Research Article

Psychological Science
1–14

DOI: 10.1177/0956797620968787
www.psychologicalscience.org/PS
SAGE

Now You Hear Me, Later You Don't: The Immediacy of Linguistic Computation and the Representation of Speech

TC

Spencer Caplan[1], Alon Hafri[2,3], and John C. Trueswell[4]
[1]Department of Linguistics, University of Pennsylvania; [2]Department of Cognitive Science, Johns Hopkins University; [3]Department of Psychological and Brain Sciences, Johns Hopkins University; and [4]Department of Psychology, University of Pennsylvania

Abstract
What happens to an acoustic signal after it enters the mind of a listener? Previous work has demonstrated that listeners maintain intermediate representations over time. However, the internal structure of such representations—be they the acoustic-phonetic signal or more general information about the probability of possible categories—remains underspecified. We present two experiments using a novel speaker-adaptation paradigm aimed at uncovering the format of speech representations. We exposed adult listeners (N = 297) to a speaker whose utterances contained acoustically ambiguous information concerning phones (and thus words), and we manipulated the temporal availability of disambiguating cues via visually presented text (presented before or after each utterance). Results from a traditional phoneme-categorization task showed that listeners adapted to a modified acoustic distribution when disambiguating text was provided before but not after the audio. These results support the position that speech representations consist of activation over categories and are inconsistent with direct maintenance of the acoustic-phonetic signal.

Keywords
language, speech processing, immediacy of computation, mental representation, acoustic maintenance, open data, open materials, preregistered

Received 5/8/20; Revision accepted 8/27/20

Variability is a constant in the world. How cognitive systems represent and process input signals to adapt to such a gradient and shifting landscape is a classic problem in psychology ranging from learning and decision-making (e.g., Erev & Barron, 2005; Gallistel, 1990) to plasticity in visual processing (e.g., Postle, 2015; Sagi, 2011). In this regard, language represents an ideal domain to study the *structure* of mental representations built up in real time and the type of information thus available for learning. We investigate the problem of adaptation and representation through the lens of speech processing by asking the following questions. How do listeners convert a gradient and variable acoustic signal into cognitive units such as phones and words in order to reconstruct the underlying meaning? What happens to the acoustic-phonetic signal after it enters the mind of a listener?

Language unfolds over time. Unlike in reading or visual search, an acoustic signal is inherently ephemeral: If cognitive computations are not made over transient and shifting information as it occurs, they cannot be made at all. This inherent constraint, which we term the *immediacy of linguistic computation*, means that listeners cannot and do not wait until the end of an utterance to begin building a representation of speech (Christiansen & Chater, 2016; following Marslen-Wilson & Tyler, 1980). Thus, a design feature of all models of speech perception (e.g., Marslen-Wilson & Welsh, 1978; McClelland & Elman, 1986) is the real-time construction of *intermediate representations*—that is, representations held in memory (irrespective of content) that outlast the stimulus itself but may be integrated with additional information over time. This intermediate structure

Corresponding Author:
Spencer Caplan, University of Pennsylvania, Department of Linguistics
E-mail: spcaplan@sas.upenn.edu

serves as a listener's working hypothesis for recognition, but given the immediacy of computation, the form of these representations is a bottleneck: Computation can occur only over the material that is constructed rather than the original, ephemeral signal.

Real-time processing involves extracting and integrating linguistic evidence from varied sources, with disambiguating information arriving in the form of multiple temporally disjoint cues, such as visual-articulatory cues (McGurk & MacDonald, 1976), prior lexical knowledge (Ganong, 1980), et cetera. Speech processing is thus a problem of handling and representing uncertainty. Experimental evidence shows that listeners maintain and update intermediate representations over time, both locally (Galle, Klein-Packard, Schreiber, & McMurray, 2019) and over long distances (Bushong & Jaeger, 2017; Connine, Blasko, & Hall, 1991; Zellou & Dahan, 2019). However, although claims in the literature (e.g., Bicknell, Jaeger, & Tanenhaus, 2016; Darwin & Baddeley, 1974; Galle et al., 2019) are varied, such work on long-distance cue integration has not directly addressed the *structure* of information included in these intermediate representations and how this is recruited for adapting to variability. We contrast two classes of theories.

Under a *signal-retention* account, listeners maintain acoustic-phonetic detail (e.g., Bicknell et al., 2016; Goldinger, 1998; McMurray, Tanenhaus, & Aslin, 2009). This would include information such as acoustic cues, among other properties. For example, Bicknell et al. (2016) noted that "recent data from speech perception and sentence processing . . . demonstrate that comprehenders can maintain fine-grained lower-level perception information for substantial durations" (p. 23). A second family of accounts—which we develop here—we term the *activation-over-categories* (AOC) theory. Under the AOC theory, listeners maintain a graded activation pattern over some set of cognitive or linguistic categories (phones, words, etc.). Crucially, this is a Markovian process: Listeners encode a state of activation but do not retain the precise sensory evidence that led to that belief. These states of activation can be understood as predictions that are updated by later linguistic input and thus support learning variation. Phonetic information is recruited for identifying higher-level categories but is not stored or isolable within the speech-processing system. Past work interpreted as evidence for maintenance of acoustic detail (Bushong & Jaeger, 2017; Connine et al., 1991; Crowder & Morton, 1969; Frankish, 2008; McMurray et al., 2009) is also compatible with the AOC account because the AOC account maintains gradience through probabilistic information about linguistic categories, not the acoustic details that gave rise to those probabilities. Such debate between these general accounts of mental representations is pervasive across psychology—for example, in the exemplar versus abstract representations of concepts and categories (e.g., Schuler, Kodner, & Caplan, 2020; Smith & Medin, 1981).

