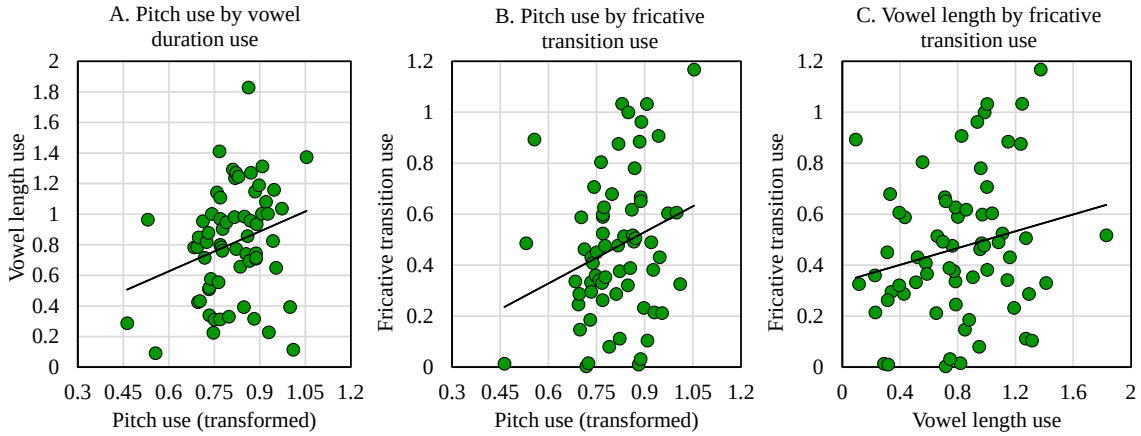


Figure 6.3 Scatterplots of different types of secondary cue use

Despite these significant (or marginal) correlations, it is important to note that these are only small to moderate effects, suggesting that individual differences in cue use may vary depending on the particular cues, as well as the phonemic contrast for which they are used.

We then examined whether phoneme gradiency is linked to secondary cue use. We computed three separate pairs of correlations: between each of the three secondary cues and VAS slope or residualized VAS slope from the corresponding stimulus (e.g., F_0 and VL use were correlated to VAS slope from the voicing continuum; formant transition use was correlated with VAS slope from the fricatives). In agreement with Experiments 1 and 2, F_0 use was significantly correlated with VAS slope ($r = -.348$, $p < .01$), as well as residualized VAS slope ($r = -.371$, $p < .01$). However, use of vowel length was not correlated with VAS slope, ($r = -.095$, $p = .47$), or residualized VAS slope, ($r = -.093$, $p = .48$), and for the fricatives, use of vowel/transition, was not correlated with VAS slope, ($r = -.163$, $p = .191$), or residualized VAS slope, ($r = -.129$, $p = .31$).

Overall, these results suggest that individuals that are highly gradient in one phoneme distinction (e.g., voiced versus unvoiced labial stop consonants) are somewhat more likely to be gradient when performing other types of phoneme distinctions (e.g., fricatives) as well. This agrees with and expands the results from Experiments 1 and 2, where we found significant correlations between the VAS slopes from two types of stop consonants (labial and alveolars), and with Kong and Edwards (submitted) who found good test/re-test reliability for VAS slopes (though within the same cue). Similarly, secondary cue use also seems to be a characteristic of individuals' speech perception

pattern, with individuals showing higher use for one secondary cue (e.g., F_0), also showing high use of other cues (e.g., vowel duration and vowel/transition information). However, both of these latter effects are somewhat small, suggesting there may be more to these differences than simply an overall gradient approach to speech. Third, as in Experiment 2, gradiency in the speech domain was not robustly correlated with visual gradiency, and the relationship between cue integration and gradiency holds even when we account for the visual VAS slope. Lastly, these results replicate the finding also reported in Experiments 1 and 2, that gradiency is correlated with multiple cue integration, with higher use of pitch information predicting higher phoneme categorization gradiency. However, it is interesting that we did not find evidence for an equivalent relationship between gradiency and the use of other types of secondary cues (i.e. vowel length for voicing, and formant transition for frication). We return to this pattern of results in the Discussion.

6.3.2 Gradiency and spoken word recognition in noise

We next examined the role of phoneme categorization gradiency in language processing using a direct measure of speech perception accuracy. We started by investigating which of the stimuli characteristics are important for accuracy in the speech-in-noise task. To do so, we fitted a logistic mixed effects model with trial-by-trial accuracy as the dependent variable. Fixed effects included 1) difficulty (determined by the authors of the test based on frequency and neighborhood density); 2) speaking rate (effect-coded into two variables, one comparing fast rate to the slow rate [FR=1, SR=-1], and the other comparing fast rate to the medium rate [FR=1, MR=-1]); and 3) the

difficulty × rate interaction. 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.

Difficulty showed a marginally significant effect, B = .308, z = 1.73, p = .084. However, speaking rate showed a more robust effect, with fast rate predicting significantly worse performance than slow rate, B = -.347, z = -3.27, p < .01, and marginally significantly worse performance compared to medium rate, B = -.173, z = -1.65, p = .099. None of the interaction terms were significant.

To assess the relationship between categorization gradiency and perception of speech in noise, we fitted a pair of models which included the same fixed and random effects as above with the addition of either VAS slope or residualized VAS slope (extracted from the labial categorization task) as a between-subject fixed effect. Neither the addition of VAS slope, $\chi^2(1) = .001$, p = .97, nor that of residualized VAS slope, $\chi^2(1) = .025$, p = .88, improved the fit of the model. The same was true in regard to the addition of the VAS slope, and residualized VAS slope extracted from the categorization of fricatives, VAS: $\chi^2(1) = .018$, p = .89; residualized VAS: $\chi^2(1) = .255$, p = .61. Thus, phoneme categorization gradiency does not appear to play a role in how well listeners can comprehend speech in a noisy background. These results are consistent with Experiment 1, where gradiency was not correlated with performance in the AzBio task.

6.3.3 Gradiency and recovery from lexical garden paths

Next we addressed whether maintaining within-category information (i.e. higher gradiency) helps listeners when they need to reconsider their initial interpretation of the input. We started by considering the listeners as a whole, both to replicate McMurray et

al. (2009) and to understand the range of measures that may be useful as correlates of individual differences.

We first investigated the strength of different effects on participants' accuracy, response times, and eye-movements in order to identify which factors should be included in the primary analyses. Next, we performed our main analyses focusing more closely on (1) how listeners dealt with the lexical garden paths induced by our stimulus manipulation and (2) whether their ability to cope with ambiguities was related to phoneme categorization gradiency.

For all analyses raw VOT step (1-7) was recoded as distance from the target (i.e. target distance or tDist), similarly to McMurray et al. (2009). For example, for a stimulus with an onset VOT step of 1 (0 ms), tDist took a value of 0 for voiced-onset targets (e.g., the *bumpercar-pumpercar* continuum) and 6 for non-voiced-onset targets (e.g., the *pumpernickel-bumpernickel* continuum), while for a stimulus with an onset VOT step 7 (48 ms), tDist was coded as 0 for non-voiced-onset targets and as 6 for voiced-onset targets. This was done to allow us to collapse the voiced and voiceless continua (e.g., we could collapse the *bumpercar→pumpercar* and the *pumpernickel→bumpernickel* continua). When this was done, an additional term indicating the voicing of the word endpoint was included in the analysis.

6.3.3.1 Preliminary analyses: Effects of distance from target and splice. Participants performed the task without problems and responded rapidly (M = 1325.43 ms, SD = 200.1 ms). We examined the mouse click (identification) responses to determine if the participants were able to recover from the garden path at all. For completely unambiguous targets 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.

As shown in Figure 6.4, as tDist increased, participants were more likely to click on the X (indicating that none of the pictures on the screen matched what they heard). However, even when the VOT was completely mismatching, participants still selected the target word 25.2% of the time. It is also crucial to note here that even when the onset of the stimulus fully matched the onset of the competitor (i.e. tDist = 6) the participants clicked on the competitor picture only 6% of the time.

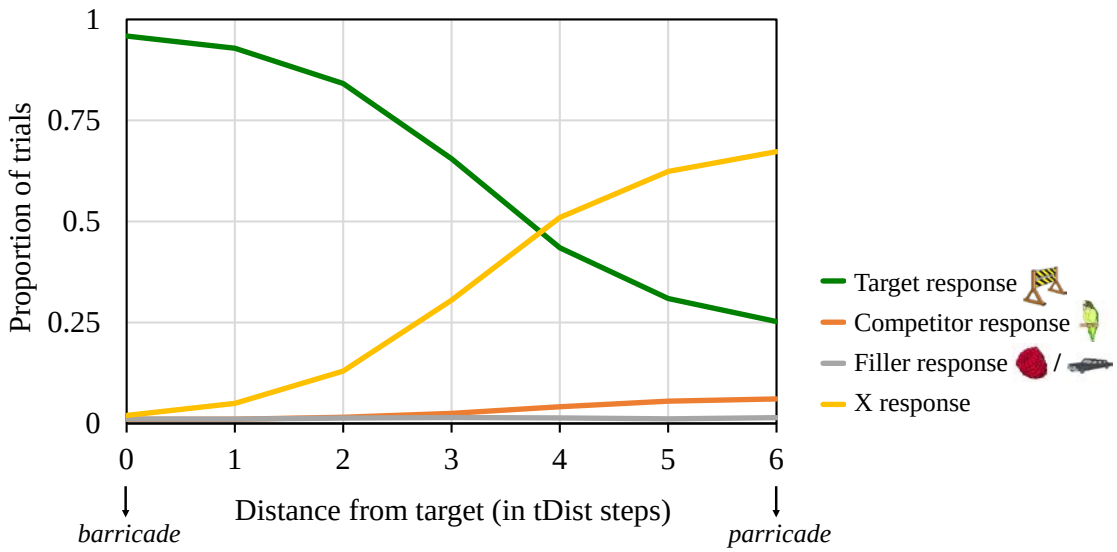


Figure 6.4 Average proportion of clicks to the target/competitor/filler/X as a function of stimulus distance from the target (tDist)

To assess these effects statistically, we examined the main effects of our stimulus manipulations on participants' accuracy and response times. We also conducted a preliminary analysis of the overall amount of fixations (though we turn to a more detailed analysis of the eye movements in the next section). Each analysis examined three factors:

1) distance from the target (tDist); 2) target voicing (whether the target started with a /b/ [e.g., *barricade*] or /p/ [*parakeet*]; and 3) splice condition (i.e. whether the onset and offset of a stimulus came from the same or a different item; see Figure 6.1). These three factors and their interactions were entered in a set of mixed effects models as fixed effects. The purpose of these analyses was to help us decide whether we should keep these factors in our main analyses, or collapse across them.

In the first analysis, we fitted a mixed effects model with random slopes of target distance (tDist) for subject and item. Target voicing and splice were effect-coded and the dependent variable was accuracy (logit-transformed; see Figures 6.5.A and 6.5.B). We found a significant main effect of distance from the target, $B = -1.795$, $t(27) = -12.32$, $p < .001$ and target voicing, $B = 1.014$, $t(8) = 2.77$, $p < .05$, but not splice $t < 1$. None of the interactions were significant.

In the second analysis, the same random and fixed effects were used as in the accuracy analyses and RT was entered as the dependent variable. Only correct trials were included and RTs were log-transformed because the distribution was substantially positively skewed. There was a significant main effect of distance from the target, $B = .054$, $t(45) = 10.18$, $p < .001$, and a marginally significant effect of target voicing, $B = -.041$, $t(8) = -2.03$, $p = .076$. Neither the splice condition, $t < 1$, nor any of the interactions were significant (see Figure 6.5.C).

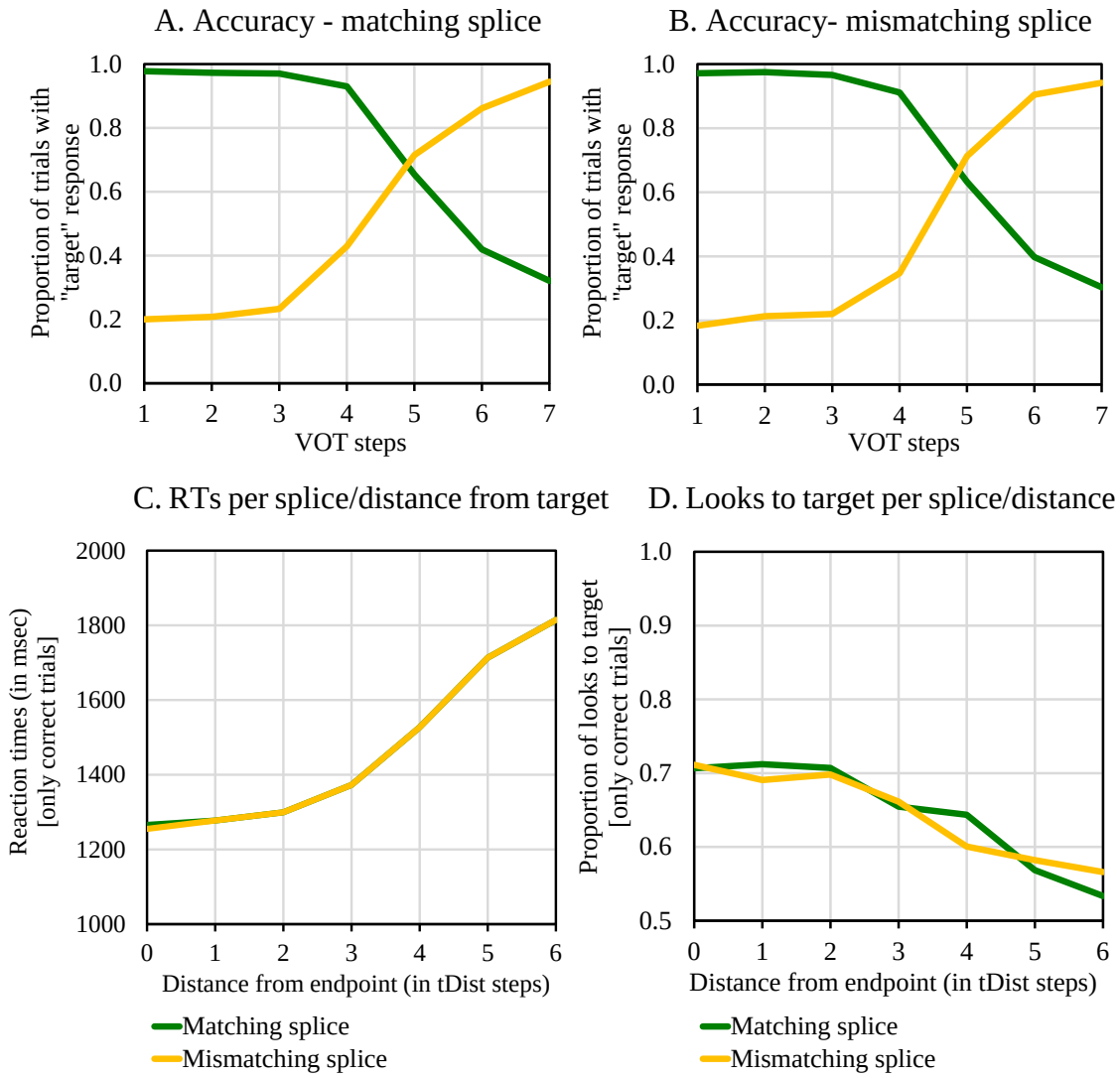


Figure 6.5 Mean reaction times as a function of splice and distance from target (tDist; panel A); proportion of looks to the target as a function of splice and distance from target (tDist; panel B); mean accuracy as a function of VOT step for matching splice (panel C)

