Attention, Perception, & Psychophysics (2019) 81:1127–1146
https://doi.org/10.3758/s13414-019-01720-9

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

Regressive spectral assimilation bias in speech perception

Amanda Rysling[1] · Alexandra Jesse[2] · John Kingston[3]

Published online: 21 May 2019

Abstract

Speech perception presents a parsing problem: construing information from the acoustic input we receive as evidence for the speech sounds we recognize as language. Most work on segmental perception has focused on how listeners use differences between successive speech sounds to solve this problem. Prominent models either assume (a) that listeners attribute acoustics to the sounds whose articulation created them, or (b) that the auditory system exaggerates the changes in the auditory quality of the incoming speech signal. Both approaches predict contrast effects in that listeners will usually judge two successive phones to be distinct from each other. Few studies have examined cases in which listeners hear two sounds in a row as similar, apparently failing to differentiate them. We examine such under-studied cases. In a series of experiments, listeners were faced with ambiguity about the identity of the first of two successive phones. Listeners consistently heard the first sound as spectrally similar to the second sound in a manner suggesting that they construed the transitions between the two as evidence about the identity of the first. In these and previously reported studies, they seemed to default to this construal when the signal was not sufficiently informative for them to do otherwise. These effects go unaccounted for in the two prominent models of speech perception, but they parallel known domain-general effects in perceptual processing, and as such are likely a consequence of the structure of the human auditory system.

Keywords Speech perception · Psycholinguistics · Grouping and segmentation

In order to comprehend the speech that they hear, listeners must solve a parsing problem: they must successfully construe information in the acoustic input that they hear as evidence for segments and words that they recognize as language. Solving this parsing problem is complicated by fact that successive speech sounds are *coarticulated* with each other. Coarticulation results from the unavoidable mutual influence and overlap of the gestures that produce successive intended speech sounds. As a result, any two speech sounds are more acoustically similar to each other when they are adjacent than they otherwise would be next to other speech sounds or in isolation. Furthermore, the articulation of speech sounds in unstressed syllables, fast, or conversational speech is often under-realized, such that gestures are not fully executed compared to their extents in careful or clear speech. Faced with these myriad challenges, work on speech perception to date has attempted to account for listeners' parsing success by understanding how they use differences between successive speech sounds to determine those sounds' identities.

Gesturalism vs. auditorism

Previous theories of speech perception fall into two prominent types, (i) gesturalist and (ii) auditorist approaches (for a more in-depth and historically detailed discussion, see Diehl, Lotto, & Holt, 2004). Gesturalist approaches (Fowler, 1986, 2006; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985a, b) hold that speech perception is the result of listeners' mapping from the acoustics of the speech signal to the articulatory gestures

✉ Amanda Rysling
rysling@ucsc.edu

[1] Department of Linguistics, University of California Santa Cruz, Santa Cruz, CA, USA

[2] Department of Psychological and Brain Sciences, University of Massachusetts Amherst, Amherst, MA, USA

[3] Department of Linguistics, University of Massachusetts Amherst, Amherst, MA, USA

that produced each apparent speech sound. These gestures are understood to be the atomic units of speech perception, and so the process of speech perception necessarily uses language-specific representations from the earliest moments of that process. In contrast to this, auditorist approaches (Diehl & Kluender, 1989; Lotto & Holt, 2006, 2015; Lotto & Kluender, 1998) hold that general mechanisms of the auditory system work at the initial stages of perception, at first treating the acoustics of speech in the same way that they treat all other acoustic stimuli. The listener's first representations of the speech signal are thus general auditory qualities, which are then mapped to representations of speech sounds as the evidence for each speech sound is identified. However, despite their differences, both gesturalist and auditorist approaches have to date framed the parsing problem in speech perception as one of leveraging differences between the speech sound that a listener is attempting to identify (i.e., the target sound) and its surrounding sounds (i.e., the context).

Perhaps the best-known example of this difference-leveraging phenomenon in the laboratory comes from Mann's (1980) demonstration that listeners categorize intermediate, ambiguous steps from a /da-ga/ continuum more often as "ga" after /al/ than after /ar/. Both gesturalist and auditorist accounts of this finding rely on listeners leveraging differences between the target sounds from the /da-ga/ continuum and the contexts, /al/ or /ar/.

Gesturalist theories have characterized parsing speech as a problem of compensating for coarticulation, whereby the listener separates the acoustic properties that are due to the articulatory gestures of a target sound from those that are due to the articulatory gestures of the sounds that surround it. With respect to Mann's findings, coarticulation with /al/ makes the target syllable gesturally and acoustically more like /da/, while coarticulation with /ar/ makes it more like /ga/ instead. Successful compensation for coarticulation undoes these context effects and renders an otherwise ambiguous target more /ga/-like after /al/ and more /da/-like after /ar/.

Auditorist theories have characterized parsing as a problem of detecting differences between successive intervals in speech. They account for Mann's findings as an effect of spectral contrast between the target and its context: next to spectrally high /al/, an ambiguous target's auditory quality is heard as spectrally low and thus more like /ga/, while next to spectrally low /ar/, the same ambiguous target's auditory quality is instead heard as spectrally high and thus more like /da/. According to auditorist theories, this immediate exaggeration of the spectral differences between target speech sounds and their contexts is not specific to speech, but would arise in the auditory response to any sequence of sounds that differ spectrally.

This auditorist explanation of Mann's results makes two predictions. Firstly, it predicts that listeners' responses to target continua that are differentiated by spectral weight should evince spectral contrast effects both when contexts are vowels and targets are consonants, and when contexts are consonants and targets are vowels. This is because the effects arise from spectral properties that any kind of speech sound can have, not from particular qualities of only one class of speech sound. Indeed, spectral contrast effects have been found for both scenarios (Holt, 1999; Holt, Lotto, & Kluender, 2000), and were replicated with different speech sounds in the present paper. Secondly, this explanation of Mann's effect predicts that listeners' responses to speech continua such as /d/-to-/g/ should also be affected by nonspeech contexts, such that a spectrally high pure tone context should bias listeners to answer with a spectrally low "g" response, and a spectrally low pure tone context should bias listeners to answer with a spectrally high "d" response to the same ambiguous middle continuum steps. In line with this prediction, Lotto and Kluender (1998) found spectrally contrastive responses to a target /d/-to-/g/ continuum after context spectrally high versus low pure tones. Gesturalist approaches can only account for such interactions between speech targets and nonspeech contexts by assuming different mechanisms from the compensation for coarticulation one that they posit to be active on all-speech stimuli (Fowler, Brown, & Mann, 2000; Viswanathan, Fowler, & Magnuson, 2009).

Under the auditorist approach to speech perception, descriptions of effects are separate from the mechanistic causes of those effects. Contrastive patterns of responses, by which listeners judge target speech sounds to differ from their contexts, can and should arise via a variety of different underlying mechanisms in the auditory system or higher-level processing (see Lotto & Holt, 2006, 2015, for more detailed discussions). This can be understood in terms of Marr's (1982) levels of explanation. What can be described at the computational level as "spectral contrast effects"—behavioral response patterns of saying that a target speech sound is spectrally different from its context—can, and likely do, have different algorithmic- or implementational-level bases. At the algorithmic level, the processes underlying spectral contrast effects could be ones that either are posited to occur immediately upon hearing a context, as a reaction in the listener's auditory system creates a bias for an immediately opposite-valency response to an upcoming new stimulus, or they can be processes that are posited to occur well after all of a context, target, and following context speech sounds are heard, as the listener retrospectively implicitly judges the target to be different from its precursor. At the implementational level, different physiological bases at different points in the auditory pathway have been proposed for contrastive response patterns to spectrally varying stimuli (see, for example, Holt, 2005, 2006a, b; Kiefte & Kluender, 2008;

Sjerps, Mitterer, & McQueen, 2011, 2012, 2013; Stilp & Anderson, 2014; Stilp, Anderson, & Winn, 2015; Stilp & Assgari, 2018; Watkins, 1991; Watkins & Makin, 1994, 1996, inter alia).

All auditorist accounts of spectral contrast effects that characterize spectral contrast between two speech sounds as the immediate product of continuous change detection make a directional prediction: the percept of a target speech sound will only be affected by a context that *precedes* the target, because a target speech sound is only a change relative to its preceding context. According to such explanations, the effect of a *following* context on a preceding target could be contrastive, but this would necessarily arise from a different underlying mechanism, just as a different underlying mechanism must be posited to account for contrast effects based on duration rather than spectral weight (Diehl & Walsh, 1989).

It is thus the case that both of the two prominent approaches to speech perception take it as a default assumption that listeners should hear a target speech sound as different from its context speech sounds. In addition, Lotto and Holt (2006) acknowledge the existence of assimilation effects, that is, of cases when listeners judge target sounds as similar to their contexts. However, these assimilation effects (Fujimura, Macchi, & Streeter, 1978; Hura, Lindblom, & Diehl, 1992; Kingston & Shinya, 2003; Repp, 1983; Wade & Holt, 2005) are comparatively far fewer than contrast effects in the literature, and no explicit account of them has been offered in auditorist terms. Meanwhile, certain gesturalist accounts have explicitly argued that such assimilations should not take place under normal speech parsing conditions. For example, Fowler (2006) has argued that compensation for coarticulation (producing contrastive effects) acts on all adjacent speech sounds, and furthermore, independent of direction.

The assumption of correctness in compensation for coarticulation

Yet another type of account explicitly argues against the existence of assimilation effects in clear speech signals. Ohala (1981) frames the parsing problem as one of correctly attributing acoustic consequences to the speech segments that caused them. Under this account, listeners will correctly disentangle coarticulated speech when they are able to implicitly reason about which segments are the sources of the acoustic properties observed in the signal. Listeners would only hear a target sound as similar to its context on a relevant dimension when acoustic information about the context is somehow impoverished, for example by being masked with noise, being realized too quietly, or being articulated with ineffectual gestures. This deprecation of contextual information would fail to provide evidence about which context segment's acoustics had blended with the target's. This, in turn, would prevent listeners from successfully reconstructing a target's intended properties, and so lead to "misparsing," by which a listener would conclude that the target sound was supposed to be the option that was more similar to its context. Ohala thus has in common with gesturalist approaches the assumption that there is a correct parse, and that this is the one that listeners will reach if at all possible. Ohala for this reason predicts that assimilation effects only occur when the context information is unclear. In all other situations, listeners will arrive at the correct parse, which attributes acoustic consequences to the segments that were their sources.

Assimilation effects to date

To our knowledge, an exhaustive list of the studies reporting spectral assimilation effects is as follows: Aravamudhan, Lotto, and Hawks (2008); Fujimura et al. (1978); Hura et al. (1992); Kingston and Shinya (2003); Mitterer (2006); Repp (1983); Sjerps et al. (2012); Wade and Holt (2005). Inspection of these results reveals an intriguing commonality: in the majority of cases, listeners judge target sounds that immediately *precede* their contexts to be similar to those contexts. Fujimura et al. investigated listeners' judgments of the identity of a consonant in vowel-consonant-vowel sequences (VCV). When the spectral information provided by the transition from the first vowel into the consonant (VC) mismatched the spectral information provided by the transition from the consonant into the second vowel (CV), listeners' judgments were consistent with the second transition interval. That is, they assimilated the consonant to its following context, not its preceding one, and this could not be explained by a spectral contrast effect with the preceding vowel. Hura et al.'s, Kingston and Shinya's, and Repp's results may be understood as a consonant-consonant version of target-context assimilation: listeners most often judged the first of two consonants in a sequence to be spectrally similar to the second consonant. Taken together, these studies suggest that assimilation is most likely to occur whenever a target speech sound precedes its context.