Statement of Relevance

A fundamental challenge for any cognitive system is to represent and process input signals into useful representations. Daily life involves a stream of rapid yet implicit categorization decisions: "What object is this?" "What word did I just hear?"—examples abound. Yet signals are embedded in a variable and noisy world, a problem especially salient in spoken-language processing. Unlike words in reading or items in visual search, acoustic signals are inherently ephemeral: If cognitive computations are not made over transient and shifting information as it occurs, they cannot be made at all. What happens to an acoustic signal after it enters the mind? We demonstrated through two experiments that listening involves no direct retention of the acoustic-phonetic signal over time. Rather, listeners process speech and adapt to variability by storing and updating probabilistic activation over cognitive and linguistic categories. At a broad level, limits to the storage of sensory input place limits on mental representations.

To investigate the contents of intermediate speech representations and evaluate the predictions of a signal-retention account against the AOC theory, we looked at how people adapt to shifts in speech when disambiguating information appears after the original signal rather than before it. We present findings from two experiments using a novel variant of the *accent-adaptation paradigm* (Norris, McQueen, & Cutler, 2003; Samuel & Kraljic, 2009): In this paradigm, after encountering a series of target words with a manipulated distribution over an acoustic cue to some phone, participants subsequently exhibit shifted criteria for categorizing phones, such as /t/ versus /d/ (Bertelson, Vroomen, & De Gelder, 2003; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Jesse & McQueen, 2011; Kraljic & Samuel, 2006; Reinisch & Holt, 2014). In the current study, we exposed participants to acoustically ambiguous audio involving minimal pairs (e.g., *time/dime*). Disambiguation was provided by a text subtitle that appeared either briefly before or after the audio and systematically biased the ambiguous audio to be interpreted either as /t/ or /d/ (Fig. 1; see the Method section for complete details).

When the disambiguating text is provided before the audio, both the signal-retention and AOC accounts

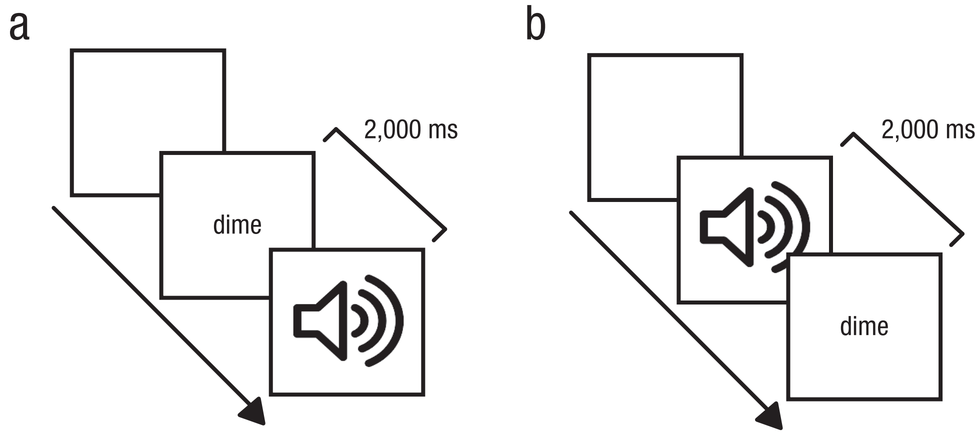

Fig. 1. Timeline of the main experimental manipulation. Participants were provided with disambiguating text either before (a) or after (b) hearing the corresponding audio.

predict that participants should adapt to the shifted phonetic distribution. When reading the word first, participants know the intended phones ahead of time and can evaluate the upcoming ambiguous audio accordingly: The signal can be evaluated given the prior hypothesis. When the text is provided after the audio, only the signal-retention account predicts that adaptation will occur (maintenance of the phonetic detail is the central tenet of the theory). The AOC account, conversely, predicts that no adaptation will occur, because while the graded activation over /t/ and /d/ allows for the proper lexical interpretation once text arrives, the reason for that particular activation state is lost. Thus, there is no pattern to generalize.

Experiment 1

Method

Design. The experiment had a 2 (shift direction: shifted /d/ vs. shifted /t/) × 2 (timing: text before vs. text after) between-subjects design. During the exposure phase, participants heard and saw a sequence of 142 words presented once in a random order. Exposure words were divided between 44 target items (22 “t” onset and 22 “d” onset) and 98 fillers. Each target word was paired with corresponding audio that had a *voice-onset time* (VOT) that was ambiguous (60 ms) or unambiguous (10 ms for “d” words and 100 ms for “t” words). VOT is the time delay between the release of a stop consonant and the onset of glottal pulses from the closed vocal folds. VOT is the primary acoustic cue for distinguishing voiced stops (e.g., /b/, /d/, and /g/) from their voiceless counterparts (/p/, /t/, and /k/). The ambiguous versus unambiguous mapping was controlled by the shift-direction condition—“t” words paired with ambiguous VOT for the shifted-/t/ group and “d” words paired with ambiguous VOT for the shifted-/d/ group. Because we used a fully crossed design, each shift direction occurred with a timing manipulation in which the subtitle appeared either 2 s before the audio (text-before condition) or 2 s after the audio (text-after condition).

In previous studies (Jesse & McQueen, 2011; Kraljic & Samuel, 2006), the interpretation of manipulated audio under an accent-adaptation paradigm was provided by local lexical context (e.g., only one interpretation of “croco[t/d]ile” results in a real word). However, adaptation induced by lexical context is not informative about the structure of intermediate representations, because listeners can resolve the [t/d] ambiguity locally, regardless of their ability to store phonetic detail. We explicitly removed information needed to disambiguate words internally by using *minimal pairs*: words that differ in exactly one phoneme. This is similar to distributional approaches to adaptation (Clayards et al., 2008; Munson, 2011), except that our method does not require hearing a large number of repeated tokens, and it allows for the direct manipulation of disambiguation timing.