Lastly, we examined the eye-movement data – specifically, participants' looks to the picture of the target (see Figure 6.5.D). Figure 6.6.A shows the proportion of trials on which the participant was fixating the target at each 4 ms time window. As seen in Figure 6.6.A, participants seemed to look more to the target when the auditory stimulus was very similar to the target (small tDist), and looks fell off gradiently as tDist increased. However, as Figure 6.6.B shows, splice condition did not seem to have an effect. To test

these statistically, we fitted a mixed effects model with the same random and fixed effects as the two previous models (random intercepts and random tDist slopes for subject and item) and looks to the target as the dependent variable. Our measure of looks to the target was the average proportion of looks to the picture of the target within a time window starting at the point of disambiguation of the stimulus (POD; corrected for 200 ms oculomotor delay) and until 2000 ms. As in the RT analyses, only correct trials were included.

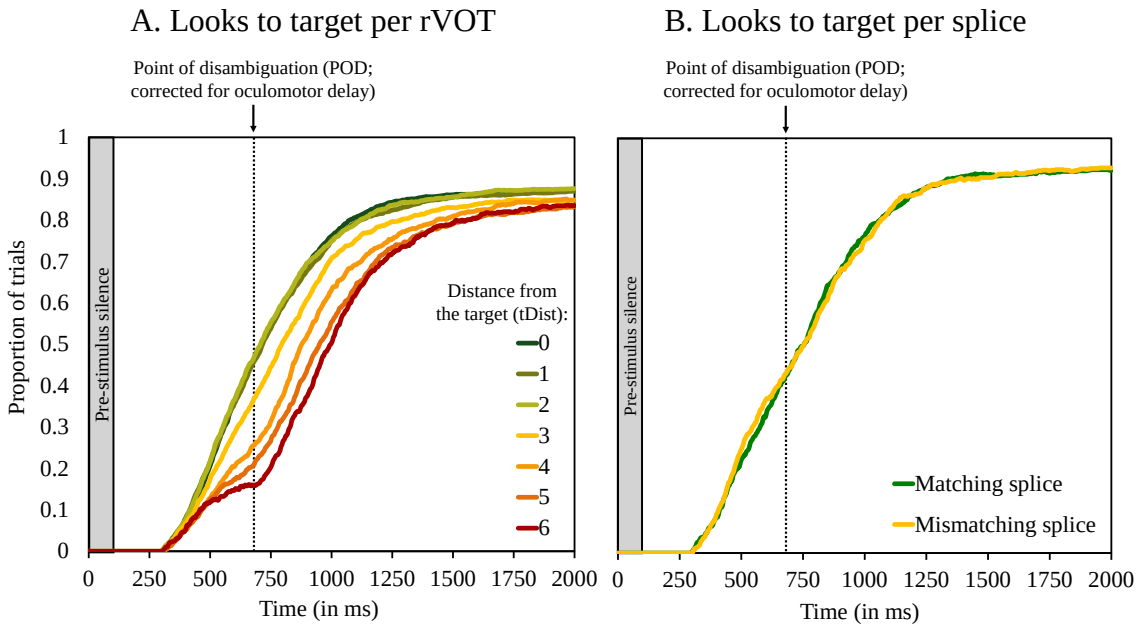


Figure 6.6 Proportion of fixations to the target as a function of: 1) time and rVOT (panel A) and 2) time and splice condition (panel B)

As expected, there was a significant main effect of distance from the target, B = -.263, t(13) = -10.38, p < .001. None of the other main effects were significant, but the three-way interaction was, B = -.031, t(5577) = -2.05, p < .05. 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 = -.282, t(5) = -9.60, p < .001, and

unvoiced-initial targets, B = -.237, t(6.5) = -5.79, p < .001. Neither splice, nor the splice × distance interaction were significant.

In sum, the preliminary analyses showed a robust effect of distance from the target for all measures and target voicing for accuracy and RT. Therefore, we decided to keep distance from the target (tDist) and target voicing in our main analyses and collapse the data across splice condition.

6.3.3.2 Primary analyses: Effects of gradiency on lexical garden paths. Next we turned to our primary question, that is, whether phoneme categorization gradiency affects how people recover from lexical garden paths. We did so in three steps, each one examining a different aspect of performance in the VWP task. First, for each trial we determined whether the participant fixated the competitor prior to the POD (a "garden path" trial), and analyzed the proportion of garden-pathed trials. Second, for each garden path trial, we determined whether the participant "recovered" by ultimately looking at and/or selecting the correct target, even though they had looked at the competitor prior to the POD (only garden path trials were included in this analysis). And third, we examined latency of recovery (i.e. how long it took participants to look to the picture of the target after the POD; only recovered trials were included in this analysis).

We used mixed effects models to evaluate the effect of gradiency (VAS slope) on all three garden path measures. For all analyses, raw proportions were logit-transformed and the latency measure was log-transformed because the distribution of raw values was positively skewed. In addition, the VAS slopes included in these models correspond to the stop-initial stimulus set (rather than the fricative-initial), because the acoustic manipulation in that set (i.e. VOT × F_0) most closely matched the stimuli used here.

The first analysis examined how likely participants were to look at the competitor item (e.g. the picture of a *parakeet* when listening *barricade*) prior to the POD. To compute this, each trial was given a value of 1, if the participant looked at the competitor for any amount of time before the POD for that specific stimulus, and a 0 otherwise. This was averaged within cell, logit-transformed, and examined as a function of 1) target distance, 2) target voicing, and 3) the participants' estimate of gradiency (either VAS slope or residualized VAS slope). In the first model, the maximal random effects structure justified by our data included random intercepts and random slopes of distance from target for both subjects and items. A second model was similar to the first, differing only in including residualized VAS slope instead of VAS slope in the fixed effects.

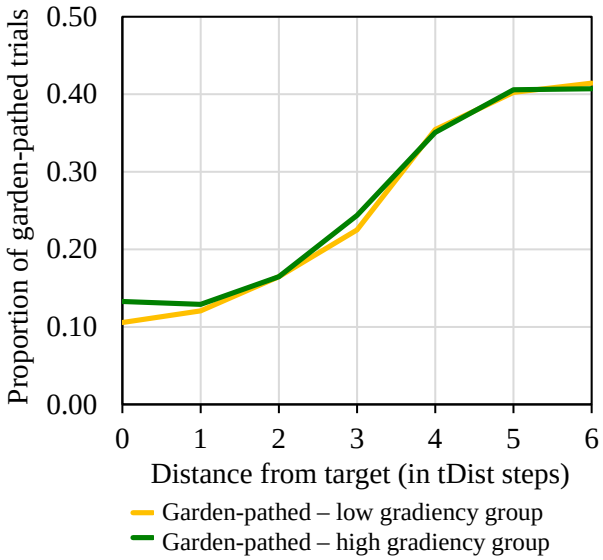


Figure 6.7 Proportion of garden-pathed trials as a function of distance from the target (tDist) for each gradiency group

These models showed that distance from the target was a significant predictor of the proportion of garden-pathed trials in both models where the other fixed effect was VAS slope, B = .673, t(11) = 9.46, p < .001, and residualized VAS slope, B = .674, t(11)

= 9.49, p < .001, with greater distance from the target predicting higher proportion of garden-pathed trials. This replicates McMurray et al. (2009) and suggests that the likelihood of committing to the incorrect option is a function of fine-grained differences in VOT. Target voicing was not significant in either model, B = -.226, t(8) = -1.63, p = .141, B = -.231, t(8) = -1.68, p = .132, suggesting results were similar on both sides of the continuum. Also, neither VAS slope, t < 1, nor residualized VAS slope, t < 1, were significant predictors in these models. Lastly, none of the interaction terms were significant. This suggests that phoneme categorization gradiency does not affect the likelihood of a listener activating a competitor word based on early misleading information (see also Figure 6.7), consistent with the within-category lexical gradiency results of Experiment 2.

Next we looked at the likelihood of recovery (i.e. proportion of recovered trials) across participants. Recovered trials were defined as trials in which participants first looked to the competitor picture some time before the point of disambiguation (i.e. garden-pathed trials as defined in the previous section), and then looked to the picture of the target sometime after the point of disambiguation. Recovered trials also included trials for which participants looked at the target, but ultimately clicked elsewhere (predominantly the X). We made the decision to include these trials because we believe that the kind of recovery we are interested in (i.e. at the level of lexical activation) 1) is better reflected by eye-movements and 2) may not directly map to the participants' ultimate decision to click on the target or not.

Raw proportion of recovered trials were logit-transformed prior to analysis. Two mixed effects models were fitted with identical fixed and random effects structures as

described above; they differed only on whether VAS slope or residualized VAS slope was used. Distance from the target was a significant predictor of recovery rate independently of whether the second fixed effect was VAS slope, B = -1.016, t(13.1) = -8.45, p < .001, or residualized VAS slope, B = -1.005, t(12.6) = -8.24, p < .001, with greater distance from the target predicting lower recovery rates (as expected). Target voicing was also significant in both models, B = 1.203, t(8) = 4.75, p < .01, B = 1.212, t(8) = 4.73, p < .01. In addition, even though VAS slope was not a significant predictor, t < 1, residualized VAS slope was found to be a marginally significant predictor of recovery rate, B = -.510, t(46.0) = -1.88, p =.067, with shallower slope (i.e. more gradiency) predicting higher likelihood of recovery. Lastly, the distance × VAS slope interaction was marginally significant, B = -.441, t(47.6) = -1.87, p = .068, while the distance × residualized VAS slope interaction was significant, B = -.155, t(46.5) = -2.27, p < .05.

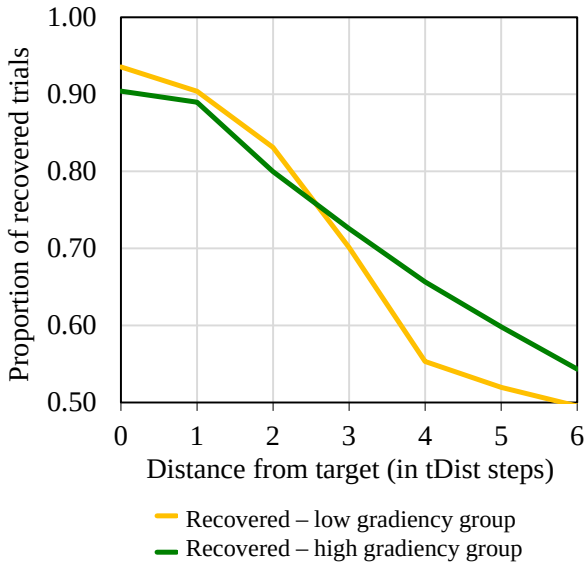


Figure 6.8 Proportion of recovered trials as a function of distance from the target for each gradiency group

In order to investigate these interactions, we split the data into high (tDist > 3) and low (tDist < 3) distance from the target. For low distance from the target, neither VAS slope, $t < 1$, nor residualized VAS slope, $t < 1$, predicted recovery from lexical garden paths. However, for stimuli that were highly divergent from the target, even though VAS slope was not a significant predictor, $B = -1.443$, $t(47.7) = -1.31$, $p = .196$, residualized VAS slope significantly predicted recovery from lexical garden paths, $B = -.696$, $t(46.5) = -2.19$, $p < .05$, with participants with shallower VAS slopes (i.e. more gradiency) showing higher likelihood of recovery (see also Figure 6.8).

Lastly, we looked at the effect of gradiency on the time it took participants to recover from lexical garden paths. This was calculated as the time from the point of disambiguation (plus 200 ms to account for the time it takes to plan an eye-movement) until the first fixation to the target. Only recovered trials were included in these analyses (i.e. trials in which participants garden-pathed sometime before the point of disambiguation, but recovered after it). Two mixed effects models were fitted with the same fixed and random effects as in the previous models.

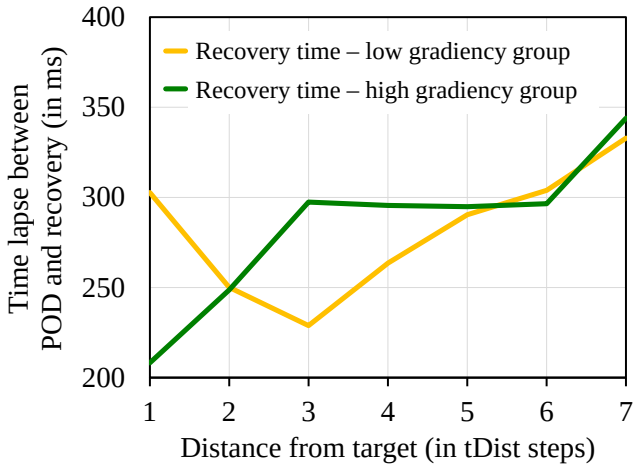


Figure 6.9 Delay of recovery as a function of distance from the target for each gradiency group.