In addition, Wade and Holt's studies further suggest that such assimilation only occurs between a preceding target and a following context if the interval between them is sufficiently continuous. Spectral assimilation between target speech sounds drawn from a spectrally high-to-low /d/-to-/g/ continuum and a following nonspeech tone was only observed when that tone could be heard as continuous with the formant transitions out of the preceding stop. However, when that tone began 50 ms later, contrast effects occurred. This delay provided a long enough interval of

discontinuity between the transitions' spectra and the tone's spectral weight that the listener no longer perceptually grouped the tone with the transitions, and so did not use the spectral weight of the tone to inform her decision about the quality of the transitions.

The present account

In this paper, we test the hypothesis that spectral assimilation is the result of a perceptual default to group together sufficiently continuous spectral information. We characterize this as a default in order to capture the intended prediction that this tendency will only determine a listener's response when no other information can be brought to bear on the identification of a target speech sound. Stated in Bayesian terms, we argue that spectral assimilation in conditions of relatively more ambiguity in the speech signal reveals the action of the speech perception system's prior in conditions of low evidence. The present paper thus advances the computational-level proposal that spectral assimilation is better understood as a default behavior, not a simple failure to accomplish compensation for coarticulation. It further advances a preliminary algorithmic-level proposal that this spectrally assimilative pattern of responses is due to domain-general time-distance grouping effects.

When the acoustic stream contains a spectrally unambiguous context sound before any ambiguous sounds, a change can be marked or discovered relative to that standard, and so the judgment of a following target sound's more ambiguous spectra can be affected by its precursor. However, when an ambiguous target has no such precursor, there is at first only uncertainty about the target's quality. If the change from that ambiguity is sufficiently continuous or gradual, then it is possible to group following information with that initial ambiguity. This grouping could lead the listener to construe later-arriving spectral information as evidence about the ambiguous interval's identity. If that later information is itself less spectrally ambiguous than the initial information, then the judgment about that initial ambiguous interval will take on the label of the unambiguous following information that was continuous with it. This predicts that contrastive versus assimilative responses in categorization should be determined by the order of contexts and targets. In context-target order, the context provides a standard against which to hear or judge a target that is available from the moment the target is heard. In target-context order, such a standard is not available while processing the target, and so similarity-based grouping can make a gradual spectral change from the target to the context sound like the target is indeed similar to its context. These two tendencies together would yield an effect of the relative order of spectrally ambiguous and unambiguous sounds on whether listeners' responses are contrastive or assimilatory: context-target orders would more often lead to contrastive responses, while target-context ones would more often lead to assimilatory responses. We acknowledge the existence of apparent backward spectral contrast effects on judgments of /s/ versus /ʃ/ before the vowels /i/ or /u/ (see, for example, Mitterer, 2006; Nittrouer & Whalen, 1989; Smits, 2001a, b; Whalen, 1981; Winn, Rhone, Chatterjee, & Idsardi, 2013; inter alia), but, as elaborated in the general discussion, we argue that these sequences do not fulfill the necessary preconditions for spectral assimilation, and so do not constitute counterexamples to the account proposed here.

In Experiments 1a/b and 2a/b, we established that judgment order determines the nature of the effect that is found, as predicted. In Experiment 3, we further demonstrated that an interval of gradual change between target and context is sufficient for assimilation. In Experiment 4a, we confirmed another prediction of this account, namely that inherently spectrally more ambiguous speech sounds are more vulnerable to assimilation than inherently less ambiguous ones, because a listener relies more on the evidence construal given by this default when she is less certain about the evidentiary value of the signal she has heard so far. In Experiment 4b, we demonstrated that a more gradual rate of spectral change between preceding target and following context sounds gave rise to more assimilation, as expected if it is indeed a continuous change between ambiguous and unambiguous components of the acoustic signal that underlies assimilation effects.

Experiment 1a

In Experiment 1a, we established a contrast effect with stimuli where a context preceded a target, that is, by using exactly the configuration that the prevailing accounts of spectral contrast effects predict should give rise to a spectral contrast effect. Listeners categorized a target continuum from spectrally high /t/ to spectrally low /p/ that occurred after an unambiguous context vowel, either spectrally high /i/ or low /u/. These consonant-vowel context-target sequences were combined with an initial /h/ to create the real word continua of *heap–heat* versus *hoop–hoot*.

Methods

Participants Eighteen undergraduate students from the University of Massachusetts Amherst participated in exchange for course credit. All of them were adult native speakers of American English who reported no history of speaking or hearing disorders, and who were not exposed to any language other than English before the age of five. No participant took part in more than one study reported in this paper.

Stimuli The endpoint stimuli were all /h$V_{context}C_{target}$/ words, *heat, heap, hoot, hoop*. The stimuli were modeled on natural productions recorded from the last author and digitized at a sampling rate of 44.1 kHz at 16-bit resolution. Formant frequencies and bandwidths, fundamental frequencies, and intensities were measured from these recordings, and those measurements were then edited and smoothed before being used as input parameters to the Klatt synthesizer (Klatt & Klatt, 1990).

Each initial /h/ was a voiceless version of the following vowel, so the frequencies of F2 and the higher formants did not change in the transition from /h/ to the following vowel steady-state. At the boundary between the /h/ and the vowel, the formants ceased to be excited by noise and began to be excited by voicing.

In order to create vowels for consistent context conditions, /i/s and /u/s were treated separately. The steady states of the /i/ vowel before /t/ and the /i/ vowel before /p/ were averaged. The final value of this average steady state provided the starting value of the vowel-consonant transitions. Formant frequencies were interpolated in order to find their values during the transitions, such that the frequencies began at the last value of the steady state and ended at the last vowel value measured for either /i/ before /t/ or /i/ before /p/. A 20-step context-vowel continuum was then synthesized so that the vowel-consonant transition trajectories would change from those appropriate for /i/ before /t/ to those appropriate for /i/ before /p/ in the same number of steps as the target consonant continuum. In this way, the steady states were consistent for /i/ as a context before /t/ and /p/, but the transitions were appropriate for /i/-/t/ transitions at one end of the continuum, /i/-/p/ transitions at the other end, and their intermediate steps in between. The same procedure was performed for the values of the /u/ steady states and /u/-/t/ versus /u/-/p/ transitions. The acoustics of /i/ and /u/ did not affect each other, because /i/ was only averaged with /i/ and /u/ was only averaged with /u/.

In order to form the bursts of the word-final target consonant continua, noise bursts that occurred at the release of the final stops were taken directly from the same recordings. To neutralize any information about the adjacent vowel in the bursts, the /t/ burst from after /i/ was added to the /t/ burst from after /u/, and similarly, the /p/ burst next to /i/ was added to that next to /u/. The 20-step target continuum was then created by adding energy from the resulting /t/ and /p/ bursts in complementary proportions. This allowed for the creation of a /t/-to-/p/ continuum that was, in terms of its distribution of energy in the stop bursts, equally appropriate after either context vowel. Stimuli were then created by appending the appropriate vowel context continua to 80 ms of silence for the final stop closure, followed by the appropriate stop for the burst continuum for each condition.

Procedure Participants were tested in sound-attenuated rooms. Sound was presented binaurally at a comfortable listening level via circumaural headphones. Participants judged all of the steps of the target continuum after each vowel context, where the stimuli formed the h$V_{context}C_{target}$ words *heat–heap* versus *hoot–hoop*. Practice trials were drawn from steps 1, 3, 18, and 20 of each of the continua in each context, and these were presented three times each in random order. After each practice trial with the endpoint steps 1 and 20, participants saw a letter on the screen for 750 ms indicating which consonant endpoint they just heard. No feedback was displayed after practice trials with steps 3 and 18. Participants proceeded from practice to experimental trials regardless of practice performance. For each context, continuum steps 1–6 and 15–20 were presented only 20 times each, while steps 7–14 were presented 30 times each. The intermediate steps along the continuum were presented more often in order to obtain better estimates of how the contexts influenced the categorization of these more ambiguous stimuli. Each block consisted of four presentations of steps 1–6 and 15–20 and six presentations of steps 7–14 of all continua in all contexts. Stimulus presentation in each block was fully randomized, and participants were given the opportunity to rest for as long as they wanted between blocks. During the experiment, listeners also saw a consonant letter on the screen after an endpoint stimulus, so as to help them maintain their representations of the clearest possible consonants that occurred in the study. These letters appeared regardless of the responses that preceded them, and participants were told that these letters were not feedback, that is, these letters did not depend on their previous responses. Participants received no feedback about the correctness of their responses at any time. For all types of trials, the response prompts "t" = /t/ and "p" = /p/ were displayed on a computer monitor, corresponding to the respective buttons of a button box. Assignment of response prompt to sides was counterbalanced across participants. Both categorical response judgments and response times were recorded. Only responses that occurred within 1500 ms of the stimulus onset were recorded; this resulted in the loss of no more than 3% of all the trials of any study reported in this paper. The experiment advanced either after a button press or after the response window had elapsed. The next trial began 750 ms after the previous trial ended.

Results

Under all accounts of speech perception to date, listeners were expected to respond spectrally low "p" more often after spectrally high /i/ contexts than after spectrally low /u/ contexts. Figure 1 shows that the mean proportion of spectrally low "p" responses is, as expected, greater after spectrally high /i/ than spectrally low /u/. A mixed

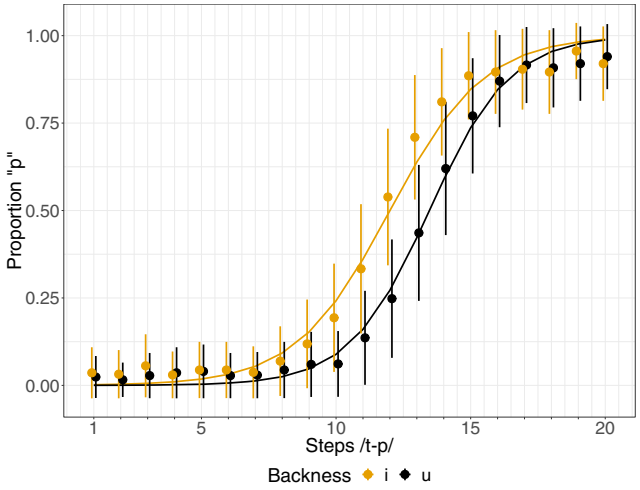

Fig. 1 Experiment 1a. *Dots*—Mean proportions (95% CI) of "p" responses across the /t–p/ continuum following /i/ (*black*) versus /u/ (*ochre*). *Lines*—predicted values from the mixed effects logistic regression model

effects logistic regression model was fitted to the "p" responses. Step values were coded as increasing from /t/ to /p/, and centered and scaled. Context was contrast-coded positively (0.5) for spectrally high /i/ and negatively (−0.5) for spectrally low /u/, because spectrally high /i/ contexts were expected to receive the greatest number of spectrally low "p" responses. Random effects in the model were decorrelated slopes and intercepts by participants for step, context, and their interaction. The statistics in Table 1 show that listeners responded "p" significantly more often relative to "t" as the consonant became more like the /p/ endpoint of the continuum (positive step), and when the preceding vowel was /i/ rather than /u/ (positive context), but that the increase in "p" responses after /i/ compared to /u/ was smaller for some continuum steps than others (negative step by context interaction).