The test phase was identical for all participants. On each test trial, participants heard a syllable beginning with an alveolar stop consonant with a particular VOT (ranging from 20 ms to 80 ms, in randomized order), and they were asked to judge whether they heard a /t/ or a /d/. The design is schematized in Figure 2. The design, analysis, and exclusion criteria for this experiment were preregistered (https://osf.io/5nvk2/).

Participants. We recruited 132 University of Pennsylvania undergraduates who received course credit for their participation. All participants were native English speakers with no reported hearing or visual impairments. As was planned in the preregistration, the final sample

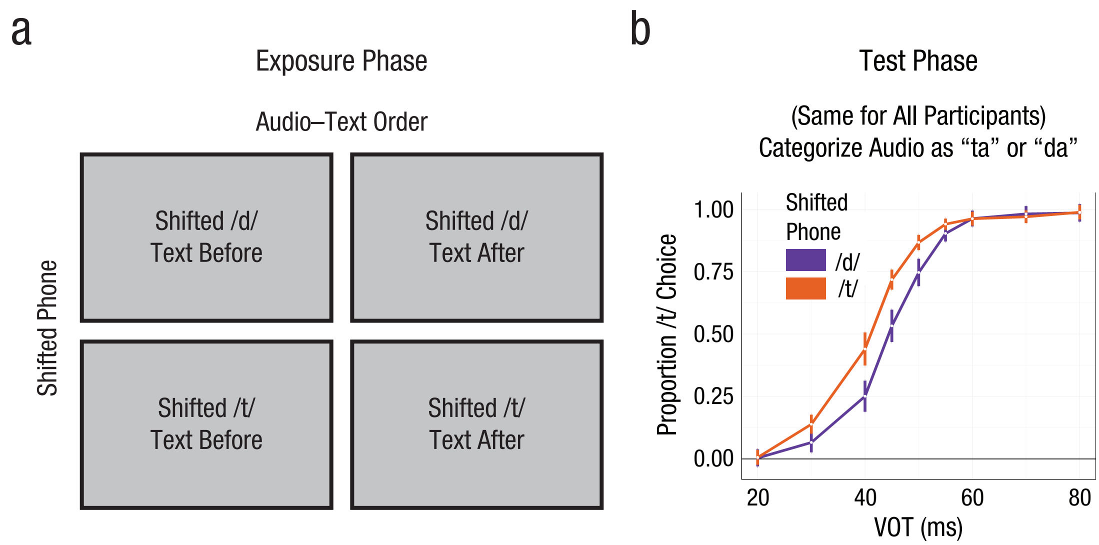

Fig. 2. Design of the exposure and test phases in both experiments. Each participant was assigned to one of four possible conditions during the exposure phase (a), which had a 2 × 2 design: shifted phone (/d/ or /t/) and audio–text order (text before or text after). All participants then completed the same task at test (b), categorizing audio on a continuum of voice-onset time (VOT) as either "ta" or "da." The graph illustrates predicted categorization patterns (separately for each shifted-phone condition) in cases in which adaptation occurs.

consisted of 128 participants after exclusions (see criteria below), which is in line with previous studies measuring similar effects (Kraljic & Samuel, 2006). Participants were approximately evenly divided among the four different exposure conditions, and test stimuli were held constant across all participant groups.

Stimuli. Target words for the exposure phase were selected by identifying minimal pairs in the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995) that are differentiated solely by an onset position /t/ versus /d/. This resulted in a list of 82 such minimal pairs, from which we manually selected 44 words (22 pairs) on the basis of part-of-speech category and approximate match of overall corpus frequency. The 98 filler words were randomly selected from the CELEX database using the following constraints: Fillers did not contain the phonemes /t/ or /d/, did not contain the orthographic letter strings "t" or "d," did not begin with a capital letter (to exclude proper nouns) or include apostrophes or hyphenation, were not longer than four syllables, were a minimum of four letters long, and had a CELEX frequency of at least 150. The full lists of both target and filler words are provided in the Supplemental Material available online.

Audio versions of each word were recorded by a 20-year-old female native speaker of American English from the Pacific Northwest who was not the experimenter. The VOT for target items was edited by splicing the onset of each "t" word onto the rime of the corresponding "d" word. The "t" onsets were trimmed in order to impose the specified VOT level (10, 60, or 100 ms) within an acceptable range of several milliseconds. Minor deviation from goal VOTs was caused by gluing onsets to rimes at zero-crossing points in order to minimize noticeable acoustic distortions. This editing procedure is consistent and generalizable but retains secondary acoustic (non-VOT) cues to voicing from the "d" rimes and thus an overall bias toward /d/ responses, which explains the higher-than-normal VOT value (60 ms) for ambiguous tokens.

Test-phase stimuli were "CV" syllables (i.e., a consonant followed by a vowel) of the following form: a /t/ or /d/ onset edited along the VOT continuum followed by the vowel /ɑ/ (pronounced as in the word "spa"). Recordings for the test items were taken from the same speaker as for the exposure stimuli, and audio manipulation was performed using the same procedure that was applied to target exposure items. As with the exposure stimuli, specified VOT levels imposed over test items varied within an acceptable range of several milliseconds.

Procedure. Participants completed the experiment in the lab with headphones. The experiment was implemented using custom JavaScript code and psiTurk (Version 2.2.3; Gureckis et al., 2016), a toolbox for conducting psychology experiments on Amazon's Mechanical Turk (MTurk). This was done to ease replication and extension using the same scripts with online participants (see Experiment 2). After providing informed consent, participants completed several questionnaires (demographics, language, attention check) before beginning the experiment. We ensured that audio was at sufficient volume by playing participants an audio captcha that required them to correctly identify numbers embedded in static noise.