Note: delay of recovery = [time of look to the target] - [point of disambiguation] – [200 ms (oculomotor delay)]

Distance from the target was again a significant predictor of recovery speed independently of whether the second fixed effect was VAS slope, B = .033, t(8.4) = 6.98, p < .001, or residualized VAS slope, B = .033, t(8.5) = 6.29, p < .001, with greater distance predicting slower recovery (as expected). In addition, the distance × target voicing interaction was significant in both models, B = .020, t(8.3) = 4.22, p < .01, B = .019, t(8.3) = 3.84, p < .01. Neither VAS slope, t < 1, nor residualized VAS slope, t < 1, were significant predictors in these models. Lastly, none of the other interaction terms were significant.

Our main analyses showed that phoneme categorization gradiency may not affect the likelihood of a listener making a lexical garden-path (Figure 6.7), or how fast they recover from it (Figure 6.9); however, it does predict the *likelihood of a listener recovering* from a lexical garden path, when it comes to stimuli that diverge greatly from the target (Figure 6.8).

6.4 Discussion

In discussing the findings from Experiment 3, we will first focus on the key results regarding the consequences of phoneme categorization gradiency for different aspects of language processing. Then we will turn to our secondary findings on phoneme categorization gradiency and multiple cue integration across different phonemic contrasts.

6.4.1. Consequences of phoneme categorization gradiency

The primary goal of Experiment 3 was to more closely examine the functional role of phoneme categorization gradiency in speech perception. Specifically, our hypothesis was that individuals with higher levels of gradiency should deal better with noise and/or temporary ambiguities.

To test whether categorization gradiency affects individuals' ability to filter out noise, we administered a speech-in-noise task, similar to that of Experiment 1, but with isolated words as stimuli instead of sentences, so as to eliminate potentially helpful top-down information and force listeners to rely more heavily on bottom-up processing. Despite this change in our stimuli, once again we did not find evidence to support the hypothesis that maintaining gradient representations partially active affects (in a positive or negative way) listeners' ability to filter out external noise. Furthermore, even though Experiment 1 did find a marginal correlation between participants' degree of categorization gradiency and their performance in a speech perception in noise task, this correlation disappeared after accounting for the variance in participants' performance explained by executive function measures. More broadly, the findings from both

experiments appear to agree that phoneme gradiency does not seem to matter for speech perception in noise accuracy.

This does not mean there are no differences between listeners in *how* they solve this problem, but it could mean that both ways (categorizing phonemes more or less gradiently) are good enough when it comes to filtering out noise. In addition, we need to consider that tasks measuring language comprehension via self-report, are also highly sensitive to the amount of effort participants put into them; no matter whether it is due to personality traits or circumstantial fluctuations in motivation, participants' performance in tasks like these is likely affected by their level of engagement.

In contrast, an alternative approach to examining the relationship between categorization gradiency and speech perception processing is to use a measure of language processing that is not outcome-based and thus, not as sensitive to participants' degree of effort. By using this kind of measurements, we can ask different questions that are more closely linked to our theoretical hypotheses regarding speech perception. Adopting this approach, we next asked whether and how different degrees of gradiency affect the way listeners deal with temporary ambiguities in the signal. To test this, we used a visual world paradigm task originally used by McMurray and colleagues (2009), in which participants are presented with auditory stimuli that have been manipulated to induce lexical garden paths (e.g. *bumpernickel*). Analyses of the eye-movement data from this task showed that in such cases participants do temporarily activate the competitor word (*bumpercar*), but are usually able to recover later and activate the correct item. Phoneme categorization gradiency did not predict the likelihood of a garden path or the time it took participants to recover, but it did relate to the likelihood of a listener

recovering from a garden path. Consistent with the notion of recovery, this effect was more robust for auditory stimuli that were highly dissimilar from the target (i.e. with higher VOT distance from the target). This pattern of findings is quite intriguing and provides valuable insights into the nature and functional role of phoneme categorization gradiency, which we discuss next.

First, all listeners, independently of their gradiency on the VAS task, seemed to activate the competitor word early on and the magnitude of this activation was linearly related to the degree of acoustic similarity between the auditory stimulus and the competitor. This suggests that perceiving speech sounds in a gradient manner and, in turn, activating lexical candidates in a manner that reflects this gradiency are fundamental aspects of speech perception. This is in line with previous studies that have found 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. Crucially, this finding is also consistent with our ERP findings from Experiment 2, where we found evidence for linear component to the encoding of speech cues across gradient and categorical listeners.

Second, however, when stimuli were highly divergent from the target, listeners with higher speech categorization gradiency were more likely to recover from lexical garden paths compared to listeners with steeper VAS slopes. This is quite intriguing, particularly when interpreted under the light of our ERP findings from Experiment 2. As we report in Chapter 5, participants showing a more categorical/step-like pattern of distinguishing between phonemes, also showed evidence for some kind of warping of the acoustic cue space around the category boundary. This could mean that the gradient

encoding of speech cues is distorted (to varying degrees between listeners) when it comes to stimuli that fall close to the category boundary. As a result, it may be more difficult (or even impossible) for some listeners to recover the original, undistorted input, which would matter in cases where the listener needs to re-process the signal in order to reconsider an initial erroneous interpretation. This is precisely the kind of case that we examined here, in our lexical garden path task, and we found evidence that listeners who are more likely to warp the speech signal also have lower likelihood of recovering from lexical garden paths.

One possible weakness of this account is that it seems intuitive to predict that more warping at the level of cue encoding (e.g., Experiment 2) should also have led to more initial fixations to the competitor when the VOT mismatched the target, which was not observed here. However, it is possible that despite them being distorted, ambiguous stimuli could still activate multiple items. In addition, the disambiguating information (in the offset) may come before any lexical activation has built up enough to drive a garden path. These two ideas together could mean that any small differences in the bottom-up support words receive may not be enough to drive significant differences in the initial commitment, even as they affect recovery.

There is also another, more indirect mechanism through which this warping may affect speech perception. Due to the weaker (or absence of) warping of ambiguous input, gradient categorizers may not fully suppress competing representations (e.g. when hearing a somewhat ambiguous *bumper*, they do not fully suppress the word *pumpernickel*), because the input is still highly consistent with both (or multiple) items. This may allow them to be better able to re-activate the more weakly activated word

(*pumpernickel*) later on. In contrast, warped input would be less consistent with the weakly activated word, which would make it more susceptible to the suppression from the more activated word.

Whatever the exact mechanism is, the ability to re-activate previously ruled-out items is particularly useful in cases where the input may be misleading early on; 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. If we consider the commonality of such conditions, it becomes clear that being able to point to the factors that may help listeners recover better from such ambiguities would have significant benefits across a wide range of circumstances.

6.4.2 Speech gradiency and multiple cue integration as properties of individuals' language processing

We now turn to our secondary findings. Here, we see a number of places in which we have extended our understanding of what exactly differs among individuals in terms of both the VAS slope and our 2AFC measures of cue integration.

First, we found that neither of the two measures of phoneme categorization gradiency (labial or fricative) were correlated with visual categorization gradiency, but they were marginally correlated to each other. This is in line with our corresponding findings from Experiments 1 and 2, showing significant correlations among measures collected via phoneme VAS tasks. In addition, this is consistent with our assumption that our VAS-based measure of gradiency reflects differences in how listeners categorize

phonemes and not arbitrary task demands; which is also supported by our findings throughout Experiments 2 and 3, where we observe a great overlap between results involving VAS slope and residualized VAS slope. In contrast, none of the two phoneme VAS slopes was significantly correlated with the visual VAS slope. This suggests that the phoneme categorization gradiency measured by our task is tapping into speech processes without being substantially affected by task-related biases. Despite the correlations among VAS slopes for the speech tasks, it is important to point out that the correlations, while significant, were small, suggesting that much of the VAS response may be geared to a specific phonetic contrast or cue, and is less a general property of the listeners' speech perception system.

Second, in line with Experiments 1 and 2, as well as previous work by Kong and Edwards, submitted), our results showed that individuals with higher categorization gradiency scores use pitch information to a higher degree. However, we did not find a significant correlation between categorization gradiency and the other two measures of secondary cue use (vowel length for stop consonants and formant transition for fricatives). This suggests that the link between categorization gradiency and use of pitch information does not apply to all secondary cues.

Distinguishing between the cases in which secondary cue use and gradiency are related versus those in which they are not, may shed light onto the nature of their relationship. For example, one of the possible interpretations for this correlation, discussed at the end of Chapter 4, is that there is a third factor (e.g. executive function) that causes both higher gradiency and higher secondary cues use. This means that the correlation should hold independently of whether the two cues are available close

together in time (e.g. VOT and F_0) or not (e.g. VOT and vowel length, which only becomes known at the end of the vocalic portion). Our results, showing that use of vowel length and gradiency are not correlated, suggest that this might not be the case and is more consistent with a direct causal link between secondary cue use and gradiency (even though we cannot say much about its direction at this point).

Alternatively, we should also consider the possibility that there is something qualitatively different in the relationship between VOT and pitch, compared to the other combinations of speech cues. For example, it could be suggested that due to their close temporal proximity, VOT and F_0 are perceptually integrated and processed as one cue. In contrast, VOT and vowel length are much more temporally separated, while frication and transition are spectrally quite independent. Thus, it could be the fact that VOT and F_0 are perceptually more integrated that is driving this relationship.

These sorts of integral (vs. separable) perceptual dimensions have been explored previously in speech. In fact, Kingston, Diehl, Kirk, and Castleman (2008) used the Garner paradigm to show that VOT and F_0 may be perceptually integral. They further showed that the critical property that drives perceptual integration is the continuation of low frequency energy across the vowel-consonant border. However, as the authors point out, such a continuation is not possible in stop-initial syllables in English because voicing always starts shortly *after* the closure's release. Therefore, since we do not have enough evidence to support such a perceptual integration, it remains unclear whether this is the case or not.

Lastly, the weak but consistently positive correlations between our three measures of secondary cue use suggest that multiple cue integration as a whole is at least partly a

property of the individual − some listeners are more likely than others to rely on additional cues. However, again, the small size of the correlations also suggest that listeners seem to adopt idiosyncratic weightings of individual cues. Therefore, further experiments manipulating the type and availability of the two cues are needed for us to achieve a better understanding of the exact nature of this link.

6.5.3 Conclusions

In conclusion, the primary goal of Experiment 3 was to investigate the potential consequences of phoneme categorization gradiency in language comprehension. Our results indicated that despite gradiency being a fundamental aspect of speech processing across listeners, individual differences do exist and they seem to affect the way in which listeners recover from initial errors in interpreting ambiguous stimuli.

CHAPTER 7: ALTERING CATEGORIZATION GRADIENCY VIA TRAINING (EXPERIMENT 4)

7.1 Introduction

The previous experiments explored the sources and consequences of phoneme categorization gradiency and its links to general, non-speech processes (such as executive function), lower-level perceptual encoding, as well as mid-level lexical processes. The primary conclusions that we can draw thus far are that (1) gradiency is a fundamental aspect of speech perception (Experiments 1-3), (2) listeners differ substantially in the degree to which they maintain gradiency in later response stages (Experiments 1-3), (3) these differences seem to stem from differences in the early encoding of speech cues (Experiment 2), and (4) greater gradiency helps listeners recover from early misinterpretations (Experiment 3). This last finding suggests that gradiency may play a *positive* role in spoken language comprehension, at least in certain cases.

This does not necessarily mean that greater gradiency is always good, but it does suggest that different degrees of gradiency may be more or less helpful depending on the task at hand and the specific difficulties listeners have to deal with. Therefore, increasing (or decreasing) categorization gradiency may have a substantial positive impact on how we process language. However, we do not know whether we can change the degree to which listeners exhibit more or less phoneme categorization gradiency. If anything, our results from Experiment 2 seem to suggest that these differences are based on early perceptual differences in how listeners encode acoustic cues (in a more or less warped way) – and this may be difficult to adjust, particularly in adulthood. Even if the way in which cues are encoded cannot be changed, however, short-term training may potentially

allow listeners to modify the way in which they use whatever information they have available. Addressing this question would not only provide useful theoretical insights as to the nature of categorization gradiency in speech perception, but it may also have great implications for the design of training paradigms that aim at improving how individuals deal with ambiguities in spoken language.

There is indeed some evidence that listeners can change the gradiency of the mapping between speech cues and categories. For instance, Clayards, et al (2008) showed that it may be possible to change how listeners map continuous acoustic cues like VOT to words via probabilistic training. During training, listeners were exposed to one of two different types of VOT probability distributions; stimuli either came from distributions with high variance (14 ms), or low variance (8 ms), while the distribution means were kept the same between groups (0 ms for voiced and 50 ms for unvoiced). Both behavioral responses and eye-movements in a following task revealed a significant difference between the two groups in terms of the sharpness of their categorization slopes, with shallower slopes observed for listeners exposed to the wide VOT distributions. These results suggest that listeners are not only able to maintain fine-grained within-category cue information, but they are also sensitive to the probabilistic properties of their distributions of occurrence.

Experiment 4 also examined whether we can train participants to change the way they categorize phonemes. Crucially, in contrast to Clayards et al (2008), we used our VAS-based measure to more closely test whether we can manipulate listeners' categorization gradiency, as well the way in which they combine multiple speech cues. In addition, we employed a more rigorous training design by adding feedback. This

feedback was experimentally manipulated to change the way in which listeners map acoustic cues to phoneme categories. Specifically, we used three training designs, each of them with a different feedback manipulation aiming at changing the cue-to-phoneme mapping in a different way. In the first group, we aimed to boost secondary cue use (bi-dimensional group) – in other words, make participants better cue integrators; this condition also aimed at testing the hypothesis that increased secondary cue use is what gives rise to more gradiency. The second training design aimed at making participants more probabilistic in the way they map cues to phoneme categories (probabilistic group; similarly to Clayards et al). Finally, the third group aimed at boosting the participants' reliance on the primary cue, and eliminate their use of the secondary cue (unidimensional group).