Discussion

The results of Experiment 1a exhibited an expected spectral contrast effect: listeners judged the ambiguous middle steps of the /t/-to-/p/ target continuum to be spectrally different from the context vowel that preceded them. Having established that these stimuli indeed yielded a spectral contrast effect in the classic context-target judgment configuration, we proceeded to test in Experiment 1b whether reversing the judgment order to target-context with these same stimuli would yield a spectral assimilation effect.

Table 1 Fixed effects table for mixed effects logistic regression on Experiment 1a. Model specification is described in the text

Fixed effect	$\hat{\beta}$	*se*	*z*	*p*
(Intercept)	−1.424	0.138	−10.336	<0.0001
Step	0.621	0.040	15.563	<0.0001
Context	1.136	0.204	5.555	<0.0001
Step:Context	−0.101	0.027	−3.726	<0.001

Experiment 1b

In Experiment 1b, the roles of target and context were reversed: listeners categorized a spectrally high to spectrally low /i/-to-/u/ vowel target continuum that preceded unambiguous either spectrally high /t/ or spectrally low /p/ consonant contexts, where the stimuli were /hV$_{target}$C$_{context}$/ continua, *heap–hoop* versus *heat–hoot*. On the basis of the prior studies, listeners were expected to evince spectral assimilation effects and judge targets to be more similar to their following contexts.

Methods

Participants Another 18 participants from the same population as Experiment 1a were tested.

Stimuli The stimuli were based on the same recordings as those used for Experiment 1a, and the same synthesis tools were used. The vowels and consonants were created differently from those of Experiment 1a, because the vowel continua varied from /i/ to /u/, while the consonant contexts remained constantly /t/ or /p/.

In order to form the target vowel continua, the steady state values of the first formant (resonance) for all four of /i/ before /t/, /u/ before /t/, /i/ before /p/, and /u/ before /p/ were averaged together resulting in one grand mean first formant. This was appropriate, because both /i/ and /u/ vowels have similar, and characteristically low, F1 values. For this reason, the second and all higher formants were manipulated so that a mean steady state was found for both /i/ tokens, that is, /i/ before /t/ and /i/ before /p/, as well as one for both /u/ tokens. The values of the second formant (and higher formants') steady states were then varied in equal Hertz intervals so as to change in 20 steps from the mean steady state for /i/ to the mean steady state for /u/. This formed an /i/-to-/u/ steady state continuum that contained acoustic information equally consistent with the presence of a following /t/ or a following /p/. Vowel-consonant transitions appropriate to the following context consonant were found for the /i/ and /u/ endpoints by interpolating from the final value of the steady state to the final value measured for each vowel before each consonant. The same equal interval procedure was then applied to find the transitions for the intermediate continuum steps that would be appropriate for each context consonant.

In order to form the context consonants, the energy of a /t/ burst after /i/ was added to that of a /t/ burst after /u/, and likewise for a /p/ burst after /i/ and a /p/ burst after /u/. Since the contexts needed to remain simply /t/ or /p/ throughout, nothing further was done, because the burst of each output consonant contained acoustic information that was equally appropriate after /i/ or /u/.

The /h/s from /i/ and /u/ contexts were added together in complementary proportions so that their spectra matched those of their following vowels. Again, the component pieces were conjoined to form word continua, now *heap–hoop* versus *heat–hoot*.

Procedure Participants judged stimuli that formed /h$V_{target}C_{context}$/ test words *heat–hoot* versus *heap–hoop*. If spectral assimilation occurs when targets precede their contexts, then listeners in Experiment 2 would be expected to respond spectrally low "u" more often before spectrally low /p/, and spectrally high "i" more often before spectrally high /t/. The procedure for Experiment 1b was the same as that reported for Experiment 1a, except that the response prompts given on screen were now "ee" = /i/ and "oo" = /u/.

Results

Figure 2 shows that the mean proportion of spectrally low "u" responses was greater before spectrally low /p/ rather than spectrally high /t/. A mixed effects logistic regression model was fitted to the "u" responses. Step values were centered and scaled, and increased from /i/ to /u/. Context was contrast-coded positively (0.5) for /p/ and negatively (-0.5) for /t/, under the expectation that an assimilation effect would yield more spectrally low "u" responses before spectrally low /p/. Random effects were again de-correlated slopes and intercepts by participants for step, context, and their interaction. The fixed effect estimates in Table 2 show that listeners responded spectrally low "u" significantly more often relative to spectrally high "i" as the vowel approached the /u/ end of the continuum (positive step), when the following stop was spectrally low /p/ rather than spectrally high /t/ (positive context), and more so for some steps before /p/ compared to /t/ than for others (positive step by context interaction).

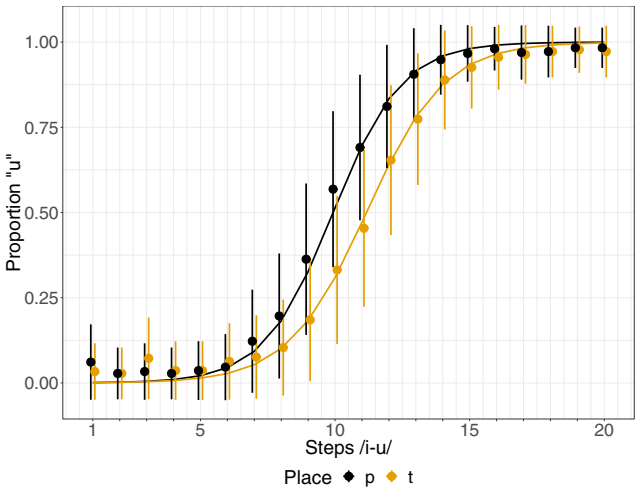

Fig. 2 Experiment 1b. *Dots*—Mean proportions (95% CI) of "u" responses across the /i–u/ continuum before /p/ (*black*) versus /t/ (*ochre*). *Lines*—predicted values from the mixed effects logistic regression model

Discussion

The results of Experiment 1b fit the pattern of spectral assimilation. When categorizing targets that preceded their contexts, listeners judged the ambiguous middle continuum steps to be spectrally similar to the context segment that followed them. Taken together with the results of Experiment 1a, these results suggest that spectral contrast versus assimilation is determined not by the identity of the speech sounds in a string, but by the order of judgment, that is, by which speech sound is ambiguous relative to which is clear. In Experiment 1a, clear preceding contexts were available to provide a standard against which the following target continuum could be judged. Under any account that posits that such contrast arises from continual change detection, this is expected. In Experiment 1b, only clear contexts followed preceding target continuum steps. Thus, ambiguous middle continuum steps were heard before a clearly spectrally high or low standard that they could be compared to. While no auditorist account positively predicts spectral assimilation in this context, none necessarily rules it out, either. Ohala's account of assimilation, under which listeners "misparse" only when following context information is not clearly conveyed, is falsified by the fact that the /t/ and /p/ contexts used in Experiment 1b had the same properties as the /t/ and /p/ endpoint steps of the target continuum in Experiment 1a. These contexts were certainly clearly audible, and listeners nonetheless did not use their presence to implicitly reason that ambiguous target stimuli should be maximally different from them. However, by reversing judgment order, we also changed whether listeners judged consonants or vowels. Experiments 2a and

Table 2 Fixed effects table for mixed effects logistic regression on Experiment 1b. Model specification is described in the text

Fixed effects	$\hat{\beta}$	*se*	*z*	*p*
(Intercept)	−0.007	0.250	−0.028	0.977
Step	4.240	0.327	12.980	<0.0001
Context	0.452	0.123	3.684	<0.001
Step:Context	0.253	0.071	3.587	<0.001

2b tested, and falsified, the possibility that such assimilation effects are only a property of vowel-consonant sequences by showing that spectral contrast versus assimilation is determined by the order of judgment in the words *dee, do/due, be/bee*, and *boo*.

Experiment 2a

In Experiment 2a, we established a contrast effect for a context-target order between clear consonantal contexts, spectrally high /d/ and spectrally low /b/, and the ambiguous middle continuum steps of a following spectrally high-to-low /i/-to-/u/ continuum. This experiment provides the foundation for demonstrating that word-initial consonant-vowel sequences can host both contrast and assimilation effects, as a function of the order in which contexts and targets occurred. If effect type, that is, spectral contrast versus assimilation, is determined by judgment order, then participants in Experiment 2a would be expected to contrast following targets with their preceding contexts, responding more spectrally low "u" after spectrally high /d/ and spectrally high "i" after spectrally low /b/.

Methods

Participants Thirty-three additional participants were recruited from the same population as for the previous experiments.

Stimuli Stimuli were the real words *dee, do/due, be/bee, boo*. Model pronunciations of stimuli were collected from the same speaker and recording conditions as for the previous studies. Also included in this task were the nasal consonant analogues of /d/ and /b/, that is, a spectrally high /n/ and spectrally low /m/. These nasals served as fillers for this study, and stimuli for another one. Target vowels were drawn from 20-step /i/ to /u/ continua created in the same way as the target vowels for Experiment 1b.

For voiced stop consonants, the consonant intervals were divided into closure and burst portions. Total durations and average F0 measures were extracted from the closures of the selected models, because closures contained periodicity as a result of voicing. The averages of these values were used as the input parameters to the Pitch-Synchronous Overlap and Add (PSOLA) synthesis function in Praat (Boersma & Weenink, 2019). PSOLA replaced the naturally occurring duration and pitch values from the original consonant tokens, so that it would be possible to create a single closure that was equally influenced by all context consonants without asynchronies in their oscillations canceling each other out or creating artifacts. This single voiced closure interval was appended to all stop bursts. Context consonants were created in the same way as the contexts from Experiment 1b, except that the continua were now, for example, from /d/ in front of /i/ to /d/ in front of /u/.

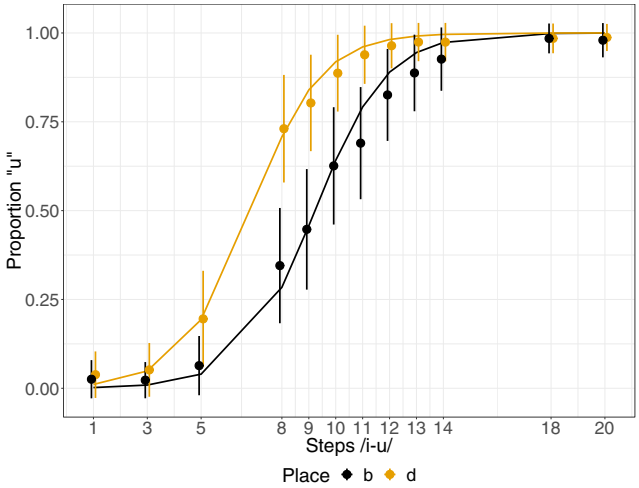

Fig. 3 Experiment 2a. *Dots*—Mean proportions (95% CI) of "u" responses across the /i–u/ continuum following /b/ (*black*) versus /d/ (*ochre*). *Lines*—predicted values from the mixed effects logistic regression model

Procedure The procedure for Experiment 2a was the same as that for Experiments 1a and 1b, except that listeners judged the final vowels in the $C_{context}V_{target}$ words *dee–do/due* versus *be/bee–boo*, and steps 1, 3, 8, 9, 10, 11, 12, 13, 14, 18, and 20 of a 20-step /i/-to-/u/ continuum were presented.