Instructions prior to the exposure phase informed participants that they were completing an experiment on word comprehension and memory. Part of the instructions encouraged participants to respond to even slightly unnatural-sounding words: "Some of the audio may sound somewhat unnatural but try to ignore this. This is designed to distract you from comparing the audio to the text." This was to encourage participants to confirm that the ambiguous target items conformed to the word displayed in the subtitle.

All items in the exposure phase were played along with an accompanying text subtitle. Participants were asked to push a button to confirm whether the text and audio matched, and they received explicit feedback after each trial. All target words—regardless of audio-ambiguity status—were paired with an accurate subtitle. Seventy-eight of the ninety-eight filler items were similarly paired with accurate accompanying text. So that participants would not be distracted by some proportion of potentially unnatural-sounding audio (for the manipulated targets) and to conceal the manipulation of interest in the experiment, we randomly assigned an unrelated text subtitle to the remaining 20 filler words (e.g., audio was "coffee" but text was "green"), to which the participant was expected to press the "NO" button. The order of word trials during exposure was randomized for each participant. The use of subtitles ensured that the intended lexical (and hence phonemic) interpretation for the manipulated targets was upheld while also affording direct control of the temporal availability of disambiguating cues for integration.

For participants in the shifted-/t/ condition, visually presented "t" words were paired with ambiguous audio (60-ms VOT), whereas visually presented "d" words were paired with unambiguous audio (10-ms VOT). For those in the shifted-/d/ condition, the opposite was true: "d" words were paired with ambiguous audio (60-ms VOT), whereas "t" words were paired with unambiguous audio (100-ms VOT). This pattern is illustrated in Figure 3.

After completing the exposure phase in their assigned condition, each participant completed the same test phase—a classic phoneme-categorization task (Liberman, Harris, Hoffman, & Griffith, 1957)—consisting of 162 trials. Participants received new instructions telling them to press a button to decide whether the audio they heard was *ta* or *da*. The side of the screen on which the *ta* and *da* choices appeared was consistent within each participant but randomized between participants. On each trial, participants were exposed to audio of a CV syllable edited along a continuum between *ta* and *da*. After listening to the audio, participants were asked to judge whether the syllable contained "t" or "d." The 162 test trials were divided between two exemplar *ta/da* tokens and nine VOT levels (20, 30, 40, 45, 50, 55, 60, 70, and 80 ms), with nine repetitions for each exemplar and level (2 × 9 × 9). The order of test items was randomized within a set of nine blocks such that every stimulus was heard once before it was repeated.

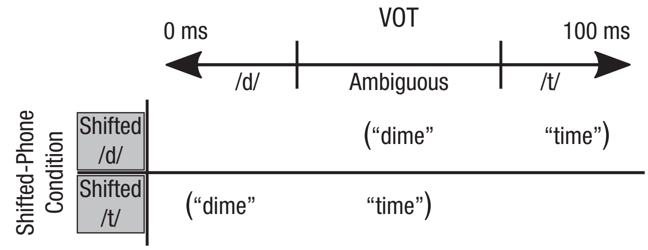

Fig. 3. Pairing of text and audio used in Experiments 1 and 2 in the shifted-/d/ and shifted-/t/ conditions. Although all participants were exposed to the same text, participants in the shifted-/d/ condition heard audio with ambiguous voice-onset times (VOTs) paired with "d" text, whereas participants in the shifted-/t/ condition heard audio with ambiguous VOTs paired with "t" text.

Predictions. In the *text-before condition*, the subtitles appeared on the screen during each trial of the exposure phase 2,000 ms prior to the start of audio (shown in Fig. 1a). Participants in the text-before condition could thus activate the correct lexical hypothesis before hearing the manipulated targets and would thus be able to map the acoustic signal to the proper interpretation independently of signal retention. Therefore, adaptation would be expected under either the signal-retention or AOC account.

In the *text-after condition*, the subtitles appeared on the screen 2 s after the audio began playing. Since the 2,000-ms gap was measured from the onset of audio, the actual gap from the end of audio to the display of text was somewhat less than 2,000 ms (between 1,000 ms and 1,500 ms, depending on the duration of the spoken word). This is illustrated in Figure 1b. If the text-after group showed the same adaptation as the text-before group, this would be in line with a signal-retention account—that is, that intermediate speech representations include information about phonetic cues. Conversely, the AOC account predicts no adaptation for the text-after group. On this view, participants are able to update their representations of the correct lexical item and properly perform the match/mismatch task during exposure, but they are unable to generalize the shifted audio because they have not stored the underlying acoustic-phonetic information required to do so.

Exclusions. We excluded participants whose match/mismatch response accuracy during exposure was less than 80% and participants whose exposure response times were less than 150 ms on more than 25% of all responses (indicating a misunderstanding or noncompliance with the task). We also excluded participants whose "da" confirmation rates during test were lower for low-VOT trials than high-VOT trials (indicating either random responses or having accidentally flipped the scale). This resulted in 128 remaining participants for analysis (exclusion rate of 3%), divided among the conditions in the following way: 33 in the text-before condition with shifted /t/, 30 in the text-before condition with shifted /d/, 36 in the text-after condition with shifted /t/, and 29 in the text-after condition with shifted /d/.