7.2 Methods

7.2.1 Participants

One hundred (100) monolingual English speakers participated in Experiment 4. Participants received course credit for participation in the study, and underwent informed consent in accord with University of Iowa IRB policies. One participant was excluded due to a technical error and six (6) participants were excluded from the analyses due to failure to perform the VAS tasks as instructed, leaving valid data from 93 participants.

7.2.2 Design

Each participant was randomly assigned to one of three experimental conditions All participants performed the exact same VAS task immediately before and after the

training task. The training task was the only portion of the experiment that differed between experimental conditions. The training stimuli, number of training trials, and training task procedure were identical across experimental conditions and the only difference was in the feedback participants received after their response.

7.2.3 VAS task

7.2.3.1 VAS task design and materials. Similarly to Experiments 1-3, a VAS task was used to measure individuals' phoneme categorization gradiency at the beginning (pre-training VAS task) and at the end of the experiment (post-training VAS task). The exact same VAS task was repeated within and across participants. For the stimuli, we used a subset of the labial-initial stimuli used in Experiment 3. Given that the results from Experiment 3 indicated a small bias towards the "unvoiced" response, we decided to use VOT steps 1-6 (out of 7) and F_0 steps 1-4 (out of 5) to achieve a more balanced pattern between voiced and unvoiced responses. Each participant was presented with all 24 (6 VOT × 4 F_0) stimuli and each step was presented six times yielding 144 trials in each of the two VAS tasks.

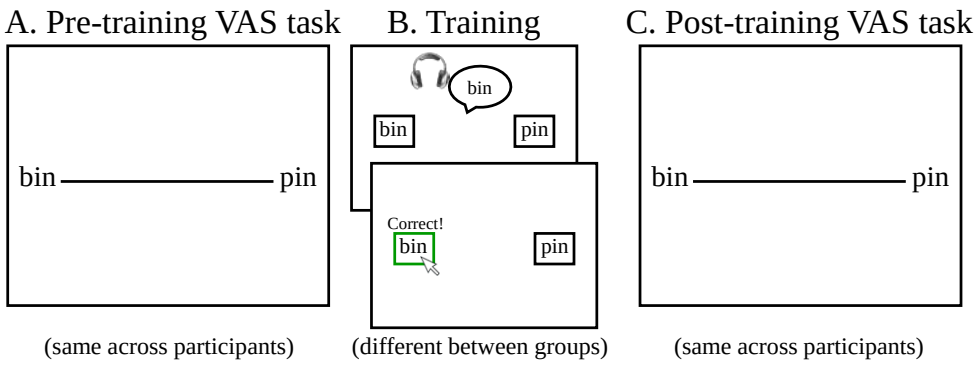


Figure 7.1 Basic structure of the training and testing tasks used in Experiment 4

7.2.3.2 VAS task procedure. Similarly to Experiments 1-3, participants were presented with a line at the two ends of which two words were presented. As in Experiments 2-3, there was no rectangular bar in the middle of the line and participants were asked to listen carefully to each stimulus and then click on the line to indicate where they thought the stimulus they heard falls on the line. As soon as they clicked, the rectangular bar appeared at the point where they clicked and then they could either change their response or press the space bar to verify it. Unless the participant had clarifying questions, no further instructions were given. Each of the two VAS tasks took approximately 11 mins.

7.2.4 Training task

7.2.4.1 Training task design and materials. For the training task, we used the structure of the 2AFC task that was used in the previous experiments, to which we added feedback at the end of each trial. The feedback was systematically manipulated between conditions to support a different mapping of speech cues to phoneme categories. The exact same stimuli used in the VAS tasks were also used in the training task. Each stimulus was presented 30 times yielding 720 (6 VOT × 4 F_0 × 30 repetitions) trials, presented in 10 blocks of 72 trials each.

7.2.4.2 Training task procedure. Participants were presented with two rectangular shapes on the two sides of the screen, each one containing one of the two words for that set (*bin, pin*). Participants heard a single stimulus and then clicked on the box that contained the word they thought best matched what they heard. Once they clicked the outline of that box would become bold and then they could either change their response

or press the space bar to verify it. As soon as they verified their response, the color of the square changed to green, indicating a correct response, or to red, indicating an incorrect response. In addition to the color change, the word "Correct!" or "Wrong!" appeared over the rectangular they had selected. Both of these types of feedback appeared at the same time and remained on the screen for 200 ms. In addition, participants were notified of their average block accuracy score at the end of each block and were encouraged to try their best throughout the training task. The training task took approximately 30 mins.

7.2.4.3 Bi-dimensional training condition. In this condition, participants received feedback that relied both on VOT and F_0 information. For example, a stimulus with a VOT step of three (3) and F_0 of two (2) was treated as "voiced", but a different stimulus with the same VOT value and an F_0 step of three (3) was now considered "unvoiced" (see Figure 7.2.A, for the structure of the feedback in the bi-dimensional training condition). This means that participants were encouraged to use *both* cues when categorizing stimuli and, once VOT and F_0 were taken into account, feedback was perfectly consistent. In terms of the rotated logistic function, this condition aimed at forcing participants to adopt a categorization pattern of a very steep slope (s), but a rotated theta (θ).

Furthermore, this condition was also relevant to one of our questions regarding the relationship between gradiency and secondary cue use; if greater gradiency is a result of greater use of secondary cues, then, if participants in the bi-dimensional training condition learn to use F_0 to a higher degree, they should also show evidence for higher gradiency.

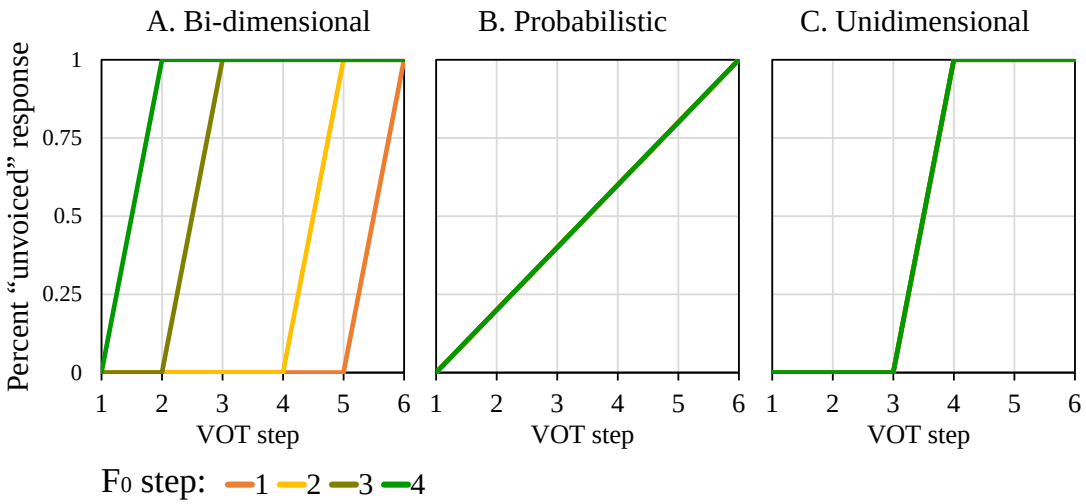


Figure 7.2 Mappings of cue values to phoneme categories for each training condition

7.2.4.4 Probabilistic training condition. In the second training condition, only VOT information was relevant (i.e. the primary cue for distinguishing between voiced and unvoiced stop consonants), while feedback was 100% consistent only for the two extreme VOT steps. In contrast, for VOT steps two (2) and five (5), the feedback was 80% consistent with a mapping to a "voiced" and an "unvoiced" response respectively, while for the two middle steps, it was only 60% consistent (see Figure 7.2.B). This training aimed at encouraging participants to follow a more *probabilistic* cue-to-phoneme mapping strategy, in which intermediate VOTs should be treated as partial evidence for both /b/ and /p/ (to varying degrees). Since the mapping here was inconsistent and F_0 use was not reinforced, we can characterize this condition as a rotated logistic function with a shallow slope (s), but a theta (θ) of 90°.

7.2.4.5 Unidimensional training condition. Finally, in the third condition, participants received feedback that was consistent with a categorization strategy relying exclusively on the VOT step of the stimulus (and in this way it was similar to the probabilistic training), while feedback was 100% consistent across VOT steps (similarly

to the bi-dimensional condition). That is, all stimuli with a VOT step of three (3) or smaller were treated as voiced (i.e. this was the only correct response for this group), while stimuli with VOT steps 4 or higher were treated as unvoiced. Crucially, F_0 information was *irrelevant* in this training condition (i.e. a stimulus with a VOT step of one [1] and an F_0 step of one [1], as well as a stimulus with a VOT step of one [1] and an F_0 step of four [4], were both considered "voiced"), thus encouraging participants to ignore F_0 information altogether (see Figure 7.2.C). This means that this condition was defined as having a very steep slope (*s*), and a theta angle (θ) of 90°.

7.3 Results

Participants performed the tasks without problems, however, careful inspection of the responses revealed that six (6) participants did not do the VAS tasks as instructed[16] and were excluded from the analyses. In total, data from 93 participants were analyzed (30 from the bi-dimensional condition, 31 from the probabilistic, and 32 from the unidimensional condition).

We fitted participants' responses in the pre- and post-VAS tasks separately using the rotated logistic function (see Section 2.1.3). Overall fits were good (R^2 = .98 and R^2 = .95 respectively).

7.3.1 Training task results

Participants overall performed the training task with high accuracy (*M* = 78.6%, *SD* = 7.0%). Specifically, if we focus on accuracy at the last block, the unidimensional

[16] They seemed to be consistently clicking on the exact same point on the line for a run of trials independently of the acoustic characteristics of the stimuli.

group reached the highest accuracy (*M* = 84.6%, *SD* = 7.9%), the bi-dimensional group came in second (*M* = 79.8%, *SD* = 7.3%), while the probabilistic group was last[17] (*M* = 72.4%, *SD* = 5.8%). However, as we see in Figure 7.3, accuracy did not show noticeable improvement with time, suggesting that participants' responses were not affected by our training manipulation.

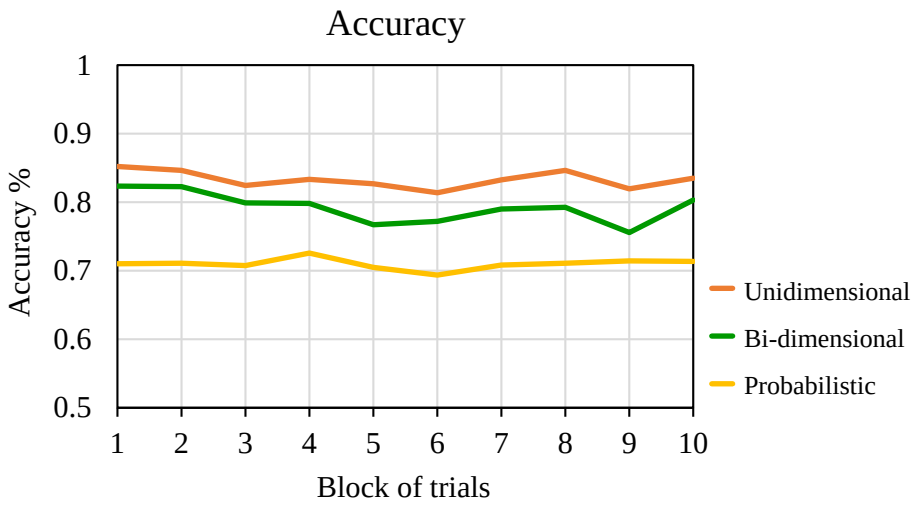


Figure 7.3 Average accuracy in time between training conditions

To test this, we evaluated the effects of VOT, F_0, and time separately for each of the three training groups using a binomial mixed effects model implemented in the *lme4* package (Bates et al., 2009) of R. To keep our analysis simple, we conducted separate analyses for each group, with the primary question of interest being whether the effect of trial block interacted with VOT and F_0. To code time, we grouped trials into five (5) blocks of 144 trials each and this variable was added to the fixed effects. VOT and F_0 were also added as fixed effects after centering and scaling (i.e. dividing each value by

[17] Remember that in the probabilistic condition maximum possible accuracy was ~80% due to the probabilistic nature of the task.

the standard deviation) the raw values. The random effects structure included random intercepts and random VOT and F_0 slopes (and their interaction) for subjects. Our dependent variable was participants' ratings (i.e. /bin/ responses were coded as zero [0] and /p/ responses as one [1])

As expected, there was a significant positive effect of VOT on rating for all three training groups (unidimensional: B = 3.357, z = 16.68, p < .001; bi-dimensional: B = 3.381, z = 16.84, p < .001; probabilistic: B = 2.685, z = 15.63, p < .001). The effect of F_0 was also significant across groups (unidimensional: B = 1.165, z = 11.24, p < .001; bi-dimensional: B = 1.491, z = 12.85, p < .001; probabilistic: B = 1.048, z = 10.85, p < .001). These two effects were strongly expected and show that participants were more likely to categorize stimuli with higher VOT and higher F_0 values as unvoiced.

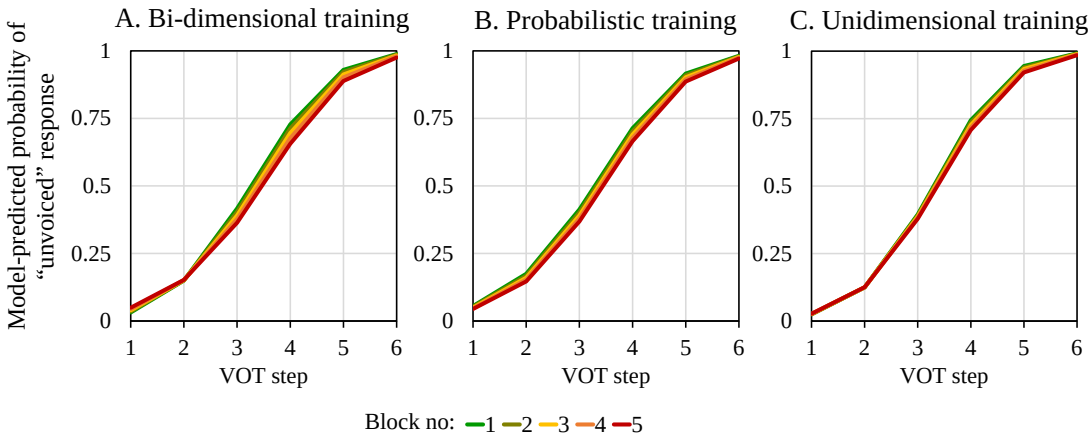


Figure 7.4 Effect of VOT on 2AFC ratings (i.e. training task) per block

Moving on to the effect of time, we found that block number had a significant negative effect across training conditions (unidimensional: z = -2.33, p < .05; bi-dimensional: z = -6.24, p < .001; probabilistic: z = -3.43, p < .001). This indicates that

participants were more likely to categorize any stimulus as unvoiced (rather than voiced) at the beginning of the training (i.e. there was a shift of the category boundary).