Results

Figure 3 shows that the mean proportion of spectrally low "u" responses was greater after spectrally high /d/ than spectrally low /b/. A mixed effects logistic regression was again fitted to the "u" responses. Step values were centered and scaled, and context was contrast-coded positively (0.5) for initial /d/ and negatively (−0.5) for initial /b/, reflecting the expectation of more "u" responses after /d/ than /b/. Random effects were again de-correlated slopes and intercepts by participants for step, context, and their interaction. The significantly positive estimate for the intercept reported in Table 3 shows that listeners responded "u" more often overall, and the significantly positive estimates for step and

Table 3 Fixed effects table for mixed effects logistic regression on Experiment 2a. Model specification is described in the text

Fixed effects	$\hat{\beta}$	Std. error	z	p
Intercept	1.762	0.217	8.112	<0.001
Step	4.138	0.243	17.052	<0.001
Context	0.925	0.132	7.008	<0.001
Step x Context	0.049	0.086	0.568	0.57

context showed that they also responded "u" more often both as the vowel became more /u/-like and after /d/ than after /b/. The effect of context did not depend on step.

Discussion

In Experiment 2a, listeners responded with more spectrally low "u" after spectrally high /d/, and with more spectrally high "i" after spectrally low /b/. A similar contrast effect was also found for the context-target order presented in Experiment 1a, where targets were consonants instead of vowels. Together, these results show that contrast effects are observed for the context-target orders, independent of whether the target is a vowel or a consonant. The results of Experiment 2a set the stage for a comparison with Experiment 2b, and a more direct test of the prediction that judgment order, not the consonantal or vocalic manner of the particular segments or their place in a word, governs whether listeners respond with spectral contrast or spectral assimilation.

Experiment 2b

In Experiment 2b, we further tested the hypothesis that the type of context effect found in listeners' responses depends on order, and not on the consonantal or vocalic nature of the target sounds. In Experiment 1b, we observed an assimilation effect for a target-context order, where the target was a vowel. In Experiment 2b, we tested a target-context order, where the target was a consonant. That is, in Experiment 2b, the roles of vowel and consonant were reversed relative to Experiment 2a. Listeners categorized word-initial target consonants in spectrally high to low continua from /d/ to /b/ that were followed by either unambiguous spectrally high /i/ or spectrally low /u/ context vowels. The stimuli were again *dee–be/bee* versus *do/due–boo*, which are word frequency-biased against assimilation. Therefore, a finding of assimilation could not be attributed to lexical biases, and instead must be a function of the target-context judgment order common to Experiments 1b and 2b. We predicted an assimilation effect if order, not manner, determines the type of effect. That is, we predicted that participants in Experiment 2b would be expected to respond with spectrally high "d" more often before spectrally high /i/, and spectrally low "b" more often before spectrally low /u/.

Methods

Participants Twenty-seven participants from the same population were tested as in the other experiments.

Stimuli Stimuli were synthesized from the same natural models of these consonants as those for Experiment 2a.

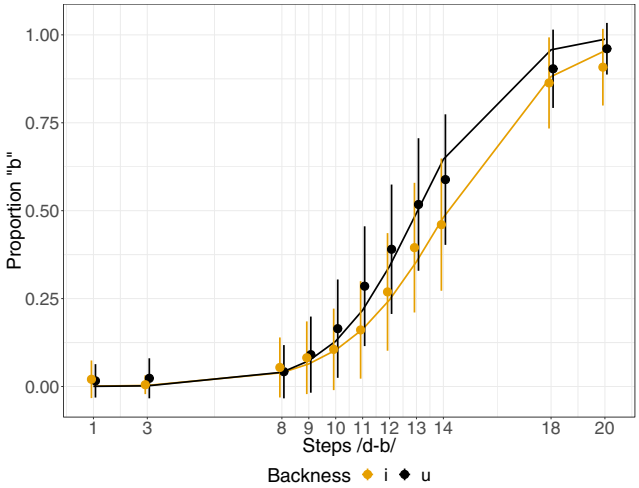

Fig. 4 Experiment 2b. *Dots*—Mean proportions (95% CI) of "b" responses across the /d–b/ continua preceding /i/ (*ochre*) versus /u/ (*black*). *Lines*—predicted values from the mixed effects logistic regression model

As for Experiments 1a, 1b, and 2a, bursts were isolated from the selected tokens of the target words. The /d/- and /b/-bursts from before each vowel context were mixed in complementary proportions to form 20-step continua of bursts from *dee* to *be/bee* and *do/due* to *boo*. Again, separate voiced oral stop continua were made, one for each context vowel, in order to preserve whatever acoustic effect the vowel may have had on preceding consonants. Vowel contexts were made in the same way as those for Experiment 1a.

Procedure Participants judged stimuli that formed C_{target} $V_{context}$ test words *dee–be/bee* versus *do/due–boo*. The procedure for Experiment 2b was the same as that for Experiment 2a, except that the response prompts given on screen were now "d" = /d/ and "b" = /b/.

Results

Figure 4 shows that mean proportions of low "b" responses were greater before low /u/ than high /i/. A mixed effects logistic regression with fixed effects of centered and scaled continuum step, vowel context (low = 0.5 vs. high = −0.5), and their interaction was fitted to listeners' responses. Again, this coding of context reflected expectations about participants' responses: the vowel /u/ is spectrally lower than the vowel /i/, and so the proportion of low "b" responses in a target-context order is expected to be higher in the context of low /u/ than high /i/ vowels if assimilation takes place. Random effects were again de-correlated slopes and intercepts by participant for step, context, and the interaction of step by context. The results are reported in Table 4.

Table 4 Fixed effects table for mixed effects logistic regression on Experiment 2b. Model specification is described in the text

Fixed effects	$\hat{\beta}$	Std. error	z	p
Intercept	−1.587	0.183	−8.650	<0.0001
Step	3.357	0.184	18.204	<0.0001
Context	0.166	0.140	1.181	0.238
Step x Context	0.315	0.074	4.290	<0.0001

The significantly negative estimate for the intercept in Table 4 shows that listeners responded "b" less often than "d" overall. The significantly positive estimate for step shows that they responded "b" more often as continuum step increased, that is, as the target became more /b/-like. The estimate for context is not significant by itself. Instead, the significant positive interaction of context with step shows that, in general, as the target became more /b/-like, the proportion of "b" responses increased more before /u/ than /i/.

Discussion

The results of Experiment 2b evinced spectral assimilation effects: listeners categorized target consonants as spectrally similar to the vowels that followed them. In conjunction with the results of Experiment 2a, which demonstrated that this same set of words could also produce spectral contrast effects in the opposite judgment configuration, these results provide further evidence that the judgment order of contexts and targets determines whether spectral contrast or assimilation is found. The combined results of Experiments 1a, 1b, 2a, and 2b are summarized in Table 5.

In words with both vowel-consonant and consonant-vowel sequences, listeners responded with assimilatory responses to target-context judgment orders, while they responded with contrastive responses to context-target judgment orders. The results presented here confirm that assimilation and contrast effects can be found across the segments of real words, word initially and word finally, and from vowels to consonants as well as consonants to vowels. Crucial to the account advanced here, this is possible, because the two speech sounds in each of these words are spectrally similar and continuous enough with each other to conform to the necessary conditions for assimilation laid out above.

Table 5 Summary table for effect of order on lexical items

Experiment	Continua	Results
1a	*heat–heap, hoot–hoop*	contrast
1b	*heap–hoop, heat–hoot*	assimilation
2a	*dee-do, bee–boo*	contrast
2b	*dee-bee, do-boo*	assimilation

Experiment 3

In Experiment 3, we further examined the sufficient preconditions for spectral assimilation: is it, in fact, the interval of continuity between targets and contexts that is sufficient for assimilation? We have framed the problem of a target followed by its context as one of an interval of ambiguity without a preceding standard against which to compare it. Our account holds that, in such target-context cases, listeners default to the parse of the signal that attributes incoming acoustic evidence to a target sound for longer than might reasonably be expected, if listeners were implicitly attempting to leverage maximal differences between targets and contexts. This occurs even when they can be certain that another speech sound will certainly follow, for example, as in Experiment 2b (where neither /d/ nor /b/ can occur in isolation in English). That is, they do not seem to begin pre-emptively reserving evidence for the following context speech sound, at least not until some point after they have already construed a large portion of the acoustic effects of that context sound's articulation as part of the target.

In the case of word-final stop consonants such as /t/ and /p/, bursts are optional, unlike word-initial stops. When word-final stops are articulated with release bursts as they were for Experiments 1a and 1b, there are two intervals of acoustic information that can provide cues to the identity of the stop consonant: the vowel-consonant formant transitions and the burst. Vowel-consonant transitions move through frequency space as a result of the articulators moving from the extremum of the vowel's articulation toward the ultimate location that they will assume during the consonant's closure. These transitions thus convey information about both the identity of the vowel and the consonant, but, because they are continuous, there is no inflection point before which information *should* be considered vocalic versus consonantal. Vowel-consonant transitions may thus be considered a region of durative between-segment ambiguity: at each moment in a smooth vowel-consonant transition, listeners are faced with the problem of whether to parse the spectra in the signal they have just heard as evidence about the vowel, or to parse those spectra as evidence about the following consonant (or, somehow, as evidence of both). No such ambiguity exists for the release bursts, because stop consonant release bursts in word-final position are unambiguously evidence about the identity of a consonant.

In Experiment 3, we presented listeners with the same stimuli as the *heat-hoot* versus *heap-hoop* words

in Experiment 1b, but we added another manipulation: match versus mismatch of the final consonantal burst with the place suggested by the vowel-consonant transitions. In spectral terms, this meant that for crucial middle-continuum trials, listeners first encountered an interval of spectral ambiguity in the steady state of the target vowel, which was neither spectrally high nor spectrally low. They then encountered an interval of between-segment ambiguity, namely the transition out of the steady state of the vowel and toward the consonant. The acoustic information in this interval moved through frequency space from the ambiguous steady state toward either unambiguous spectrally high energy, if a /t/ sound followed, or unambiguous spectrally low energy, if a /p/ sound followed. When they then encountered a matched stop burst, they reached a part of the signal that confirmed that this vowel-consonant transition could ultimately be construed as evidence about both the vowel and the consonant, because that burst matched the spectral quality of the end of the transitions' trajectory. However, when they encountered a mismatched stop burst, they reached a disambiguating point that told them that the spectral information in the vowel-consonant transition could only be consistent with the vowel, since the consonant was ultimately of the opposite spectral weight value. If continuously-changing transitions are sufficient for assimilation to occur, then listeners should assimilate just as much with mismatched as with matched following bursts. Alternately, it could be the case that assimilation is driven (primarily) by the spectral quality of the final sound that listeners hear because, when faced with spectral ambiguity early in a stimulus, they choose whatever response option is spectrally consistent with the last energy they hear. That is, assimilatory responses could be the result of a pure recency effect, not any kind of evidence grouping. If the burst interval were used in this way, then it would be necessary for the burst to match the transition interval's final spectral weight in order for listeners to decide on the assimilatory responses we have observed so far. In such a case, if transitions and bursts mismatch, then listeners would assimilate less (if at all) to the spectral value at the end of transition intervals in mismatched than matched conditions, because in mismatched conditions they would be using the burst's spectral weight, not the transition's.