Analysis. A mixed-effects logistic regression analysis was conducted on trial-level data. The main dependent variable was "t" responses—whether participants chose the "t" or "d" item on each trial of the categorization task. The independent variables were experimental condition: shifted phone (shifted /t/ vs. shifted /d/, sum-coded), timing (text before and text after, sum-coded), and their interaction. VOT (continuous variable, scaled and centered) and test half (first vs. second, sum-coded) were included as main effects and interaction terms with experimental conditions to test whether the effects of interest changed over the course of the test phase; this followed previous observations (e.g., Liu & Jaeger, 2018) that perceptual adaptations may be unlearned, to some degree, throughout testing. We attempted to include block number (1 to 9, centered) as a factor, but no models with this factor converged, so we used test half (first four blocks vs. last five blocks) instead. We used the maximal random-effects structure that converged; this structure included random intercepts for participants and test exemplars, VOT as random slopes for participant and test exemplars, and condition (shifted phone and timing) as random slopes for exemplar (Barr, 2013). Full model structures are available in the Supplemental Material. We tested for significance of factors in models by using likelihood-ratio tests on the χ^2 values from nested model comparisons with the same random-effect structure (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). We computed Bayes factors (BFs), where appropriate, to quantify the degree of support in favor of accepting or rejecting null hypotheses. All BFs were computed in R using the *brms* package (Bürkner, 2017) with default parameters, except where required for accurate estimation of posterior probabilities (see the Supplemental Material). All data and R analysis code are available on the Open Science Framework at https://osf.io/wg6de/.

Results

In the exposure phase, performance of the participants included in the analysis was high and was comparable across conditions: Accuracy in confirming the audio–subtitle match was above 99% on unambiguous target items, above 96% on ambiguous targets, and above 97% on fillers. This suggests that for the included participants, the matching task at exposure was not notably more difficult within one set of exposure conditions than another. Indeed, a mixed model with a main effect of timing was not a better fit to exposure accuracy on ambiguous targets than one that included only random effects, $\chi^2(1) = 0.42$, $p = .519$ (BF = 0.32). This high accuracy (above 96%) on ambiguous targets in the text-after condition suggests that participants held an intermediate representation over time between hearing the word and seeing the text. What type of representation this was can be revealed only by examining the adaptation patterns from the test phase.

Results from the test phase appear in Figure 4 (split by shifted phone and timing). As can be observed, adaptation was successful: The psychometric functions are different between the shifted-/t/ and -/d/ ranges. Remarkably, and as predicted by the AOC account, such an effect was observed only in the text-before condition; the categorization functions are not reliably different in the text-after condition as a function of shift direction (i.e., adaptation did not occur in the text-after condition). The adaptation additionally began to fade over time: The magnitude of adaptation (in conditions in which it was present) was larger in the first half of testing than in the second half of testing (see the Supplemental Material for additional visualizations). This reduction in the adaptation effect over time is in line with previous findings (Liu & Jaeger, 2018) and is perhaps not surprising given the remarkably limited sample during exposure (only 22 edited tokens out of 142 total) and the comparatively long testing phase.

These results were confirmed in comparisons of mixed-effects models. First, we compared models over all of the data. The best-fitting model was one that included a main effect of VOT and a main effect of test half, with main effects and interactions of shifted phone, timing, and test half. This model was a better fit than one that did not include the interaction of shifted phone and timing and the three-way interaction of shifted phone, timing, and test half, $\chi^2(2) = 6.42$, $p = .040$, and better than a model without the three-way interaction of shifted phone, timing, and test half, $\chi^2(1) = 5.66$, $p = .017$. These modeling results demonstrate that adaptation was higher in the text-before than in the text-after condition and that the adaptation effect faded over time during the test phase.

Given the significant three-way interaction of shifted phone, timing, and test half, we next tested for the effects of interest (shifted phone and timing) in each test half separately. In the first half, the best-fitting model was one that included main effects and interactions of VOT, shifted phone, and timing (Table 1). This

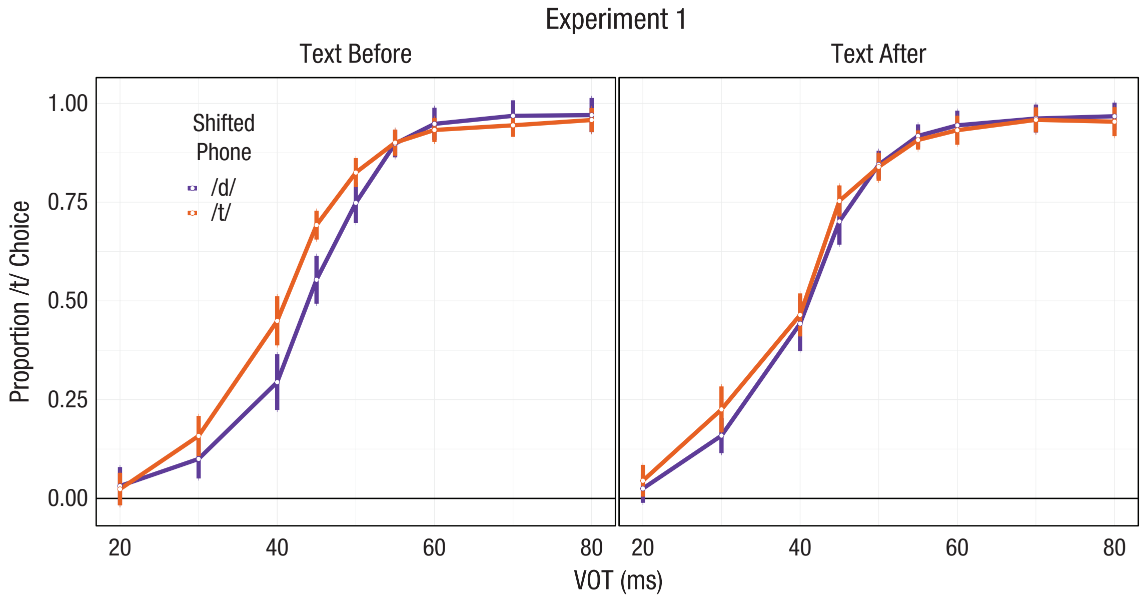

Fig. 4. Psychometric functions for Experiment 1: proportion of /t/ choices as a function of voice-onset time (VOT) and shifted-phone condition (/t/ or /d/), plotted separately for the text-before and text-after conditions. Data points are the average of participant means, and error bars are within-subject 95% confidence intervals.