Interestingly, we also found that for the unidimensional and bi-dimensional groups there was a significant VOT × block interaction, $z = -2.89$, $p < .01$; $z = -6.65$, $p < .001$. This was marginally significant for the probabilistic group, $z = -1.66$, $p = .098$. This indicates that the effect of VOT on participants' ratings grew weaker (shallower) over training for these groups. Lastly, a significant F_0 × block interaction was also found, but only for the bi-dimensional group, $z = -4.57$, $p < .001$; unexpectedly, this indicates a stronger effect of F_0 on participants' ratings *earlier* in the training (unidimensional: $z < 1$; probabilistic: $z = -1.53$, $p = .126$).

Overall, these results suggest that participants' responses were not modified by our experimental manipulation throughout the course of training.

7.3.2 Pre-training VAS task results

We started our analyses of the VAS data by examining whether VAS slope (i.e. a measure of categorization gradiency) or theta angle (i.e. a measure of secondary cue use) differed significantly between experimental groups *before* our training manipulation (i.e. in the first VAS task). To get at this, we conducted two one-way ANOVAs with VAS slope (s) and theta angle (θ) respectively as the dependent variable and training condition (unidimensional, bi-dimensional, and probabilistic) as the independent variable, including only data from the pre-training VAS task. Training condition was effect-coded in two variables, one comparing the unidimensional (coded as 1) to the bi-dimensional (coded as -1) condition, and another one comparing the unidimensional (coded as 1) to the

probabilistic (coded as -1) condition. The results showed a significant effect of training condition on VAS slope, $F(2,87) = 4.17$, $p < .05$. Post-hoc pairwise comparisons (Bonferroni corrected) showed that the pre-training VAS slope in the unidimensional condition was significantly higher ($M = 1.92$, $SD = .29$) than the bi-dimensional condition ($M = 1.77$, $SD = .15$, $p < .05$), and marginally higher than the probabilistic condition ($M = 1.79$, $SD = .18$, $p = .066$), but the two latter conditions were not significantly different from each other. No significant effect of training condition was found for theta angle, $F(2,87) = 1.16$, $p = .32$.

These results suggest there were some group differences in categorization gradiency between the groups *before* the training. Therefore, to evaluate our primary hypothesis, we decided to include the pre-training VAS slope as a covariate in the main analyses (see Section 7.3.3 next).

7.3.3 Pre- versus post-training VAS results

This set of analyses addressed our primary question; whether our training manipulation led to a significant difference in participants' responses between the first and second VAS task. We did this by first examining the effect of training within a task by comparing pre- and post-test measures; next we evaluated the effect of group by comparing post-test scores between groups.

7.3.3.1 Effect of training on phoneme categorization within subject. We first evaluated the effect of time independently of training condition (i.e. overall difference between pre- and post-training VAS tasks). A paired t-test showed that participants' average boundary (x_0 in Eq.1) significantly shifted from an average of 4.15 ($SD = .63$) to

4.82 (*SD* = .69), t(91) = 8.26, p < .001. Given that participants' boundary in the pre-training VAS task was to the left of the center (i.e. there was a bias towards categorizing ambiguous stimuli as *pin*), this boundary moving to the right suggests that participants' boundaries shifted in a direction consistent with the training.

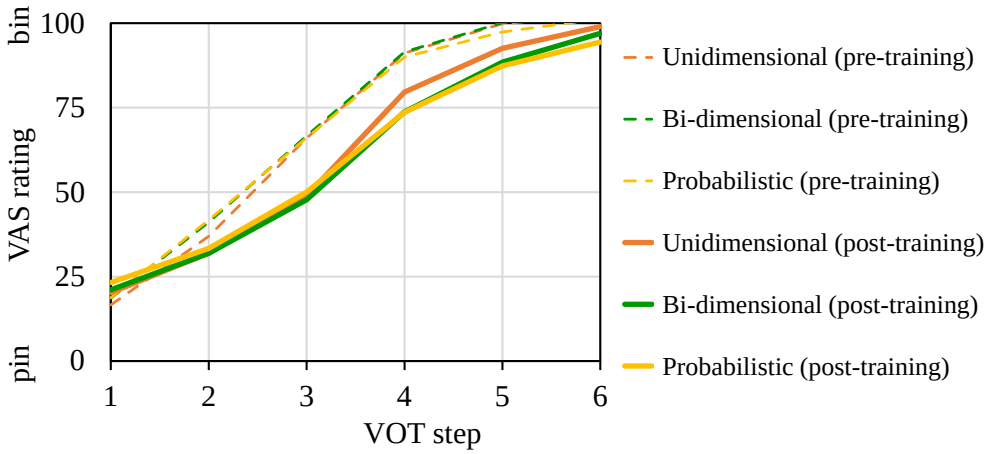


Figure 7.5 Pre- and post-training effect of VOT on VAS ratings

VAS slopes were also significantly different, with overall steeper slopes observed in the pre-training (*M* = 1.85, *SD* = .03) compared to the post-training VAS task (*M* = 1.75, *SD* = .04), t(91) = -3.00, p < .01. This suggests that, across groups, participants' categorization gradiency increased in the post-training task.

Finally, theta angle, was marginally significantly different between the two VAS tasks, with overall higher theta angles in the pre- (*M* = 60.33, *SD* = 7.37) compared to the post-training (*M* = 58.75, *SD* = 10.08), t(91) = -1.82, p = .072. This is consistent with an increase of the use of secondary cue information.

7.3.3.2 Effect of training on phoneme categorization. Next, we moved on to addressing our key question; whether our training manipulation affected the way listeners categorized phonemes. We conducted three analyses of covariance (ANCOVA) with

each of the three VAS parameters (crossover, slope, and theta angle) extracted from the post-training VAS task as the dependent variable in three separate analyses. Training condition was entered as a fixed effect (dummy-coded into two contrasts comparing the first and second condition to the last one), while the corresponding VAS parameters from the pre-training were used as a covariate.

The effect of training condition was not significant for any of the parameters (crossover: $F(2,91) = 1.41$, $p = .25$, $\eta^2 = .03$; slope: $F < 1$; theta angle: $F(2,91) = 1.32$, $p = .27$, $\eta^2 = .03$).

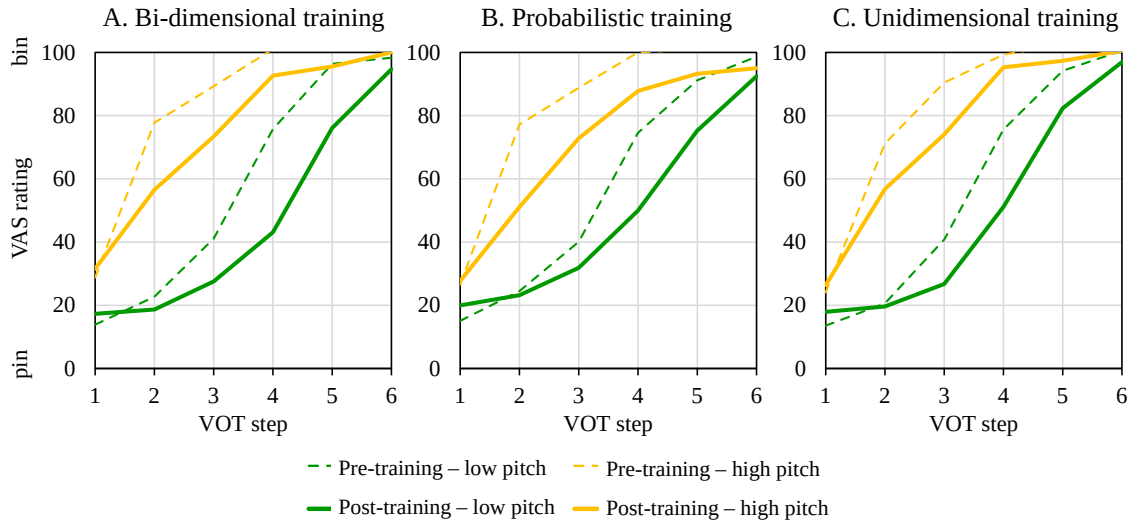


Figure 7.6 Effect of training on VAS ratings per training group

7.4 Discussion

The results from Experiment 4 are not consistent with the hypothesis that we can use feedback-based training to change how listeners map acoustic cue information to phoneme categories. Specifically, our training manipulation did not seem to affect any of the parameters that we extracted from the VAS task.

That said, an interesting –and not predicted– finding of this experiment was that across training conditions, listeners' VAS slope steepness decreased in the post-training VAS task while secondary cue use marginally increased. This could mean that, similarly to Clayards et al. (2008), sole exposure to a variety of VOT and F_0 values may be enough to make listeners more sensitive to differences along these dimensions. Furthermore, this may be telling us that speech perception can in principle be susceptible to change via training, but at the same time may not be particularly sensitive to feedback manipulations – that is, mapping of cues to categories can be shaped by exposure to distributions of cues, but not feedback. We return to this issue in the General Discussion.

CHAPTER 8: GENERAL DISCUSSION

There appears to be a divide in the speech perception literature with respect to the issue of whether phoneme categories are gradient or discrete. On one hand, the idea of categorical perception—which argued for a discrete encoding of the signal—has lost support as a plethora of basic research studies documented that the typical listeners maintain within-category information and use it to activate lexical items in a gradient way. On the other hand, there is a widely accepted idea among researchers studying atypical populations according to which, sharp, step-like categorization of phonemes is the desirable outcome, and any information that does not serve this goal should be treated as noise.

In the middle stands our evidence (along with that of Kong and Edwards) showing robust individual differences among typical listeners; some individuals exhibit a more gradient pattern, while others are more categorical. The primary purposes of this dissertation was to determine what factors give rise to individual differences in phoneme categorization gradiency; and what are the consequences of individual differences in phoneme categorization gradiency for downstream language processes. Ultimately, a more thorough understanding of these issues may allow us to reconcile the two seemingly discrepant accounts of speech perception described above.

In this final chapter, I will briefly review the key results of Experiments 1-4, highlight systematic patterns of results across tasks, and discuss possible interpretations and conclusions we can draw in regard to the specific questions we laid down in Chapter 1, as well as broader insights in regards to the mechanisms underlying speech perception as a whole.

8.1 Brief summary of results

Before turning over to the broader discussion and conclusions, I will list the key findings across experiments. In sum, the present study found evidence in support of the following hypotheses:

- Phoneme categorization gradiency (as measured by the VAS slope) is distinct from internal noise or inconsistency in the cue-to-category mappings (see Experiment 1).
- Phoneme categorization gradiency is positively linked with secondary cue use, but this does not apply to all combination of cues (see Experiments 1, 2, and 3).
- Phoneme categorization gradiency seems to be only weakly linked to aspects of executive function (see Experiments 1 and 2).
- Individual differences in phoneme categorization gradiency seem to stem from differences in the early perceptual encoding of speech cues (see Experiment 2).
- Phoneme categorization gradiency does not seem to be linked to individuals' ability to comprehend speech in noise (see Experiments 1 and 3).
- Phoneme categorization gradiency can be helpful in situations where maintaining competing representations active is desirable (see Experiment 3).
- Phoneme categorization gradiency cannot be modified via brief feedback-based training (see Experiment 4).

Next, I discuss these findings and their significance in regard to the specific theoretical questions addressed here, as well as broader issues in the study of speech perception.

8.2 Measuring phoneme categorization gradiency

In order to address our theoretical questions regarding individual differences in phoneme categorization, we first needed to establish valid measures of different aspects of speech perception. As argued in Chapter 1, the main limitation of the commonly used 2AFC task is that it only allows participants to make binary responses. This means that the steepness of the slope may reflect both gradiency at the mapping between cues and categories, as well as noise in the encoding and/or mapping of the cues to phoneme categories. VAS-based measures (see Kong & Edwards, submitted), offer a way of disentangling these by examining both the shape of the response function and the continuous variation around it. However, other limitations apply to the way this task is currently used. For example, even with continuous responses, it is difficult to estimate the degree of gradiency independently of other aspects of speech perception, such as multiple cue integration.

To address these issues, we developed and evaluated an alternative paradigm for measuring speech categorization gradiency. This paradigm is based on the VAS task, thus allowing us to collect precise measurements of participants' percept on individual trials, but we developed a novel statistical approach that deconfounds gradiency and multiple cue integration.

Interestingly, our findings from Experiment 1b showed that, contrary to a common belief, 2AFC slope is not strongly related to phoneme categorization gradiency (as assessed via our VAS-based method). In addition, inconsistency in the responses (i.e. internal noise) was found to be marginally *negatively* correlated with 2AFC slope, but not significantly correlated with gradiency (even though the direction was negative). This pattern of results, as a whole, suggests that phoneme categorization gradiency is an aspect of speech perception that is theoretically independent of noise and, if anything, gradient categorizers are more likely to exhibit less noise/inconsistency in their responses.

8.3 Phoneme categorization gradiency and multiple cue integration across stimuli

The results across Experiments 1-3 show that phoneme categorization gradiency is a relatively stable aspect of speech perception within individuals that drives (at least in part) categorization patterns across different phoneme contrasts. Table 8.1 presents a summary of the correlations between different phoneme and visual measures of categorization gradiency (VAS slopes).