Experiment 3 thus tested whether assimilation indeed depends on the formant transitions into or out of a vowel, by comparing cases when the bursts of following stop consonants match versus mismatch the spectral information in the preceding transitions. If assimilation depends on construing vowel-consonant transition information as evidence about a preceding ambiguous target sound, then listeners should assimilate even when the last spectral weight value they hear is inconsistent with the spectral weight of the preceding formant transitions' trajectory.

Methods

Participants Another twelve listeners from the same population were tested as those in the previous experiments.

Stimuli Listeners were presented with the same spectrally high to low /i/-to-/u/ vowel target continuum as that employed in Experiment 1b, just divided into 10 rather than 20 steps. The following context consonant was a spectrally high /t/ or a spectrally low /p/. The stop identity conveyed by the release burst was the same as that conveyed by the vowel-to-consonant formant transitions in the matching condition, and it was the other stop identity in the mismatching condition.

Procedure Matching and mismatching stimuli were presented together to all listeners. In the matching condition, the stimuli were thus the same as in Experiment 1b, *heat–hoot* versus *heap–hoop*, while in the mismatching condition, they were *heat/p–hoot/p* versus *heap/t–hoop/t*, where "t/p" and "p/t" represent the mismatch between the spectral information conveyed by the transitions and bursts. In each context, steps 1 and 10 were presented 8 times each and steps 2–9 16 times each to each listener. The response prompts were identical to those used for Experiment 1b.

Results

Figure 5 shows the mean proportions of "u" responses as a function of the spectral weight of the following vowel-to-consonant formant transitions, whether the stop burst's

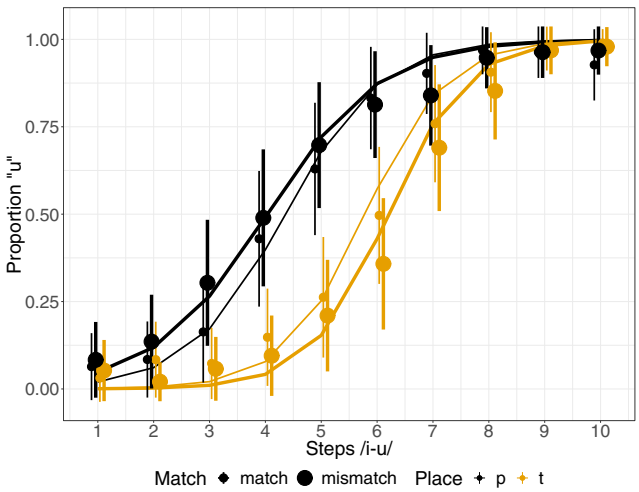

Fig. 5 Experiment 3. *Dots*—Mean proportions (95% CI) of "u" responses across the /i–u/ continua preceding /p/- (*black*) versus /t/-transitions (*ochre*), with match (*thin*) versus mismatch (*thick*) of the spectral weight conveyed by the transitions and the stop burst. *Lines*—predicted values from the mixed effects logistic regression model

weight matched or mismatched the vowel-to-consonant transition's place, and step along the /i–u/ continuum. Listeners responded "u" more often when the vowel-to-consonant formant transitions were those for /p/ rather than /t/, no matter whether the burst conveyed the same or different place of articulation for the stop as the vowel-to-consonant formant transitions. However, the difference between low-conveying and high-conveying transitions was greater when the transitions and bursts conveyed different spectral weights—the mismatching condition—than when they conveyed the same spectral weight—the matching condition.

In the model, centered and scaled step values increased from /i/ to /u/, context was contrast-coded positively for /p/-transitions and negatively for /t/-transitions (0.5 and −0.5), and match was contrast-coded positively (0.5) when the transitions and burst conveyed the same spectral weight and negatively (−0.5) when they conveyed different spectral weights. Random effects were again de-correlated slopes and intercepts by participant for step, context, match, and their interactions. The results reported in Table 6 show that there was no bias toward "u" or "i" responses overall (non-significant intercept) and that listeners responded "u" more often as the vowel became spectrally lower (positive step), before /p/ compared to /t/ (positive context), and when the stop burst had the opposite spectral weight from the vowel-to-consonant transitions (negative match). They responded "u" less often as the vowel became spectrally lower before /p/ than /t/ (negative step by context interaction), but the effect of step did not differ significantly between stimuli with matching versus mismatching bursts. However, when the burst and transitions conveyed different spectral weights, listeners responded "u" much more often before /p/ and much less often before /t/ than when the burst and transitions conveyed the same spectral weight (positive context by match interaction). Finally, listeners responded "u" more often as the vowel became spectrally lower before /p/ than /t/ and when the burst and transition conveyed the same spectral weight, than as the vowel became spectrally lower before /t/ than /p/ when the burse and transition conveyed different weights (negative step by context by match interaction).

Table 6 Fixed effects table for mixed effects logistic regression on Experiment 3. Model specification is described in the text

Fixed effects	$\hat{\beta}$	Std. error	z	p
(Intercept)	0.342	0.457	0.749	0.454
Step	3.007	0.268	11.213	<0.0001
Context	1.042	0.114	9.125	<0.0001
Match	−0.122	0.052	−2.356	0.018
Steps:Context	−0.401	0.116	−3.463	<0.001
Step:Match	−0.080	0.065	−1.229	0.219
Context:Match	0.177	0.062	2.842	0.004
Steps:Context:Match	−0.136	0.064	−2.141	0.032

Discussion

The results of Experiment 3 confirm that the between-segment ambiguous interval of vowel-consonant transitions is sufficient for spectral assimilation effects. When listeners receive evidence that the final stop consonant is incompatible with the spectral weight of the end of the transition interval, they construe the acoustic information at the end of the transitions as evidence about the identity of the target vowel even more than they do when that spectral weight evidence is compatible with both the hypothesized identity of the ambiguous vowel target and the identity of the unambiguous following context consonant.

Two predictions about the extent of assimilation arise from the finding that it is the interval of gradual change, the vowel-consonant formant transitions, that listeners use as evidence for their assimilation responses: the target will assimilate more to the following context when the identity of the target itself is more indeterminate, and when the change between the target and the transition to the following context is more gradual. The first of these predictions is tested in both Experiments 4a and 4b, and the second in Experiment 4b.

Experiment 4a

The results of Experiment 3 established that it was the interval of between-segment ambiguity, in that case the vowel-consonant transitions from the target vowel to the context consonant, that drove the assimilation effects found for target-context stimuli in Experiments 1b, 2b, and 3. We argue that another kind of ambiguity, *within-segment ambiguity*, is also necessary for spectral assimilation effects to arise. We define within-segment ambiguity for our purposes as the degree of ambiguity inherent in a speech sound's spectral weight. If a speech sound is neither spectrally high nor spectrally low, it is inherently within-segment spectrally ambiguous. This argument predicts differences in the likelihoods of undergoing assimilation among the phonemes in natural languages, because some segments are inherently less extreme in their usual acoustic values, and therefore are more ambiguous than others. Experiment 4a explicitly addressed this, by comparing two kinds of American English vowels: *lax* vowels, /ɛ/ as in *bet* and /ʌ/ as in *but*, with *tense* vowels, /e/ as in *bait* and /o/ as in *boat*.

The lax vowel /ɛ/ and tense vowel /e/ are both spectrally high, while the lax vowel /ʌ/ and the tense vowel /o/ are

both spectrally low. Lax vowels /ɛ/ and /ʌ/ are articulated less peripherally, that is, by reaching less extreme articulator positions, than tense vowels /e/ and /o/. The acoustic differences between any given pair of lax vowels are measurably smaller than the acoustic differences between any given pair of tense vowels. For this reason, the steady states of lax vowels are more within-segment ambiguous than the steady states of tense vowels. Furthermore, the extremum of the articulation of lax vowels is held for less time than that of tense vowels, with the result that the steady states of lax vowels are shorter relative to their transitions than those of tense vowels. Lax vowels thus introduce more between-segment ambiguity than tense vowels, as well. Our account therefore clearly predicts that lax vowels should undergo assimilation effects of greater magnitude than tense vowels do, all else equal.

In order to directly compare categorization of lax versus tense vowels, it was necessary to use nonsense words, because the lexicon of English does not provide eight minimally-different words. Targets were from either a spectrally high to low /ɛ/-to-/ʌ/ lax vowel continuum or a spectrally high to low /e/-to-/o/ tense vowel continuum, between a relatively spectrally neutral /kl/ and before either spectrally high /t/ or low /p/ contexts. This resulted in the /klV$_{target}$C$_{context}$/ continuum endpoint nonsense words *klet, klut, klep, klup, klate, klote, klape, klope*.

Methods

Participants An additional 21 participants from the same population as those for Experiments 1–3 were tested.

Stimuli Model recordings were produced and recorded by the same speaker as the other experiments reported here. The same synthesis tools were used.

In order to form the following unambiguous context stop consonants, the same procedure was used as in Experiment 1b, with the exception that separate /t/ and /p/ tokens were made to follow tense vowels versus lax vowels. In this way, the context consonants did not contain information about the spectral weight of the vowel that preceded them, but were appropriate to the vowel kind that they followed.

Separate instances of the initial /k/ burst noise intervals were extracted from each of the eight endpoint stimuli, /klep, klet, klop, klot, klɛp, klɛt, klʌp, klʌt/. The noise intervals from two endpoints from each spectral weight and vowel quality, e.g., from /klɛp/ and /klʌp/ or from /klep/ and /klop/ were then mixed together in complementary proportions, to produce spectrally high to spectrally low pre-lax or pre-tense /k/ continua consisting of the same number of steps as the /l/ and vowel continua described below. These intervals thus preserved any acoustic effects of coarticulation both with the adjacent /lɛ, lʌ, le, lo/ intervals and with the non-adjacent /p, t/. In all three experiments, each step along these /k/ continua was then concatenated with the corresponding step along the /l/ and vowel steady state continua.

In order to produce the stimulus /l/ and target vowels, the same procedures were used as those in Experiments 1b and 3 to create four multi-step continua, /lɛt–lʌt, lɛp–lʌp, let–lot, lep–lop/—"l" represents the voiced portion of the /l/ and "p" and "t" represent the transitions into these stops. Because the formant frequencies and bandwidths for the voiced portion of the /l/ were taken from spectrally high and low vowel contexts for each of these continua, the spectrum of every /l/ covaried directly with the spectral weight and tenseness of the following vowel's steady state. In order to form the vowel-consonant transitions of the target vowels, the same interpolation procedures were used as were applied to formant transitions in Experiments 1b, 2a, and 3 for each of the lax versus tense vowels separately. That is, at no point did lax vowels' values influence those of the output tense vowels, or vice versa. These procedures therefore preserved both the inherently shorter durations of the lax vowels relative to the tense vowels, and the fact that lax vowels' steady states are shorter relative to their transitions than the steady states of the tense vowels'.

Procedure Procedures in Experiment 4a were the same as those for Experiments 1 through 3, except that listeners categorized the eleven odd-numbered steps from 21-step lax and tense vowel target continua. Response prompts were given as "E" = /ɛ/ or "U" = /ʌ/ for the lax vowels or as"A" = /e/ or "O" = /o/ for the tense vowels. Lax or tense vowels were presented in alternating blocks, and whether the first block consisted of lax or tense vowels was counterbalanced between listeners.