model was a better fit than one that did not include the interaction of shifted phone and timing or their interaction with VOT, $\chi^2(2) = 13.79$, $p = .001$, and better than one that did not include the three-way interaction of VOT, shifted phone, and timing, $\chi^2(1) = 11.28$, $p < .001$. In contrast, in the second half, the best-fitting model was one that included main effects of VOT, shifted phone, and timing, and interactions of VOT and shifted phone and of VOT and timing but no interaction of shifted phone and timing. A model with the additional interaction of shifted phone and timing was not a significant improvement, $\chi^2(1) = 0.26$, $p = .613$, nor was one with the additional three-way interaction of VOT, shifted phone, and timing, $\chi^2(2) = 1.37$, $p = .503$. These modeling results confirm that the timing-specific adaptation effect was present only in the first half of the test phase and faded in the second half. Additionally, the interaction between VOT and other fixed effects was expected because adaptation is understood to represent a change in participants' criteria for "t" versus "d" categorization. Rather than remaining consistent throughout the VOT continuum (as might occur if, instead, participants had learned a general bias toward one phone or the other), this shift manifests most strongly for otherwise ambiguous stimuli.

Next, we directly compared the effect of shifted phone separately in the two timing conditions (text before and text after) to confirm that the effect was indeed only present in the text-before condition and not in the text-after condition (first half of test phase only). For the text-before condition, the best-fitting model was one that included main effects of VOT and shifted phone. This model was a better fit than one that did not include the effect of shifted phone, $\chi^2(1) = 6.92$,

Table 1. Output of the Best-Fitting Model Predicting /t/ Responses on the First Half of Test Trials in Experiment 1

Predictor	β	*z*	*p*	Odds ratio
Intercept	1.84 [0.45, 3.23]	2.6	.009	6.29 [1.57, 25.17]
VOT	3.47 [3.31, 3.63]	43.13	< .001	32.2 [27.5, 37.7]
Shifted phone	−0.12 [−0.34, 0.1]	−1.08	.282	0.89 [0.71, 1.1]
Timing	−0.23 [−0.45, −0.01]	−2.08	.038	0.79 [0.64, 0.99]
VOT × Shifted Phone	0.24 [0.1, 0.37]	3.41	< .001	1.27 [1.11, 1.45]
VOT × Timing	0.02 [−0.12, 0.15]	0.24	.811	1.02 [0.89, 1.17]
Shifted Phone × Timing	−0.26 [−0.48, −0.04]	−2.33	.02	0.77 [0.62, 0.96]
VOT × Shifted Phone × Timing	−0.23 [−0.37, −0.1]	−3.33	< .001	0.79 [0.69, 0.91]

Note: Bracketed values are 95% confidence intervals. VOT = voice-onset time.

$p = .008$ (BF = 19.13). In contrast, in the text-after condition, the best-fitting model was one that included only the main effect of VOT. A model with the additional main effect of shifted phone was not a better fit, $\chi^2(1) = 0.01$, $p = .906$ (BF = 0.32). These modeling results demonstrate that the adaptation effect was not simply greater in the text-before condition than in the text-after condition but that no adaptation effect was statistically detectable in the text-after condition in our data.

We additionally performed several secondary analyses to investigate factors that could instead contribute to the lack of adaptation in the text-after condition. Overall, we found no notable differences in participants' behavior during the exposure task: Accuracy on target items during the exposure phase was consistently high across conditions (see above), and further analyses (available in the Supplemental Material) showed no relationship between exposure-trial response times and test behavior, nor any evidence of bimodality in participant categorization performance within exposure condition (see the Supplemental Material). Indeed, these kinds of lexically guided adaptation effects are surprisingly easy to induce in a range of tasks with different demands, including word counting, syntactic judgments, or loudness judgments (Drouin & Theodore, 2018; McQueen, Norris, & Cutler, 2006), provided that listeners properly resolve ambiguous audio to the right phonological categories.

Lastly, the adaptation attested in text-before participants was mainly driven by the shifted-/d/ condition and not the shifted-/t/ condition. In model comparisons using data from each shifted phone condition separately (first half of test phase only), a model with a main effect of timing was significant for the shifted-/d/ group, $\chi^2(1) = 8.69$, $p = .003$, but not the shifted-/t/ group, $\chi^2(1) < 0.001$, $p = .997$. Perhaps this was due to interference from secondary acoustic cues to voicing, such as pitch or vowel length. Indeed, an examination of exposure accuracy (i.e., confirming the subtitle as a match to the audio) on ambiguous target items across /d/ and /t/ conditions was consistent with such an interpretation: Prior to participant exclusions, the mean accuracy in the shifted-/t/ groups (both text before and text after) was 93%, whereas for shifted-/d/ groups, it was 98%. Nevertheless, this /t/ versus /d/ asymmetry does not impact the main theoretical interpretation with respect to the signal-retention or AOC accounts, and we took steps to address this in Experiment 2, which we discuss below.

Discussion

Overall, we observed adaptation effects: The condition of shifted phone (/t/ vs. /d/) during exposure was successful at modulating participants' psychometric functions in a phoneme-categorization task. Crucially, this adaptation to the exposure phase occurred only when participants received disambiguating information before the acoustic input (text-before condition). Such adaptation did not occur in the text-after condition, when the acoustic stimulus ended before the disambiguating information was viewed. These results support the AOC theory and are inconsistent with signal retention.