What is interesting to note here is that the strongest correlations appear between gradiency measures of similar phonemic contrasts (i.e. voicing of labial and alveolar stimuli). In contrast, even though categorization gradiency in a voicing contrast does not seem to robustly predict fricative place of articulation gradiency (see marginal correlation between slopes in Experiment 3), it is still higher than any correlation we found between any measure of phoneme gradiency and an equivalent measure of gradiency from a control visual categorization task (see two right columns of Table 8.1). This supports the specificity of our measure to speech categorization.

At the same time, the marginal correlation we found between measures of speech gradiency for different phonemic contrasts (i.e. voicing in stops versus place of articulation in fricatives) suggests that for the most part gradiency is not a global aspect of speech perception – rather it seems to be closely tied to the acoustic characteristics of the input. This idea ties nicely with our findings from Experiment 2 showing that differences in gradiency are likely to reflect differences in the perceptual encoding of acoustic cues. If this is the case, then it makes sense why gradiency would depend on differences among types of cues.

Table 8.1 Correlations among VAS slopes across Experiments

Experiment (N)	phoneme × phoneme VAS slope		phoneme × visual VAS slope	
	labial × alveolar	labial × fricative	visual × stop	visual × fricative
Experiment 1 (123)	.307**	–	–	–
Experiment 2 (68)	.407**	–	.208 .164	–
Experiment 3 (61)	–	.231^{+}	.205	.107

This dependence of gradiency on acoustic details is also shown in regard to the relationship between individuals' categorization gradiency for voicing categories and the degree to which they use pitch as a secondary cue. As shown in Table 8.2, categorization gradiency for voicing in labials is consistently (i.e. across experiments) related to the degree to which F_0 information is used in voicing judgments, even though this relationship is much weaker in alveolars.

Table 8.2 VAS slope × secondary cue use correlations

Experiment (N)	labial (F_0)	labial (vowel length)	alveolar (F_0)	fricative (transition)
Experiment 1 (122)	.297**	–	.158^{+}	–
Experiment 2 (68)	.346** (.373**)	–	.159 (.167)	–
Experiment 3 (61)	.348** (.371**)	.095 (-.095)		.163 (-.129)

Notes: The direction of the correlations has been adjusted so that positive direction describes a positive relationship between gradiency and secondary cue use; correlations with residualized VAS slope shown in parentheses

To interpret this difference, it may be useful to point out that we consistently observed weaker use of F_0 in alveolars compared to labial stimuli in both experiments (Experiments 1 and 2) that used these two types of stimuli (see Figures 3.6 and 5.3). This could mean that alveolar categorization in general relies less on pitch, or that the range of pitch values we used to construct our stimuli was not the appropriate for alveolars. No matter what the cause of this is, what becomes clear is that whenever pitch information is needed for categorization, the degree to which it is actually used by an individual can be predicted to an extent by that individual's degree of gradiency.

Furthermore, the findings from Experiment 3 are again consistent with the idea that the relationship between gradiency and multiple cue integration is highly dependent on the specific cues. For example, as shown in the last row of Table 8.2, in contrast to pitch information, neither vowel length (for voicing) nor fricative transition seem to be robustly correlated with gradiency.

What these findings show is that figuring out when and how secondary cue use and gradiency are related may be crucial for determining the nature of their relationship. For example, it could be that this link is stronger when secondary cues are processed in

close dependency to primary cues (as may often be the case for VOT and F_0) or when the two manipulated cues really represent a single (integral) cue to the perceptual system.

Related to this issue of the relationship between gradiency and multiple cue integration, is the question of whether secondary cue use is stable within an individual and/or across different speech cues. As shown in Table 8.3, the pattern of correlations is overall positive, but not very robust. This is not surprising, given that, as pointed out earlier, the degree of secondary cue use seems to be highly dependent on stimuli characteristics. Despite this, the overall positive direction suggests that somewhat stable individual characteristics may also play a role in how listeners integrate information across multiple cues.

Table 8.3 Correlations between different types of secondary cue use

	labial F_0 × alveolar F_0	labial F_0 × labial vowel use	labial F_0 × fricative transition	labial vowel use × fricative transition
Experiment 1 (126)	.077	–	–	–
Experiment 2 (69)	(.349**)	–	–	–
Experiment 3 (65)	–	.264*	.260*	.210^{+}

Note: Correlations between theta-based estimations of secondary cue use is shown in parentheses

Overall, what we can conclude from these patterns of correlations is that there are some relatively stable aspects of speech perception processing that differ among individuals and affect how they categorize phonemes. At the same time, however, there are also stimuli characteristics that we need to take into account when assessing different aspects of speech perception at the individual level.

The consistently positive relationship between pitch use and gradiency in the categorization of voicing-defined categories possibly reflects a functional link between

these two. This could mean that higher sensitivity to within-category information (reflected by higher gradiency) also allows for better integration of VOT and pitch information. An alternative hypothesis we laid down in Chapter 1 is that the causal link has the opposite direction: better integration of cues may lead to higher gradiency. In evaluating these two accounts it may be helpful to also consider: 1) the lack of correlations between the use of other secondary cues (e.g. vowel length) and gradiency, and 2) our ERP results from Experiment 2, showing that differences in gradiency are likely due to differences in early perceptual warping.

Specifically, if gradiency relies on the low-level perceptual encoding of cues, then any surface characteristics of the signal are likely to differ substantially in how they are encoded, which may have an effect on gradiency. For example, one possibility is that when two cues are temporally adjacent, or when they are connected via a continuation of frequency energy, they may be perceptually encoded together, or in tight interdependence. In such cases, better perceptual integration of cues may allow for higher gradiency. However, as pointed out by Kingston et al. (2008), perceptual integration of VOT and pitch is unlikely the case, at least in English, given the pause between the offset of the first and the onset of the second. An alternative idea is that the perceptual system may allow for categorically-driven warping of certain cues (e.g. VOT), and not others (e.g. vowel length). Thus, while VOT may be encoded in a more or less warped way among different listeners, vowel length may always be encoded in a veridical way. As a result, multiple cue integration may be trickier when fine-grained information for at least one of the cues has been distorted.

Therefore, together our results seem to support a causal link from gradiency to multiple cue integration, according to which more veridical/gradient encoding of cues allows for better cue integration.

8.4 Sources of phoneme categorization gradiency

One of the key motivations of the present set of experiments was to identify the sources of individual differences in phoneme categorization gradiency. We focused on two broad categories of potential sources: within and outside the language system.

8.4.1 Non-linguistic sources of phoneme categorization gradiency

Our rationale for looking into general cognitive processes was that the way in which we use speech cues to activate phonemes may be to some degree modulated by higher level cognitive processes. For example, working memory could set limits to the extent to which we can keep acoustic information active so it can alter downstream processes; or top-down inhibitory control could allow listeners to suppress competing representations to act more categorical. The counter-argument would be that the kinds of processes that underlie speech perception may be for the most part automatic, without requiring substantial top-down cognitive control.

To test these two possibilities, we assessed the relationship between phoneme categorization processes and several different aspects of executive function, such as inhibitory control (Experiments 1 and 2) and working memory (Experiment 1). The results from these experiments seem to suggest that general cognitive processes play a

weak if any role in speech perception. That said, wherever an effect was found, its direction suggested that better executive function predicted more gradient categorization.

Specifically, Experiment 1 revealed a marginally significant correlation between VAS slope and a measure of executive function tapping primarily into working memory (N-Back). The direction of this correlation suggests that higher working memory capacity predicts shallower VAS slopes (i.e. more gradient categorization). One way to interpret this finding would be that higher working memory allows for better multiple cue integration, which then leads to higher gradiency. However, as argued in the previous section, it is unlikely that multiple cue integration drives gradiency. In addition, the findings of Experiment 1b do not support this possibility, because multiple cue integration was not predicted by N-Back performance. An alternative interpretation is that working memory does not play a role during the categorization process itself, but in *later* stages, when the outcome(s) of that process need(s) to be maintained active.

Furthermore, even though Experiment 1b did not reveal a correlation between inhibitory control (measured by the Flanker task) and gradiency, Experiment 2 did show a significant correlation between inhibitory control (measured by the spatial Stroop task) and gradiency. Interestingly, according to the direction of this correlation, better inhibitory control (i.e. weaker congruency effect) predicted shallower VAS slopes (i.e. more gradient categorization). To interpret this finding, we need to consider what exactly may be reflected by the spatial Stroop congruency effect. The rationale of this task is that the participant needs to suppress/inhibit the wrong option and activate the correct one as quickly as possible. Therefore, a higher congruency effect reveals greater difficulty in switching between competing representations. How could such an effect translate to

speech perception? In this case, it is the different phoneme categories (and/or words) that are the competing representations and, in that sense, switching between them may be more difficult for individuals with a general difficulty in doing so. However, this would mean that this correlation is not due to a direct causal link from inhibitory control to gradiency; in contrast, it is more likely that higher gradiency allows for multiple representations to become partly activated and, in those cases, higher flexibility in switching between them (i.e. better inhibitory control) is necessary for gradiency to be reflected in the listener's response. This would also mean that gradient listeners may be better in recovering from lexical garden paths (as we saw in Experiment 3), but this should also depend on their ability to switch effectively between alternative options.

Table 8.4 Correlations between phoneme categorization gradiency (VAS slopes) and measures of executive function

Experiment (N)	Task	Measure of	r	Direction of relationship btw gradiency and executive function
Experiment 1 (112)	N-Back	working memory	.169^{+}	Positive
Experiment 1 (118)	Trail Making	cognitive flexibility	0.101	(Positive)
Experiment 1 (120)	Flanker	inhibitory control	-0.104	(Negative)
Experiment 2 (68)	Spatial Stroop	inhibitory control	0.337**	Positive
Kong & Edwards (30)	Trail Making	cognitive flexibility	~ 0.4*	Positive

Note: The direction of the correlations has been adjusted to reflect the relationship between executive function and phoneme categorization gradiency such that a positive correlation means that better executive function predicts more gradiency

The foregoing interpretations regarding the roles of working memory and inhibitory control are based on the assumption that different aspects of executive function are linked to gradiency in distinct ways. This idea is consistent with the lack of robust

correlations among the different executive function measures in Experiment 1b. However, if we examine the results cumulatively, across tasks and experiments, including the findings reported by Kong and Edwards (submitted), a relatively consistent pattern pops out; higher executive function usually predicts more gradient categorization (see Table 8.4). Therefore, it remains to be seen whether this overall positive relationship relies on a number of qualitatively distinct links between specific executive function aspects and speech gradiency, or whether it reflects a broader positive relationship that is instantiated in different ways.

In conclusion, when we consider all of the results from the different executive function measures, it appears that the role of executive function in phoneme categorization processes does not appear to be very robust. However, we cannot rule out the possibility that such high-level processes do have a role to play especially in later stages – that is, after phoneme categorization, when a set of representations (phonemes or words) have been activated and the listener needs to either maintain them in memory or flexibly switch between them.

8.4.2 Language-related sources of phoneme categorization gradiency

Another possible source of variability in speech gradiency may be variation in processes that are higher level than speech categorization, but nonetheless within the language system. To test this, we examined whether differences in inter-lexical inhibition may lead to different degrees of sensitivity to within-category differences. Our prediction was that stronger inhibition between words would lead to faster suppression of competitive items (at the word and the phoneme level). This suppression would make it

difficult to maintain alternative interpretations of the signal active, resulting in a more categorical response. However, the results from Experiment 2 (Chapter 5) did not provide evidence for such a link.

This finding, was at first surprising, however, it is in a way consistent with our argument made earlier in regard to the role of executive function; it seems that higher processes, both within and outside the language system, may play a role in handling the *output* of the phoneme categorization processes (what we are measuring with the VAS), but they do not seem to determine how gradient that output is. In contrast, as we discuss shortly, our findings point to a different locus of categorization gradiency at an earlier processing stage.

8.4.3 Perceptual sources of phoneme categorization gradiency

Next, we examined the role of the early perceptual encoding of a primary speech cue (VOT). In particular, our hypothesis was that differences at a somewhat low level of perception are related to the variability we observe in the VAS task. As a measure of VOT encoding we used the amplitude of an early ERP component found to be linearly related with VOT: the N1 (Toscano et al., 2010). The linearity of the relationship between VOT and N1 amplitude has been taken as evidence in support of continuous/gradient perception of acoustic cues, and also documents that it is a useful measure of pre-categorical encoding of acoustic cues (at least VOT).

Thus, our prediction was that, if differences in VAS gradiency are caused by differences in the early encoding of speech cues, this linear pattern should be disrupted in the case of individuals with steeper VAS slopes. Indeed, this is what we found;

individuals with steeper VAS slopes showed a different function linking VOT to N1. Specifically, we found that for those individuals, a hybrid model combining a linear with a *step-function* describes better VOT encoding. In contrast, when we only looked at the brain responses of individuals with more gradient VAS response functions, that step-function of VOT did not explain a significant portion of the N1 variance over and above the purely linear model. It should also be noted that for steep categorizers, the said step-like function was centered at each individual's category boundary – thus reflecting a category-driven warping effect. This provides evidence for the first time that for some individuals, encoding of speech cues may be more strongly affected by category-related information and that the locus of this effect is perceptual. Despite the evidence for a significant effect of a step-function, we did not find evidence to support a pure categorical perception model. Specifically, even for listeners who showed this effect, we found that a hybrid linear/step-function model was a better fit of the data compared to an exclusively step-function model.

In addition to the N1, we also used the P3 ERP component as a marker of later processing, which is thought to reflect categorization rather than early perceptual encoding of acoustic information. In this case, we expected to find a more robust marker of categorization for steep-slope categorizers (i.e. stronger P3). However, what we found was the opposite: a higher amplitude component for the gradient categorizers. In addition, the expected effect of response (i.e. stronger P3 for trials with "target" response") was robust only for gradient categorizers. This finding was again quite surprising, however, one possible interpretation is that for more categorical listeners, categorization is partially performed earlier (via the perceptual warping described earlier). Thus, if the P3 reflects

to some degree the effort of the system to generate a categorical output, then it would make sense that, if the input to this process is already "pre-warped" during the previous perceptual stage, less effort is required.