Results

Figure 6 shows that listeners gave spectrally low lax "ʌ" or tense "o" responses more often before spectrally low /p/ than spectrally high /t/. It also shows that the effect of low versus high spectral weight in the context affects the lax vowels more than the tense ones across their respective continua. A mixed effects logistic regression model was fitted to the spectrally low responses, "ʌ" for lax continuum and "o" for the tense continuum. Fixed effects in the model were centered and scaled step, context (spectrally low /p/ coded as 0.5, spectrally high /t/ as −0.5), and vowel type (lax coded as 0.5, tense as −0.5). Random effects were again de-correlated slopes and intercepts by participant for all the fixed effects and their interactions. Table 7 shows that listeners responded with the spectrally low categories "ʌ" or "o" more often overall (positive intercept), and also as the vowel targets became spectrally lower (positive

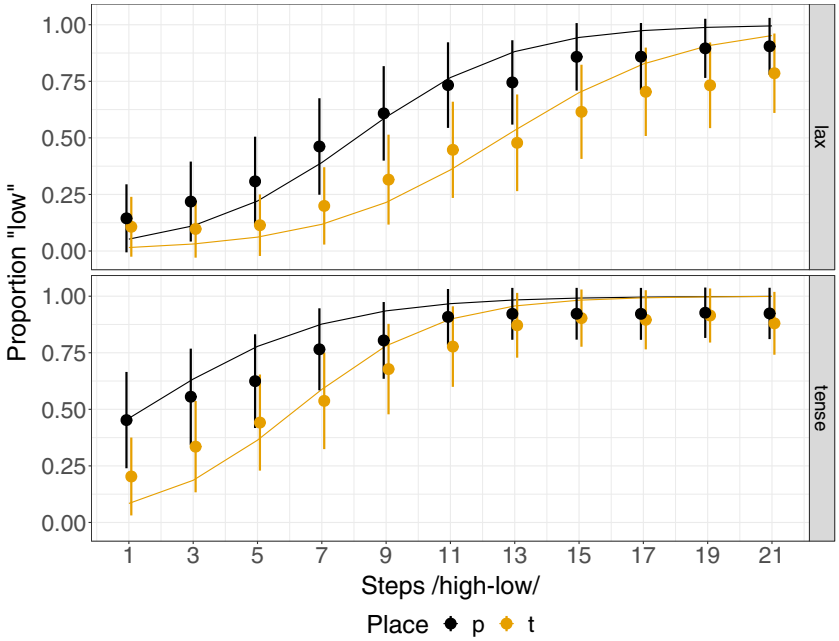

Fig. 6 Experiment 4a. *Dots*—Mean proportions (95%CI) of spectrally low lax "ʌ" (*top*) or tense "o" (*bottom*) responses across the /ɛ–ʌ/ or /e-o/ continua preceding /p/ (*black*) versus /t/ (*ochre*). *Lines*—predicted values from the mixed effects logistic regression model

step) and before the spectrally low /p/ than the spectrally high /t/ (positive context). Spectrally low responses were, however, given less often to the tense than the lax vowels (negative vowel type). None of the two-way interactions were significant, but the three-way interaction among step, context, and vowel type was. Its positive sign indicates that listeners gave the spectrally low lax response "ʌ" more often before spectrally low /p/ than spectrally high /t/ across the steps of that continuum than they gave the spectrally low tense response "o" before /p/ than /t/ across the steps of that continuum.

Discussion

Experiment 4a tested whether greater ambiguity in target sounds would yield more assimilation. We argued that lax vowels are both more within- and between-segment ambiguous than tense vowels. They are more within-segment ambiguous, because even their most extreme endpoint acoustic values are not as different as those of tense vowels. They are also more between-segment ambiguous, because their steady states are shorter relative to their transitions than tense vowels' steady states, so relatively more of lax vowels' formants is taken up by an interval of spectral change. As predicted, listeners in Experiment 4a demonstrated greater spectral assimilation effects in their responses to the more within- and between-segment ambiguous lax vowels than the clearer tense vowels. In conjunction with the results of Experiment 3, we find support for two predictions of our hypothesis. The first prediction is that the degree of within-segment ambiguity determines how vulnerable to assimilation a given speech sound will be, such that more ambiguous sounds will undergo more assimilation. The second prediction is that the interval of gradual change or between-segment ambiguity allows listeners to construe later, clearer information as evidence about the quality of the more ambiguous acoustics that came before it. In our final study, we investigated whether rates of assimilation monotonically increase with the gradualness of change in vowel-consonant transitions.

Table 7 Fixed effects table for mixed effects logistic regression on Experiment 4a. Model specification is described in the text

Fixed effects	$\hat{\beta}$	Std. error	z	p
(Intercept)	1.538	0.353	4.352	<0.0001
Step	2.136	0.266	8.021	<0.0001
Context	0.742	0.145	5.127	<0.0001
Vowel type	−1.238	0.226	−5.482	<0.0001
Step:Context	−0.06956	0.07090	−0.981	0.32649
Step:Vowel type	−0.059	0.161	−0.370	0.712
Context:Vowel type	0.145	0.173	0.840	0.401
Step:Context:Vowel type	0.210	0.074	2.817	0.005

Experiment 4b

In Experiment 4b, we returned to the second of the particular conditions that support spectral assimilation.

Our hypothesis explains spectral assimilation as listeners' defaulting to the parse of the signal that attributes acoustic evidence to the first of two segments when they have received no clear evidence that they should not do so. This hypothesis makes the prediction that the more gradual the spectral change from the ambiguous steady state of a target vowel into its following consonant, the more assimilation listeners' responses should evince, because it is even easier to *not* realize that there is a spectral change underway.

Experiment 4b tests this prediction by adding a manipulation of the ratio between the duration of the target vowel's steady state and its vowel-consonant transition, henceforth called its steady-state:transition ratio, to the conditions of Experiment 4a. Inspection of the model recordings for Experiment 4a revealed that the average of the naturally occurring ratios of the length of the steady-state to its following transition in the vowel interval was 70:30. This average represented a slightly longer steady state than that of lax vowels, and a slightly shorter steady state than that of tense vowels. This ratio was altered by keeping the duration of the vowel interval constant across all the vowels of the study, and compressing the temporal duration of the frequency values of the steady state while complementarily extending the temporal duration of the frequency values of the transition. This resulted in a total vowel duration that was slightly longer than the lax vowels, and slightly shorter than the tense vowels. Crucially, the starting and ending frequency values of the transitions remained the same, to yield experimental steady state:transition ratios of naturally-occurring 70:30, more gradual 50:50, and most gradual 30:70.

Methods

Participants Twenty-four additional participants were tested from the same population as those for Experiments 1– 4a.

Stimuli Stimuli for Experiment 4b were based on those for Experiment 4a. For this reason, only divergences from the Experiment 4a stimuli are noted.

The durations of all vowels were equalized to their mean durations across spectrally high and low, tense and lax vowels, before both /t/ and /p/. The steady-state:transition duration ratios were manipulated so as to produce three distinct ratios, schematized in Fig. 7. The first was 70:30, which equaled the mean of the originally larger tense and smaller lax ratios used in Experiment 4a. Two other ratios were created by lengthening the transition intervals and shortening the steady states proportionally, 50:50, and 30:70. Thus the total duration of the sum of the steady-state and transition intervals remained constant across stimuli. If listeners attributed the ambiguous acoustic information in the vowel-consonant formant transitions to the vowel, then more and more assimilation was expected as the transitions lengthened at the expense of the steady states.

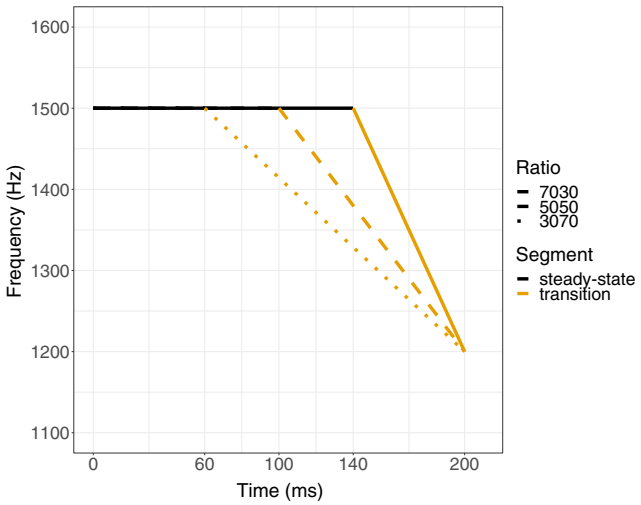

Fig. 7 Schematic representation of F2 trajectories before /p/ as a function steady-state:transition ratio

Unlike in Experiment 4a, /k/ burst intervals were first mixed in equal proportions between the final /p/ and /t/ endpoints for each of the four endpoint vowel qualities /e, o, ɛ ʌ/, thus eliminating any difference that might be due to long-distance coarticulation with the final /p/ or /t/, while preserving any differences due to coarticulation with front versus back and tense versus lax vowels.

Procedure Procedures were identical to those used in Experiment 4a, except for the stimulus steps and endpoint training. Stimuli consisted of steps 1, 4, 8, 12, 16, and 20 from the 20-step tense and lax continua. Fewer, more widely spaced steps were used so that adding three steady-state:transition ratios did not triple the number of trials. Unlike the experiments reported above, participants in Experiment 4b were given endpoint training with both the tense and lax continua before each block of trials to ensure that the categories corresponding to the endpoints were well established perceptually before the listeners were tested on intermediate steps along the continua. Establishing these categories perceptually before testing forestalls any attempt to explain the findings as a failure on the part of listeners to learn the categories.

Results

Figure 8 shows, as expected, that as the steady-state:transition ratio decreases from 70:30 to 50:50 to 30:70, the proportion of spectrally low lax "ʌ" responses increases before spectrally low /p/ relative to spectrally high /t/ far more than "o" responses do.