Experiment 2

In Experiment 2, we aimed to replicate the main findings from Experiment 1 while confirming that the effects of interest are robust to minor experimental modifications.[1] The design was the same except that we additionally manipulated pitch to remove the main secondary acoustic cue to voicing, utilized a norming study to select the maximally ambiguous VOT level for target items, made minor adjustments to display timing to better equate conditions, and sampled participants from an online subject pool.

Method

Design. Experiment 2 matched the design from Experiment 1 but with a change to the display timing. The timing for the exposure phase in Experiment 1 was as follows. In the text-before condition, participants saw text for 2,000 ms before the corresponding audio was played. However, the text remained on screen throughout the presentation of the audio until the participant responded with a match/mismatch judgment. In the text-after condition for Experiment 1, the audio was played first, and then after a gap of 2,000 ms (from the onset of audio), the text subtitle appeared and remained on screen until a match/mismatch judgment was provided. There was thus an asymmetry in the duration of text availability between conditions: Text-before participants in Experiment 1 saw the subtitles for longer than the text-after participants. To address this, we adjusted the display timing for Experiment 2. For text-before participants in Experiment 2, the subtitle appeared on screen for a fixed duration of 875 ms. Then there was a gap of 1,125 ms during which a blank screen was displayed prior to the audio. Audio was then played with nothing on screen. Immediately following the end of the audio, instructions were shown prompting participants for a match/mismatch judgment (which did not include the original subtitle). In the text-after condition for Experiment 2, participants first heard the audio (while viewing a blank screen). After a gap of 2,000 ms from audio onset, the subtitle appeared for a fixed duration of 875 ms, after which participants saw instructions to provide a match/mismatch judgment that, as in the text-before condition, did not include the original subtitle. The design, exclusions, and analyses were all preregistered (https://osf.io/x6r5t/).

Participants. Power analyses of the results from Experiment 1 suggested that we would have 90% power to detect the effect with approximately 37 participants in each condition, or 148 participants total. Given additional expected dropout from running the experiment online rather than in the lab, we recruited 194 participants using MTurk and divided them among the same four exposure conditions as in Experiment 1 (text-before condition with shifted /d/, text-before condition with shifted /t/, text-after condition with shifted /d/, and text-after condition with shifted /t/). Somewhat more participants were assigned to the text-before condition overall (n = 106) than the text-after condition (n = 88) because of an initial glitch in the online platform. Participants were paid $2.41 for taking part in the experiment.

Stimuli. The materials were the same as in Experiment 1, except that target items in the exposure phase were pitch corrected according to the following procedure. The audio for target stimuli in Experiment 1 was created by gluing different portions of "t"-word onsets onto the rime of the "d" words. Pitch contour (F_0), which is a secondary cue to voicing (Dmitrieva, Llanos, Shultz, & Francis, 2015), is realized on the following vowel, which means that although the VOT values were edited, all the target stimuli retained secondary information consistent with voicing (i.e., the "d" interpretation). To correct for this, we edited new versions of the target audio that were corrected for pitch (F_0). We manually extracted the pitch contours for each word pair and selected a new F_0 onset value at two thirds of the gap between the "d"-onset and "t"-onset words. We resynthesized the pitch contours of the "d"-onset words with a new contour that began at the designated two-thirds-boosted F_0 value and followed a smooth cline (using pseudolinear interpolation with a step-size of 10 ms) down to the original "d"-word pitch at 160 ms into the vocoid.

We conducted a norming study on a separate group of 44 participants (see the Supplemental Material for the full design) to identify the ideal ambiguity point for VOT. For the new pitch-corrected target stimuli, we identified the median VOT at which items were classified equally often as the corresponding word beginning with /t/ or /d/ (46.9 ms) and used the VOT from our tested range closest to this (45 ms) as the cutoff for ambiguous targets in Experiment 2. The test stimuli remained unchanged from Experiment 1 (without pitch correction) in order to minimize cross-experiment differences.

Procedure. Participants completed the experiment in a Web browser using the same interface as in Experiment 1. The only change to the procedure was that we enforced headphone use through a more stringent audio captcha (Woods, Siegel, Traer, & McDermott, 2017). Specifically, participants were asked to provide loudness judgments on a sequence of tones that were either in matching phase or in antiphase between the stereo channels. Because phase differences are greatly attenuated over loudspeakers, accurate performance on the captcha task was possible only with headphone use. The remainder of the procedure was identical to that of Experiment 1. The changes were only to the audio stimuli used for target items during the exposure phase, and the differences were imposed to better equate the display duration of text in the text-before and text-after conditions. Exclusions and analyses were identical to those in Experiment 1. This resulted in 169 remaining participants for analysis (exclusion rate of 13%), divided among the conditions in the following way: 50 in the text-before condition with shifted /t/, 44 in the text-before condition with shifted /d/, 36 in the text-after condition with shifted /t/, and 39 in the text-after condition with shifted /d/. The gap in the final distribution of participants across conditions was because of an initial difference in assignment, with exclusion rates remaining similar (11.3% for text-before participants and 14.8% for text-after participants). The increased exclusion rates in Experiment 2 were primarily driven by participants whose exposure response times were less than 150 ms on more than 25% of all responses. Exclusion rates for match/mismatch inaccuracy were about 3% and were comparable with those in Experiment 1 across both timing conditions.

In our preregistered analysis plan for Experiment 2, we had an additional criterion to exclude participants whose performance at the extrema of the VOT distributions (20 ms and 80 ms) was more than 0.15 away from ceiling or floor. We added this additional exclusion to the preregistration after observing that some participants' psychometric functions in Experiment 1 did not conform to the usual S shape because of deviance from floor or ceiling performance at the extrema. However, we ultimately decided to diverge from this preregistered criterion because there was no theoretical reason to expect categorizations at our chosen extrema (e.g., 20 ms VOT) to necessarily be at floor or ceiling. Excluding these participants did not qualitatively change the reported results in any of the analyses, and results with this exclusion criterion are reported in the Supplemental Material.