Given the lack of a robust P3 component, we will build our theoretical discussion around the results from the stronger N1 component. Overall, these results provide invaluable insights into the processes subserving phoneme categorization and speech perception more generally. As argued in the Experiment 2 Discussion (Chapter 5), our findings seem to be consistent with some kind of early perceptual warping of the acoustic space close to the category boundary—for a subset of listeners. In other words, the acoustic input may be distorted during early perceptual stages of processing in a way that between-category differences are amplified. This account, we believe, is quite viable because it is in line with a wide range of behavioral and neuroimaging research findings showing better discrimination of acoustic differences that fall in different phoneme categories (Chang et al., 2010; Dehaene-Lambertz, 1997; Liberman & Harris, 1961; Phillips et al., 2000; Pisoni & Tash, 1974; Repp, 1984; Sams et al., 1990; Schouten & Hessen, 1992).

Critically, however, warping does not mean extinguishing sensitivity to within-category information. That is, the warping we demonstrate is not consistent with strong forms of categorical perception. As it has been demonstrated by a number of studies, typical listeners are sensitive to within-category differences and it has been a challenge in the past to reconcile these studies with the findings showing better between-category discrimination. Our evidence for warping in some listeners may thus offer an integrative account that shows how both of these aspects of perception (better between-category

discrimination *and* sensitivity to within-category differences) can coexist. Listeners can have enhanced discrimination at the boundary without losing the benefits of encoding fine-grained detail. In that way, our findings seem to support a type of model much like that proposed by Pisoni and Tash (1974) in suggesting that listeners use *both* continuous and categorical information. In addition, our results extend this account by showing that the relative strength (i.e. weighing) of each of these two facets of speech processing may differ substantially between individuals.

The question that emerges from these findings is: why do listeners differ in that respect? According to recent findings from neuroscience, there is evidence that speech processing may be served by multiple pathways (Blumstein et al., 2005; Hickok & Poeppel, 2007; Myers & Blumstein, 2009). These routes could correspond to different aspects of perceptual processing. For example, Myers and Blumstein (2009) argue for a distinct role of different brain areas with the inferior frontal gyrus (IFG) being linked to categorical effects, while the superior temporal gyrus (STG) is associated with more continuous processing of the speech input. This kind of dissociation in the role of different areas is one possible source behind the pattern observed here, with evidence for both linear and step-like effects.

Another issue that remains to be addressed is in what way both linear and step-like types of processing are necessary. In other words, what does each of these two aspects of perceptual processing offer to speech perception, and do we need both of them? In addressing these questions, we need to reconsider the different goals of speech perception. One would argue that phoneme categorization is the primary goal of the system. Thus, any kind of process that facilitates the generation of a sharp categorical

output (such as perceptual warping) could be viewed as beneficial. However, at the same time, as argued in Chapter 1, maintaining within-category information could also have significant benefits, for example, in terms of allowing for better integration of multiple cues, or maintaining alternative representations partially active in case they are needed later on. Given the variety of different goals, it seems that the best strategy would be to have a flexible system with both types of processing available and which is able to find the most efficient way of *combining* them in a way that best serves language processing across different situations. We will come back to this issue in the next section, in which we discuss in more detail the conclusions drawn in regard to the consequences of gradiency for language processing.

In conclusion, our findings support an early perceptual locus of differences in categorization gradiency; they seem to be determined by the degree to which the input is warped early on.

8.5 Consequences of phoneme categorization gradiency for language processing

Research on atypical populations seems to favor sharp categorization of phonemes (e.g. Werker & Tees, 1987). The classic logic behind this is that maintaining irrelevant (i.e. within-category) differences is basically maintaining noise – detail that is irrelevant to downstream processes. Under this view, all that matters is that listeners get the right category. Evidence in support of this idea comes mainly from studies using 2AFC tasks, in which individuals with atypical patterns of language processing have been found to have shallower categorization slopes. This is not unexpected given that shallower 2AFC slopes likely stem from inconsistency in the encoding of continuous

cues like VOT. Supporting this assumption, Experiment 1 found a marginally significant negative correlation between the sharpness of participants' 2AFC slopes and our measure of noise/inconsistency in cue-to-phoneme mapping. Crucially, however, this does not mean that higher inconsistency stems from greater gradiency – at least not in the way we define categorization gradiency here, as the sensitivity to within-category differences. In other words, someone who is highly sensitive to such differences could still encode speech cues with fidelity, even as their categories feature a graded mapping. What becomes clear is that prior findings from 2AFC tasks show how noise can be harmful for speech perception (since this noise seems to be related to communicative disorders), but they cannot say much about whether more or less gradient categorization is harmful.

Our VAS measure allows us to address this question more directly. In accordance with our assumption about the dissociation of gradiency from encoding noise, Experiment 1 found that, if at all, VAS slope is negatively correlated with inconsistency in the responses. Having established the validity of our measure, we next examined the role that gradiency may play in downstream speech perception. We looked at two situations in which differences in speech perception processes may offer an advantage: 1) perception of speech in noise, and 2) recovery from lexical garden paths.

8.5.1 Phoneme categorization gradiency and perception of speech in noise

It could be argued that gradiency may be helpful in conditions with high background noise; for example, if information about one cue is missed (or misperceived) due to noise, then a more fine-grained representation of a different cue may prove to be quite helpful. However, in this case we were interested in situations where the noise is

relatively uniformly distributed across the different portions of the input. Therefore, it was perhaps not surprising that we did not find a link between gradiency and listeners' ability to deal with or filter out noise. We saw this when top-down information from the sentence level was available (Experiment 1), but also when listeners were forced to rely more heavily on bottom-up input in the form of isolated words (Experiment 3).

This consistent lack of correlation perhaps speaks to the fact that noise does not correlate in any way with speech-related information. Therefore, it could be argued that maintaining or discarding within-category differences does not change in any way the signal-to-noise ratio. That said, it is possible that different types of noise (for example, in cases where part of the input remains unmasked), may create conditions in which gradiency can have a positive effect.

8.5.2 Phoneme categorization gradiency and recovery from lexical garden paths

In order to evaluate further the role of gradiency in language processing, we presented participants with stimuli that were manipulated to induce lexical garden paths (e.g. *bumpernickel*; see Experiment 3). In this case gradiency did seem to play a role in listeners' ability to deal with ambiguities and temporarily misleading information; however this was only observed in specific aspects of the process.

First, listeners' degree of gradiency was not linked to the likelihood of activating the lexical item that was early on most consistent with the speech signal (i.e. the likelihood of garden-pathing). This null effect may seem counter-intuitive, given our evidence for perceptual warping discussed earlier; one would argue that warping of the speech cues should lead to stronger activation of the category that is most consistent with

the input, which in turn should lead to stronger activation of the corresponding lexical representation, and thus stronger commitment early on. However, these two may not be mutually exclusive. Specifically, it is possible that the perceptual warping does not affect the degree to which the consistent category is activated, as much as it affects the degree to which the other category is suppressed. Following from this, both types of listeners (warpers and non-warpers) may go ahead and activate to a similar degree the phoneme category (and in turn the lexical item) that is most consistent with the signal.

Second, gradiency was also not related with the time it took listeners to recover from that garden path and activate the target word, once they had more information. Once again, this may seem surprising, but it can inform our understanding as to how exactly listeners with more or less gradiency differ from each other. In this case, one could argue that, since the initial commitment to the non-target is similar across listeners, the time needed to suppress should not differ significantly between warpers and non-warpers.

In contrast to these findings, listeners with more gradient phoneme categorization were more likely to recover from the ambiguity at all. This was evident by gradient participants' higher likelihood of looking at the correct item after the point of disambiguation. This finding is consistent with the idea that gradient activation of phonological and lexical representations allows for multiple options to be maintained and considered simultaneously for longer. Thus, when new information arrives that is inconsistent with the initial interpretation of the input (as is the case in lexical garden path situations), listeners with higher gradiency have not fully suppressed alternative options, which makes it easier for them to activate them.

What is particularly intriguing is that only the likelihood of recovery was found to be predicted by gradiency, not the delay of recovery. However, this pattern of results fits quite nicely with the TRACE simulations reported by McMurray et al. (2009); they found that even when TRACE was able to recover from such garden paths (which was only possible when phoneme-level lateral inhibition was completely eliminated), recovery latency was much less affected by VOT step.

At this point, it needs to be clarified that we do not argue that categorization gradiency is beneficial across the board and in all possible situations. However, when the signal is ambiguous, it makes sense why one would want to maintain different items active and not commit too early to one of them. The present study highlights the need to examine the role of gradiency in a variety of different situations in order to determine when and how it can be used in a way that is beneficial for language processing.

8.6 Malleability of phoneme categorization gradiency

Our findings showing how gradiency can be beneficial are theoretically valuable, since they inform our understanding of the role of gradiency in speech perception, but they may also have significant implications in terms of their application. Specifically, once we determine the circumstances in which different degrees of gradiency may be helpful for speech processing, or the manner in which it facilitates perception or language comprehension, we can apply this knowledge to increase language processing efficiency. In addition to this being helpful for typical populations, it may also apply to certain atypical patterns of language processing that are linked to non-optimal degrees of gradiency (see, for example, less sensitivity to between-category differences in a Broca's

patient with a left temporoparietal lesion; Wolmetz, Poeppel, & Rapp, 2011), but may also prove to be helpful in building an alternative route of processing when the primary one has been compromised. Therefore, we next turned to the question of whether the degree to which an individual maintains within-category information is modifiable.

To examine this, we asked participants to categorize a set of labial-initial stimuli varying along two dimensions (VOT and F_0) while we manipulated the feedback they received. The goal of the differential feedback was to reinforce participants to change the way in which they mapped cues to phoneme categories so that by the end of training they would 1) rely both on a primary (VOT) and a secondary cue (F_0), 2) rely exclusively on a primary cue (VOT), or 3) follow a probabilistic cue-to-phoneme mapping approach.

Our results, however, showed no significant difference between our experimental groups in how they performed the VAS task before and after training. This result may not necessarily mean that gradiency is a stable, unmodifiable aspect of the speech perception system – just that our manipulation was not appropriate and/or the duration of the training was not sufficient for such a change to be observed.

In fact, despite the null effect of training condition, we observed that some aspects of performance did change with training – so it is not the case that people simply didn't learn anything; rather they didn't learn to be more or less gradient due to our training manipulations. Specifically, across conditions participants' gradiency increased as did (marginally) the degree to which they used a secondary cue. This shift to a more gradient approach would be consistent with a distributional-based kind of perceptual learning similar to that of Clayards et al. (2008), who reported evidence for a training-induced modification of phoneme categorization via the manipulation of the probability

distributions of VOT. Critically, in contrast to the more normal-like distributions used by Clayards et al, in our case, each cue value appeared the exact same number of times (and this was the case across training conditions). Therefore, listeners could have adjusted to the distributional statistics of our stimuli. According to this account, speech perception is more susceptible to a manipulation of the distributional characteristics of the input rather than feedback. Moreover, this idea makes a lot of sense if we consider how we learn language in natural conditions: through mere exposure. In other words, a different, more passive kind of training may be more compatible with the way in which the language system has been originally shaped and adjusted over the course of development.

Alternatively, our feedback manipulation may have put participants into a state of high uncertainty; they perform a task of categorizing *bin* and *pin*, which is something you may expect a native listener to perform 100% accurately. Despite this, due to our experimental manipulation, a perfect score was extremely difficult, while in the case of the probabilistic training condition, it was impossible. Listeners may respond to this uncertainty by becoming more gradient. However, this latter account seems less viable given that if this were the case we would expect to find a stronger increase of gradiency in the probabilistic training condition.

Overall, since Experiment 4 was not designed to manipulate neither the distributional characteristics of the speech cues, nor the degree of uncertainty, all we can conclude for now is that brief, feedback-based training does not seem to have an effect on gradiency. Therefore, it remains to be determined what the exact conditions are that can lead listeners to change their pattern of phoneme categorization and in what ways.

8.7 Overarching conclusions and future work

Our findings provide invaluable insights into the mechanisms that underlie speech perception and bridge together seemingly contrastive views of phoneme categorization gradiency. Specifically, our results show that even though the system is fundamentally sensitive to within-category differences, at the same time there are also substantial individual differences in regard to how bottom-up (acoustic) and top-down (categorical) sources of information are weighed. In addition, when looking at the findings from the different experiments together, the emerging pattern supports the idea that higher gradiency allows for better integration of multiple cues. This finding holds great theoretical value, as it informs our understanding of how these two aspects of speech perception are linked to each other. Lastly, despite gradiency being commonly considered detrimental for speech perception, here we show that there is at least one case in which the opposite seems to be the case.

As an exploratory endeavor, this work was largely correlational and, thus, cannot speak definitively to the causal links between the different facets of speech perception and other processes, as well as to their role in language comprehension. However, our findings do reveal informative patterns of correlations that can help us evaluate contrastive accounts and can be used as a basis for future work. Among the various directions that can be taken, we believe that one of the most critical aims of future research will be to validate our findings on the perceptual warping effect, determine its neural substrate, and examine whether and how it may vary among typical and atypical populations.

Another critical issue is how the system learns to be more or less gradient. In this direction, it would be useful to study the differences in phoneme categorization patterns across populations that are known to have different patterns of experience with language (e.g. bilinguals) and test hypotheses as to how exposure to different conditions may affect speech processing in the long term. Furthermore, we need to strive for a more comprehensive description of the mechanisms that underlie different patterns of speech perception using not only correlational and experimental paradigms, but also computational tools that allow us to manipulate different aspects and parameters of the system in a more precise and systematic way.

In conclusion, our results speak to the flexibility of the speech perception system in using both bottom-up and top-down sources of information. It is up to us to show a similar sense of flexibility in its study that will allow us to better understand the cognitive and neural mechanisms that underlie it.

REFERENCES

Abramson, A., & Lisker, L. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word*, *20*(3), 384–422.

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), 2031–9.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*(3), 163–87.

Apfelbaum, K., Blumstein, S., & McMurray, B. (2011). Semantic priming is affected by real-time phonological competition: evidence for continuous cascading systems. *Psychonomic Bulletin & Review*, *18*(1), 141–9.

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.