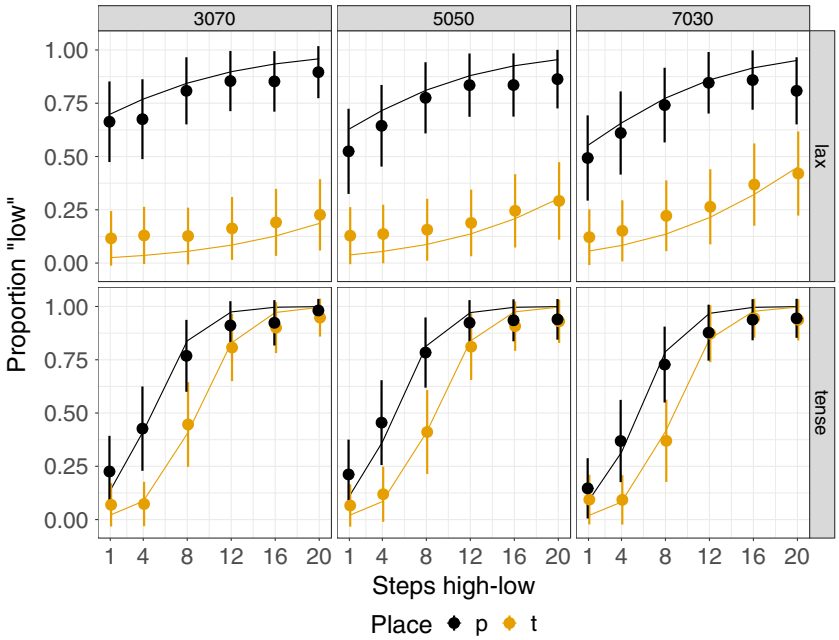

Fig. 8 Experiment 4b. *Dots*—Mean proportions (95% CI) of low "ʌ" (*top*) or "o" (*bottom*) responses across the /ɛ–ʌ/ or /e-o/ continua preceding /p/ (*black*) versus /t/ (*ochre*) at three steady-state:transition ratios. *Lines*—predicted values from the mixed effects logistic regression model

A mixed effects logistic regression model was fitted to the spectrally low "ʌ" and "o" responses. The fixed effects were centered and scaled step (increasing from the spectrally high to the spectrally low endpoints of the continua), context (spectrally low /p/ coded as 0.5, spectrally high /t/ as −0.5), vowel type (lax coded as 0.5, tense as −0.5), ratio (30:70 coded as 0.5, 50:50 as 0, and 70:30 as −0.5). Random effects were, as in the other experiments, de-correlated slopes and intercepts by participants for the fixed effects and their interactions. Table 8 shows that listeners responded with the spectrally low categories "ʌ" and "o" more often overall (positive intercept), that these spectrally low responses became more frequent as the vowel became spectrally lower (positive step), and in the context of a following spectrally low /p/ (positive context). Spectrally low responses were less frequent overall when the vowel was lax rather than tense (negative vowel type). The negative estimate for the interaction of step with vowel type shows that spectrally low lax "ʌ" responses increased less with step than spectrally low tense "o" responses. The positive estimate for the interaction of context by vowel shows that "back" responses increased more before /p/ relative to /t/ for lax than tense vowels. The positive estimate for the interaction of context with ratio shows that more spectrally low responses were given before spectrally low /p/ as the steady-state:transition ratio got smaller, while the positive estimate for the three-way interaction of vowel type by context by ratio shows that this effect increases when the vowels are lax, but decreases when they are tense.

Table 8 Fixed effects table for mixed effects logistic regression on Experiment 4b. Model specification is described in the text

Fixed effects	$\hat{\beta}$	Std. error	z	p
(Intercept)	0.720	0.118	6.108	<0.0001
Step	1.751	0.178	9.860	<0.0001
Context	1.423	0.159	8.958	<0.0001
Vowel	−0.897	0.140	−6.415	<0.0001
Ratio	−0.045	0.052	−0.867	0.386
Step:Context	0.019	0.065	0.298	0.765
Step:Vowel	−1.037	0.110	−9.451	<0.0001
Step:Ratio	−0.053	0.033	−1.609	0.108
Context:Vowel	0.490	0.119	4.122	<0.0001
Context:Ratio	0.186	0.042	4.455	<0.0001
Vowel:Ratio	−0.084	0.046	−1.825	0.068
Context:Vowel:Ratio	0.109	0.045	2.417	<0.05

Discussion

In Experiment 4b, listeners assimilated more as the gradualness of the spectral change during the between-segment ambiguous transition interval increased. This effect was much greater for lax than tense vowels, as expected given how much more similar the acoustics of lax vowel endpoints are relative to those of tense ones. These results confirm that some speech segments, in this case lax vowels, are more susceptible to undergoing assimilation than others. They also confirm that more gradual change between a

spectrally ambiguous interval and an ultimately spectrally unambiguous one causes listeners to use the spectrally unambiguous value more in determining their judgment of the spectrally ambiguous one. The predictions of our hypothesis are thus borne out.

General discussion

We argue that spectral assimilation effects should be understood as cases of listeners parsing later-arriving information as evidence about a preceding target speech sound. Listeners do so only when (i) the change between a target and its following context is smooth and continuous, and (ii) the information conveyed in the acoustic signal is not sufficiently informative for them to do otherwise. That is, listeners assimilate when they judge an ambiguous sound before any possible standard against which to compare it.

Experiments 1a/b and 2a/b established that it is the order of an ambiguous target and an unambiguous context that determines whether listeners will respond contrastively or assimilatory to spectrally ambiguous continuum steps. Experiment 3 isolated the interval of between-segment ambiguity, that is, the gradual transition between a target and its following context, as the component of the acoustic signal that is sufficient for spectral assimilation to occur. The results of Experiment 3 converged with those of Wade and Holt's (2005) nonspeech analogue study to confirm that an interval of continuous change between spectrally ambiguous and unambiguous information is both a necessary and sufficient precondition for spectral assimilation. Experiment 4a tested and substantiated the prediction that those speech sounds that are inherently more spectrally ambiguous are more likely to host assimilation than those that are more spectrally extreme. Experiment 4b provided evidence that more gradual change between a spectrally ambiguous steady state and a spectrally unambiguous transition end value causes listeners to more readily identify a target vowel as the continuum endpoint with the unambiguous transition end's spectral weight.

Together, these experiments provide support for the hypothesis that listeners perceptually group together sufficiently continuous spectral information when they do not have evidence that would lead them to do otherwise. Contra Ohala's (1981, 1993) account of assimilation effects, listeners do so even when the acoustic information about the context of judgment is sufficiently clear that they could hypothetically, under his view, have implicitly reasoned about that context's likely effect on its preceding target. Listeners consistently responded with assimilatory judgments when they heard a spectrally ambiguous target that continuously changed into a spectrally unambiguous context. They did this when judging vowels before context consonants, consonants before context vowels, and, as Wade and Holt (2005) showed, when judging consonants before nonspeech analogue tones. We argue that this should be understood as the listener's default behavior in conditions of ambiguity about the appropriate construal of spectral evidence. Our argument echoes two frequently appealed-to arguments throughout cognitive science. The first, which provided the foundation of much of the field of sentence processing, is that the action of a parser in conditions of ambiguity reveals that parser's default behavior. The second, which informs reasoning wherever in the field Bayesian methods are employed, is that the action of a system in conditions of little or no evidence reveals the contents of that system's prior. We argue that the conditions in which listeners responded in the studies presented here fit these characterizations of ambiguity, and that our findings are consistent with the existence of a *regressive* spectral assimilation bias, so named because incoming acoustics are construed as evidence about a speech sound that has already begun occurring earlier in time, as though looking backward.

By characterizing this regressive spectral assimilation bias as listeners' default behavior, we do not intend to predict that listeners' final conclusions about the speech string should always, or even most frequently, be consistent with regressive spectral assimilation. We posit that, as the acoustic signal unfolds, listeners implicitly consider all possible segmental parses of the information that they receive, but implicitly assign greater likelihood to some parses than others given the evidence they have heard. At the beginning of a signal, identifying a target vowel on the basis of a regressive spectral information grouping is only one of many possible construals of the evidence that a listener implicitly entertains. As time goes on, the listener *may* encounter highly informative acoustic evidence that uniquely identifies the upcoming sound after the ongoing vowel. Once such uniquely identifying information were available, the listener could implicitly select a parse that specified the segments of the current speech stream in more detail, and so did not need to rely on default evidence construal to identify what they were hearing.

Which acoustic information uniquely identifies a segment is relative to the language in which speech sounds are occurring. Cutler and colleagues' work provides suggestive evidence that listeners' parsing strategies differ as a function of the sound inventories of their native languages (see Cutler, 2012 for an overview). For example, Wagner, Ernestus, and Cutler (2006) demonstrated that native listeners of a language with both /f/ and /θ/ (as in theta), such as European Spanish, require both a fricative's noise and its following formant transitions in order to quickly and successfully identify which of these two fricatives they have heard. However, native listeners of a language such as Italian, the inventory of which contains /f/, but not /θ/, do not

require following formant transitions in order to attain the same speed and accuracy at identifying /f/. Wagner et al. attribute this pattern to the fact that /f/ and /θ/ are very spectrally similar, more so than any other two fricatives in the languages that they studied, and so the noise portions of these sounds are not enough to differentiate them with a high degree of confidence. This is only a problem for listeners who must regularly perform this differentiation, and so listeners who do not face this problem (because their native languages do not contain /θ/ in the phoneme inventories) can be more efficient in their segmental parsing by not waiting for upcoming transition information.

To understand how uniquely identifying acoustic information could be used predictively, another example is necessary. In a language with only one nasal speech sound, a consonant, the presence of nasalization during a target vowel would uniquely identify the upcoming segment as the language's single nasal. In such a case, listeners are faced with far less between-segment ambiguity than in the studies presented here, because the uniqueness of a single nasal would allow the listener to be certain of that nasal's spectral properties well before they had actually begun to occur in the incoming acoustics. The degree of certainty that listeners could have on the basis of such nasalization during the vowel-consonant transitions could be the same as the degree of certainty that only became available at the end of the word to the listeners in our Experiment 3. Assuming naturalistic speaking and listening conditions, such uniquely identifying coarticulation would resolve as the predicted sound, and so listeners could immediately and with a high degree of confidence implicitly select a parse of the acoustics that contained the nasal and its effects on a preceding vowel sound. They would *not* need to rely on or select the parse that is the product of grouping incoming spectral information little by little with a continuously linked preceding sound, if that were incompatible with the parse for which they have a high degree of confidence. That is, uniquely identifying coarticulation or covariation between two speech sounds should lead to listeners being certain or nearly certain about the identity of an upcoming sound during an interval that is otherwise spectrally ambiguous. This could allow listeners to be more informed about how to construe the spectral evidence that they encounter, and so lead them to not default to selecting the product of the kind of incremental and gradual perceptual grouping that we have argued underlies spectral assimilation. We therefore argue that spectral assimilation is the output of a default perceptual response, which is always available to listeners, but is only used as the listener's ultimate parse on those occasions when no other, better information is available. An example of this kind can be found in the historical development of the Chinese languages (Chen & Wang, 1975), where earlier three-way place contrasts between syllable-final nasals /m, n, ŋ/ and the corresponding oral stops /p, t, k/ have merged into a single nasal or oral stop, both velar /ŋ/ or /k/ in some innovative languages, and in even more innovative ones, have been replaced by vowel nasalization or a placeless glottal stop. In these innovative languages, a listener can reliably predict what the syllable-final consonant's place is, if there is a syllable-final consonant with a place of articulation any longer, and has only to distinguish syllables that end in nasalization from those that do not.

Under this view, it is unsurprising that the sibilant fricatives, spectrally high /s/ and relatively spectrally lower /ʃ/, do not assimilate to following high /i/ or low /u/ (Whalen, 1981; Smits, 2001a, b; Nittrouer & Whalen, 1989; Mitterer, 2006; Winn et al., 2013, inter alia). While this case may at first appear to be a counterexample to our proposal, we argue that there are several reasons to conclude that it is not. The first reason is that the change from the aperiodic frication of a sibilant to the sonorous periodicity that conveys a vowel's formant structure may already be a sharp enough acoustic discontinuity that listeners' auditory systems would not group vowel transitions with preceding sibilants by default. However, even if this discontinuity alone were not enough, the spectral values of sibilants are so much higher than the rest of the segments of speech that they might well be expected to auditorily stream above the rest of the speech signal, as observed by McMurray (this volume). This kind of difference in spectral weight would suggest that sibilants are not only discontinuous with their following contexts, they are also spectrally dissimilar from them in a way that would forestall auditory grouping into the same interval. Finally, Nittrouer and Whalen (1989) demonstrated that children categorize /s/ versus /ʃ/ differently from adults. They employ different cues, and seem to converge on adultlike behavior only in late childhood. This supports the view that the apparent backward contrast effects observed with /s/ and /ʃ/ before /i/ and /u/ are the product of adult listeners having learned the patterns of covariation between these sibilants and vowels. According to our account, it would then also be unsurprising that these are not standard spectral contrast effects: listeners are not employing the parse of these stimuli that would arise from a default to exaggerate the change between a target sound and its context, they are instead implicitly using more specialized knowledge.

Unlike auditorist approaches, all gesturalist accounts would require significant modification to accommodate the fact that listeners systematically, and in the presence of clear evidence about a target's context, failed to attribute spectral information to the gestures of the speech sound that produced it. Our account of spectral assimilation effects is compatible with all of the prevailing auditorist accounts of spectral contrast. Without a precursor, no adaptation or change detection could have taken place in order to

bias target responses, and so the fact that target-context order gives rise to assimilation is not problematic for these previous explanations. In addition to its compatibility, our account advances the auditorist approach to speech perception, by positing the action of a default that is not speech-specific in explaining listeners' behavior with speech stimuli. Grouping effects based on proximity, whether temporal or spatial, are observed in modalities besides audition. For example, the Bezold illusion (Helson, 1963), pictured in Fig. 9, is a case of visual assimilation. The grey bars, while the same shade of grey on both sides of the figure, appear to be lighter when they occur next to white lines than when they occur next to black lines. The only difference between the preconditions for spectral assimilation that we argued for here and those for the Bezold illusion that we have just described is the dimension of proximity: temporal in speech, but spatial in vision. Just as in the auditory analogue, visual assimilation dissipates if the white and black lines are farther apart. If regressive spectral assimilation is the product of grouping sufficiently similar and temporally close spectra together, then its basis need not be a property of the auditory system, specifically.

In order to construct a comprehensive theory of speech perception, much more investigation of spectral assimilation effects will be necessary. Such future work must determine the physiological basis or bases of spectrally assimilatory responses, while building an understanding of how large spectral assimilation effects are in different conditions, over what span(s) of segments assimilation effects can occur, and why. Ideally, such a model would be able to predict effect sizes. However, effect size estimates from the present experiments are unlikely to be useful in this greater endeavor, for two reasons. All of the real word stimuli in Experiments 1– 3 were frequency biased against assimilatory responses, and so the sizes of the assimilatory effects in Experiments 1b and 2b may have been depressed relative to what they would have been without a lexical bias. However, even in Experiments 4a and 4b, as an anonymous reviewer points out, the ambiguity inherent to lax vowels effectively means that they contain acoustic values that would occur in only the middle steps of the tense vowel continua. Since it is usually the case that middle continuum steps are the only ones that evince effects, listeners' responses to the lax vowels relative to the tense vowels is not remarkable. More work needs to be done in order to understand the generality of these effects, for example testing whether spectrally assimilatory tendencies influence speech parsing behavior in other tasks, such as discrimination. However, we argue that evidence from the phonological typology of the world's languages (that is, how frequently different languages evince various patterns in their sound systems) suggests that regressive spectral assimilation is a pervasive tendency. Javkin (1977) reported that among the place of articulation assimilation processes in the languages of the word, regressive assimilation is overwhelmingly more common. Place of articulation assimilation occurs when two adjacent speech sounds are required to be articulated with the same articulator(s) in the same location in the oral cavity as each other. Spectral weight is the primary auditory cue to place of assimilation, and so a regressive spectral assimilation bias in speech perception would be expected to give rise to a regressive place assimilation prevalence in phonological typology. This has since been confirmed by Jun (1995, 2004) for consonant-consonant sequences, and Bybee and Easterday (in preparation) for consonant-vowel and vowel-consonant sequences across large samples of the languages of the world. We thus argue that the regressive spectral assimilation bias proposed here has already affected the languages heard by perceivers throughout the world today.

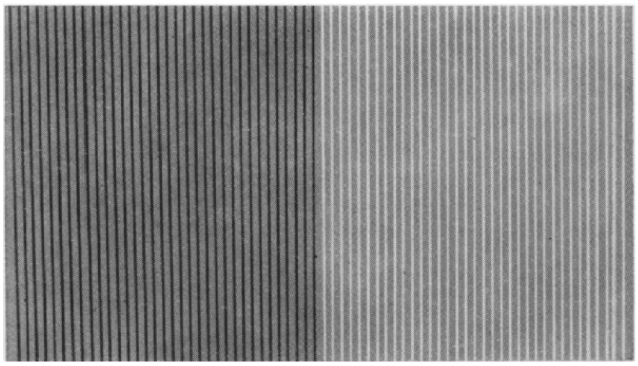

Fig. 9 The Bezold illusion, a visual assimilative effect

References

Aravamudhan, R., Lotto, A. J., & Hawks, J. W. (2008). Perceptual context effects of speech and nonspeech sounds: The role of auditory categories. *Journal of the Acoustical Society of America*, *124*(3), 1695–1703.

Boersma, P., & Weenink, D. (2019). Praat: Doing phonetics by computer [Computer program]. Version 6.0.52, retrieved 2 May 2019 from http://www.praat.org/

Bybee, J., & Easterday, S. (in preparation). The prominence of palatal articulation: A crosslinguistic study of assimilation and strengthening.

Chen, M. Y., & Wang, W. S. (1975). Sound change: Actuation and implementation. *Language*, *51*, 255–281.

Cutler, A. (2012). *Native listening*. Cambridge: MIT Press.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*, 121–144.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *Journal of the Acoustical Society of America*, *85*(5), 2154–2164.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct realist perspective. In I. G. Mattingly, & N. O'Brien (Eds.) *Status report on Speech Research*, (pp. 139–169). New Haven: Haskins Laboratories.

Fowler, C. A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Attention, Perception, and Psychophysics*, *68*(2), 161–177.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(3), 877–888.

Fujimura, O., Macchi, M. J., & Streeter, L. A. (1978). Perception of stop consonants with conflicting transitional cues: A cross-linguistic study. *Language and Speech*, *21*(4), 337–346.

Helson, H. (1963). Studies of anomalous contrast and assimilation. *Journal of the Optical Society of America*, *53*(1), 179–184.

Holt, L. L. (1999). Auditory constraints on speech perception: An examination of spectral contrast (Unpublished doctoral dissertation). University of Wisconsin Madison.

Holt, L. L. (2005). Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312.

Holt, L. L. (2006a). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *The Journal of the Acoustical Society of America*, *120*(5), 2801–2817.

Holt, L. L. (2006b). Speech categorization in context: Joint effects of nonspeech and speech precursors. *Journal of the Acoustical Society of America*, *119*(6), 4016–4026.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *Journal of the Acoustical Society of America*, *108*(2), 710–722.

Hura, S. L., Lindblom, B., & Diehl, R. L. (1992). On the role of perception in shaping phonological assimilation rules. *Language and Speech*, *35*(1–2), 59–72.

Javkin, H. (1977). Phonetic Universals and Phonological Change (Unpublished doctoral dissertation). University of California Berkeley.

Jun, J. (1995). Perceptual and articulatory factors in place assimilation: An optimality theoretic approach (Unpublished doctoral dissertation). University of California Los Angeles.

Jun, J. (2004). Place assimilation. In B. Hayes, R. Kirchner, & D. Steriade (Eds.) *Phonetically based phonology* (pp.58-86). Cambridge: Cambridge University Press.

Kiefte, M., & Kluender, K. R. (2008). Absorption of reliable spectral characteristics in auditory perception. *Journal of the Acoustical Society of America*, *123*(1), 366–376.

Kingston, J., & Shinya, T. (2003). Markedness asymmetries in place perception in consonants. In *Proceedings of the 15th International Congress of Phonetic Sciences*, (pp. 399–402). Causal Productions: Barcelona.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, *87*(2), 820–857.

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

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

Liberman, A. M., & Mattingly, I. G. (1985). Perception of the speech code. *Science*, *243*, 489–494.

Lotto, A. J., & Holt, L. L. (2006). Putting phonetic context effects into context: A commentary on Fowler (2006). *Perception and Psychophysics*, *68*, 178–183.

Lotto, A. J., & Holt, L. L. (2015). Speech perception: The view from the auditory system. In G. Hickok, & S. Small (Eds.) *The Neurobiology of Language*, (pp. 185–194). New York: Academic.

Lotto, A. J., & Kluender, K. R. (1998). General effects in speech perception: Effect of preceding liquid on stop consonant identification. *Attention, Perception, and Psychophysics*, *60*(4), 602–619.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Cambridge: MIT Press.

Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics*, *68*, 1227–1240.

Nittrouer, S., & Whalen, D. H. (1989). The perceptual effects of child-adult differences in fricative-vowel coarticulation. *Journal of the Acoustical Society of America*, *86*, 1266–1276.

Ohala, J. J. (1981). Articulatory constraints on the cognitive representation of speech. *Advances in Psychology*, *7*, 111–122.

Ohala, J. J. (1993). Sound change as nature's speech perception experiment. *Speech Communication*, *13*(1-2), 155–161.

Repp, B. H. (1983). Bidirectional context effects in the perception of VC-CV sequences. *Perception & Psychophysics*, *33*(2), 147–155.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Constraints on the processes responsible for extrinsic normalization of vowels. *Attention, Perception, and Psychophysics*, *73*, 1195–1215.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2012). Hemispheric differences in the effects of context on vowel perception. *Brain and Language*, *120*, 401–405.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2013). Evidence for precategorical extrinsic vowel normalization. *Attention, Perception, and Psychophysics*, *75*, 576–587.

Smits, R. (2001). Evidence for hierarchical organization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(5), 1145–1162.

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

Stilp, C. E., & Anderson, P. W. (2014). Modest, reliable spectral peaks in preceding sounds influence vowel perception. *Journal of the Acoustical Society of America*, *136*(5), EL383–EL389.

Stilp, C. E., Anderson, P. W., & Winn, M. B. (2015). Predicting contrast effects following reliable spectral properties in speech perception. *Journal of the Acoustical Society of America*, *137*(6), 3466–3476.

Stilp, C. E., & Assgari, A. (2018). Perceptual sensitivity to spectral properties of earlier sounds during speech categorization. *Attention, Perception, and Psychophysics*, *1*, 1–11.

Viswanathan, N., Fowler, C. A., & Magnuson, J. S. (2009). A critical examination of the spectral contrast account of compensation for coarticulation. *Psychonomic Bulletin & Review*, *16*(1), 74–79.

Wade, T., & Holt, L. L. (2005). Effects of later-occurring nonlinguistic sounds on speech categorization. *Journal of the Acoustical Society of America*, *118*(3), 1701–1710.

Wagner, A., Ernestus, M., & Cutler, A. (2006). Formant transitions in fricative identification: The role of native fricative inventory. *Journal of the Acoustical Society of America*, *120*(4), 2267–2277.

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *90*(6), 2942–2955.

Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *96*(3), 1263–1282.

Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America*, *99*(6), 3749–3757.

Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the English [s]-[š] boundary. *Journal of the Acoustical Society of America*, *69*(1), 275–282.

Winn, M., Rhone, A., Chatterjee, M., & Idsardi, W. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Frontiers in Psychology*, *4*, 824.