Bates, D., Maechler, M., & Dai, B. (2009). Lme4: linear mixed-effects models using s4 classes. hhttp. *CRAN. R-Project. org/package¼lme4i*.

Blomert, L., & Mitterer, H. (2004). The fragile nature of the speech-perception deficit in dyslexia: natural vs synthetic speech. *Brain and Language*, *89*(1), 21–6.

Blumstein, S. E., Myers, E. B., & Rissman, J. (2005). The perception of voice onset time: an fmri investigation of phonetic category structure. *Journal of Cognitive Neuroscience*, *17*(9), 1353–66.

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

Bogliotti, C., Serniclaes, W., Messaoud-Galusi, S., & Sprenger-Charolles, L. (2008). Discrimination of speech sounds by children with dyslexia: comparisons with chronological age and reading level controls. *Journal of Experimental Child Psychology*, *101*(2), 137–55.

Burke, D., & Shafto, M. (2008). Language and aging. *The Handbook of Aging and Cognition*.

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

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., & Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature Neuroscience*, *13*(11), 1428–1432.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of

speech reflects optimal use of probabilistic speech cues. *Society*, *108*, 804–809.

Coady, J. A., Evans, J. L., Mainela-Arnold, E., & Kluender, K. R. (2007). Children with specific language impairments perceive speech most categorically when tokens are natural and meaningful. *Journal of Speech, Language, and Hearing Research*, *50*(1), 41–57.

Coady, J. A., Kluender, K. R., & Evans, J. L. (2005). Categorical perception of speech by children with specific language impairments. *Journal of Speech, Language, and Hearing Research*, *48*(4), 944–59.

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.

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: evidence for lexical competition. *Language and Cognitive Processes*, *16*(5-6), 507–534.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology. General*, *134*(2), 222–41.

Dehaene-Lambertz, G. (1997). Electrophysiological correlates of categorical phoneme perception in adults. *NeuroReport*, *8*(4), 919 –24.

Du, Y., Buchsbaum, B. R., Grady, C. L., & Alain, C. (2014). Noise differentially impacts phoneme representations in the auditory and speech motor systems. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(19), 7126–31.

Elman, J., & McClelland, J. (1988). Cognitive penetration of the mechanisms of perception: compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, *27*(2), 143–165.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Farris-Trimble, A., & McMurray, B. (2013). Test–retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1328–1345.

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), 308–27.

Frye, R., Fisher, J., & Coty, A. (2007). Linear coding of voice onset time. *Journal of Cognitive Neuroscience*, *19*(9), 1476–1487.

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

Godfrey, J., & Syrdal-Lasky, K. (1981). Performance of dyslexic children on speech perception tests. *Journal of Experimental Psychology*, *32*(3), 401–424.

Goldinger, S. D. (1998). Echoes of echoes? an episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: evidence from tongue twisters. *Language and Cognitive Processes*, *21*(6), 649–683.

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

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, (8.5), 393–402.

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

Hillenbrand, J. M., Getty, L. A., Wheeler, K., & Clark, M. J. (1995). Acoustic characteristics of american english vowels. *The Journal of the Acoustical Society of America*, *95*(5), 2875–2875.

Jensen, A. R., & Rohwer, W. D. (1966). The stroop color-word test: a review. *Acta Psychologica*, *25*, 36–93.

Joanisse, M. F., Manis, F. R., Keating, P., & Seidenberg, M. S. (2000). Language deficits in dyslexic children: speech perception, phonology, and morphology. *Journal of Experimental Child Psychology*, *77*(1), 30–60.

Kapnoula, E., & McMurray, B. (2016a). Training alters the resolution of lexical interference: evidence for plasticity of competition and inhibition. *Journal of Experimental Psychology: General*, *145*(1), 8–30.

Kapnoula, E., & McMurray, B. (2016b). Newly learned word-forms are abstract and integrated immediately after acquisition. *Psychonomic Bulletin and Review*, *23*(2), 491–499.

Kapnoula, E., Packard, S., Gupta, P., & McMurray, B. (2015). Immediate lexical integration of novel word forms. *Cognition*, *134*, 85–99.

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.

Kirchner, W. (1958). Age differences in short-term retention of rapidly changing information. *Journal of Experimental Psychology*, *55*(4), 352–8.

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. (submitted). Individual differences in categorical perception of speech: cue weighting and executive function. *Journal of Phonetics*.

Kong, E. J., & Edwards, J. (2011). Individual differences in speech perception: evidence from visual analogue scaling and eye-tracking. In *Proceedings of the XVIIth International Congress of Phonetic Sciences*. Hong Kong.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–8.

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2013). Lmertest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package). *R Package Version 1.0-2*.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: when adults learn new words. *Cognitive Psychology*, *55*(4), 306–53.

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*, 379–88.

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.

López-Zamora, M., Luque, J. L., Álvarez, C. J., & Cobos, P. L. (2012). Individual differences in categorical perception are related to sublexical/phonological processing in reading. *Scientific Studies of Reading*, *16*(5), 443–456.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: a tale of two systems? *Cognitive Science*, *27*(5), 801–805.

Mahr, T., McMillan, B. T. M., Saffran, J. R., Ellis Weismer, S., & Edwards, J. (2015). Anticipatory coarticulation facilitates word recognition in toddlers. *Cognition*, *142*, 345–50.

Maiste, A. C., Wiens, A. S., Hunt, M. J., Scherg, M., & Picton, T. W. (1995). Event-related potentials and the categorical perception of speech sounds. *Ear and Hearing*, *16*(1), 68–90.

Marslen-Wilson, W. D., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychological Review*, *101*(4), 653–675.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: a new test. *Speech Communication*, *2*(1), 15–35.

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

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*(8), 363–9.

McMurray, B., & Aslin, R. (2004). Anticipatory eye movements reveal infants' auditory and visual categories. *Infancy*, *6*(2), 203–229.

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., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, *15*(6), 1064–71.

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). Oxford, UK.

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.

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), 219–46.

McMurray, B., & Jongman, A. (2015). What comes after /f/? prediction in speech derives from data-explanatory processes. *Psychological Science*.

McMurray, B., Munson, C., & Tomblin, J. B. (2014). Individual differences in language ability are related to variation in word recognition, not speech perception: evidence from eye movements. *Journal of Speech, Language, and Hearing Research : JSLHR*, *57*(4), 1344–62.

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.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category vot affects recovery from lexical garden-paths: evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*(1), 132–158.

Messaoud-Galusi, S., Hazan, V., & Rosen, S. (2011). Investigating speech perception in children with dyslexia: is there evidence of a consistent deficit in individuals? *Journal of Speech, Language, and Hearing Research : JSLHR*, *54*(6), 1682–701.

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.

Moberly, A. C., Lowenstein, J. H., & Nittrouer, S. Word recognition variability with cochlear implants: “perceptual attention” versus “auditory sensitivity”. *Ear and Hearing*, *37*(1), 14–26.

Munson, B., & Carlson, K. U. (submitted). An exploration of methods for rating children's productions of sibilant fricatives. *Speech, Language, and Hearing*.

Myers, E., & Blumstein, S. (2009). Inferior frontal regions underlie the perception of phonetic category invariance. *Psychological Science*, *20*(7), 895–903.

Näätänen, R., Teder, W., Alho, K., & Lavikainen, J. (1992). Auditory attention and selective input modulation: a topographical erp study. *Neuroreport*, *3*(6), 493–6.

Nasman, V. T., & Rosenfeld, J. P. (1990). Parietal p3 response as an indicator of stimulus categorization: increased p3 amplitude to categorically deviant target and nontarget stimuli. *Psychophysiology*, *27*(3), 338–350.

Nearey, T., & Hogan, J. (1986). Phonological contrast in experimental phonetics: relating distributions of production data to perceptual categorization curves. In J. J. Ohala & J. J. J (Eds.), *Experimental Phonology* (pp. 141–146). Orlando, FL: Academic Press.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: feedback is never necessary. *The Behavioral and Brain Sciences*, *23*(3), 299–325; discussion 325–70.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–91.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1718–25.

Ojemann, G., Ojemann, J., Lettich, E., & Berger, M. (1989). Cortical language localization in left, dominant hemisphere: an electrical stimulation mapping investigation in 117 patients. *Journal of Neurosurgery*, *71*(3), 316–326.

Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M. ., & Roberts, T. (2000). Auditory cortex accesses phonological categories: an meg mismatch study. *Journal of Cognitive Neuroscience*, *12*(6), 1038–1055.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *The Journal of the Acoustical Society of America*, *55*(2), 328–33.

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

Polich, J., & Criado, J. R. (2006). Neuropsychology and neuropharmacology of p3a and p3b. *International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology*, *60*(2), 172–85.

Repp, B. (1982). Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*(1), 81.

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

Robertson, E., Joanisse, M., Desroches, A., & Ng, S. (2009). Categorical speech perception deficits distinguish language and reading impairments in children. *Developmental Science*, *12*(5), 753–67.

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.

Sams, M., Aulanko, R., Aaltonen, O., & Näätänen, R. (1990). Event-related potentials to infrequent changes in synthesized phonetic stimuli. *Journal of Cognitive Neuroscience*, *2*(4), 344–357.

Samuel, A. G. (2001). Knowing a word affects the fundamental perception of the sounds within it. *Psychological Science*, *12*(4), 348–51.

Schellinger, S., Edwards, J., Munson, B., & Beckman, M. E. (2008). Assessment of children's speech production 1: transcription categories and listener expectations. poster presented at the. In *ASHA Convention*. Chicago, IL.

Schouten, B., Gerrits, E., & Hessen, A. van. (2003). The end of categorical perception as we know it. *Speech Communication*, *41*(1), 71–80.

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

Scott, S. K. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400–2406.

Serniclaes, W. (2006). Allophonic perception in developmental dyslexia: origin, reliability and implications of the categorical perception deficit. *Written Language & Literacy*, *9*(1), 135–152.

Serniclaes, W., & Sprenger-Charolles, L. (2001). Perceptual discrimination of speech sounds in developmental dyslexia. *Journal of Speech, Language, and Hearing Research*, *44*(2), 384–399.

Serniclaes, W., Van Heghe, S., Mousty, P., Carré, R., & Sprenger-Charolles, L. (2004). Allophonic mode of speech perception in dyslexia. *Journal of Experimental Child Psychology*, *87*(4), 336–61.

Sharma, A., & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *The Journal of the Acoustical Society of America*, *106*(2), 1078–83.

Sharma, A., Marsh, C. M., & Dorman, M. F. (2000). Relationship between n1 evoked potential morphology and the perception of voicing. *The Journal of the Acoustical Society of America*, *108*(6), 3030–5.

Shum, D. H. K., McFarland, K. A., & Bain, J. D. (1990). Construct validity of eight tests of attention: comparison of normal and closed head injured samples. *Clinical Neuropsychologist*, *4*(2), 151–162.

Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., & Cook, S. (2012). Development and validation of the azbio sentence lists. *Ear and Hearing*, *33*(1), 112–7.

Streeter, L., & Nigro, G. (1979). The role of medial consonant transitions in word perception. *The Journal of the Acoustical Society of America*, *65*(6), 1533–1541.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643–662.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology. Human Perception and Performance*, *7*(5), 1074–95.

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *The Journal of the Acoustical Society of America*, *62*(2), 435.

Sussman, J. (1992). Perception of formant transition cues to place of articulation in children with language impairments. *Journal of Speech, Language, and Hearing Research*, *36*(6), 1286–99.

Tombaugh, T. N. (2004). Trail making test a and b: normative data stratified by age and education. *Archives of Clinical Neuropsychology : The Official Journal of the National Academy of Neuropsychologists*, *19*(2), 203–14.

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. (2012). Cue-integration and context effects in speech: evidence against speaking-rate normalization. *Attention, Perception & Psychophysics*, *74*(6), 1284–301.

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), 1532–40.

Urberg-Carlson, K., Kaiser, E., & Munson, B. (2008). Assessment of children's speech production 2: testing gradient measures of children's productions. poster presented at the. In *ASHA Convention* (pp. 20–22). Chicago.

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.

Warren, P., & Marslen-Wilson, W. D. (1987). Continuous uptake of acoustic cues in spoken word recognition. *Perception & Psychophysics*, *41*(3), 262–175.

Werker, J. F., & Tees, R. C. (1987). Speech perception in severely disabled and average reading children. *Canadian Journal of Psychology*, *41*(1), 48–61.

Whalen, D. H. (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Perception & Psychophysics*, *35*(1), 49–64.

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2013). Roles of voice onset time and f0 in stop consonant voicing perception: effects of masking noise and low-pass filtering. *Journal of Speech, Language, and Hearing Research : JSLHR*, *56*(4), 1097–107.

Winn, M. B., & Litovsky, R. Y. (2015). Using speech sounds to test functional spectral resolution in listeners with cochlear implants. *The Journal of the Acoustical Society of America*, *137*(3), 1430–42.

Wolmetz, M., Poeppel, D., & Rapp, B. (2011). What does the right hemisphere know

about phoneme categories? *Journal of Cognitive Neuroscience*, *23*(3), 552–69.

Wu, Y., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (submitted). Psychometric functions of dual-task paradigms for measuring listening effort. *Ear and Hearing*.

Wühr, P. (2007). A stroop effect for spatial orientation. *The Journal of General Psychology*, *134*(3), 285–94.

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.

Zekveld, A., & Kramer, S. (2014). Cognitive processing load across a wide range of listening conditions: insights from pupillometry. *Psychophysiology*, *51*(3), 277–284.

APPENDIX

Table A.1 Word Pairs Used in Lexical Inhibition Task in Experiment 2

Target word	Competitor
bait	bake
bat	back
bride	bribe
bug	bud
carp	cart
cat	cap
chick	chip
dart	dark
dot	dock
fork	fort
grad	grab
heap	heat
hub	hug
job	jog
knot	knock
leap	leak
mug	mud
net	neck
part	park
pick	pit
pope	poke
rod	rob
shake	shape
steak	state
suit	soup
tarp	tart
web	wed
zip	zit

Table A.2 List of Images Used in Lexical Inhibition Task in Experiment 2

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

Table A.3 Triplets (in IPA) Used in Lexical Inhibition Task in Experiment 2

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