Results

In the exposure phase, performance of the participants included in the analysis was high and was comparable across conditions: Accuracy in confirming the audio–subtitle match was above 97% on unambiguous target

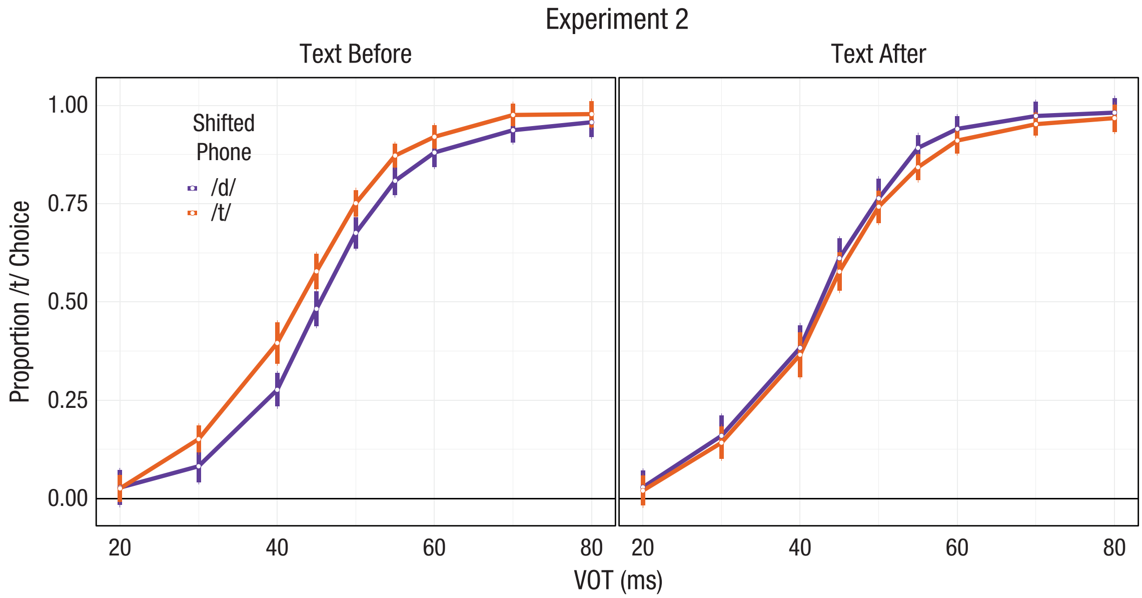

Fig. 5. Psychometric functions for Experiment 2: proportion of /t/ choices as a function of voice-onset time (VOT) and shifted-phone condition (/t/ or /d/), plotted separately for the text-before and text-after conditions. Data points are the average of participant means, and error bars are within-subject 95% confidence intervals.

items, above 95% on ambiguous targets, and above 96% on fillers. This suggests that for the included participants, the matching task at exposure was no more difficult in one condition than another. A mixed model with a main effect of timing was not a better fit to exposure accuracy on ambiguous targets than a model including only random effects, $\chi^2(1) = 2.38$, $p = .123$ (BF = 1.35). Of particular note is that, as in Experiment 1, high accuracy on ambiguous targets in the text-after condition suggested that participants held an intermediate representation over time between hearing the word and seeing the text. The content of this representation can be revealed only by examining the adaptation patterns from the test phase.

Data from the test phase appear in Figure 5 (split by shifted phone and timing). As can be observed, adaptation was successful: The psychometric functions are different between the shifted-/t/ and -/d/ ranges. Remarkably, and again as predicted by the AOC account, such an effect was observed only in the text-before condition; the categorization functions were not reliably different in the text-after condition as a function of shift direction (i.e., adaptation did not occur in the text-after condition). Unsurprisingly, and as in Experiment 1, this effect faded over time: The magnitude of adaptation was numerically larger in the first half of the test phase and diminished by the second half.

The results were confirmed in comparisons of mixed-effects models. First, we compared models over all of the data. The best-fitting model was one that included a main effect of VOT and a main effect of test half, with main effects and interactions of shifted phone and timing. This model was a better fit than one that did not include the interaction of shifted phone and timing, $\chi^2(1) = 6.05$, $p = .014$. A model with interactions of shifted phone and timing with test half did not improve the fit, $\chi^2(2) = 4.45$, $p = .108$, nor did one with the three-way interaction of shifted phone, timing, and test half, $\chi^2(3) = 5.04$, $p = .169$. These modeling results demonstrate that adaptation was higher in the text-before condition than the text-after condition and that the effect was relatively consistent throughout the test phase.

Next, although a model with the three-way interaction of shifted phone, timing, and test half was not a significantly better fit, we nonetheless tested for the effects of interest (shifted phone and timing) in each test-phase half separately, as we did in Experiment 1. In the first half of the test phase, the best-fitting model was indeed one that included a main effect of VOT and main effects and interactions of shifted phone and timing (Table 2). This model was a better fit than one that did not include the interaction of shifted phone and timing, $\chi^2(1) = 5.69$, $p = .017$, and better than one that included only a main effect of VOT, $\chi^2(3) = 8.44$, $p = .038$. Likewise, in the second test half, there was still an interaction effect of shifted phone and timing: A model that included a main effect of VOT and main effects and interactions of shifted phone and timing was significantly better than one without the interaction of shifted phone and timing, $\chi^2(1) = 3.95$, $p = .047$. However, this effect was more subtle; this model was

Table 2. Output of the Best-Fitting Model Predicting /t/ Responses on the First Half of Test Trials in Experiment 2

Predictor	β	*z*	*p*	Odds ratio
Intercept	1.68 [0.09, 3.28]	2.06	.039	5.37 [1.09, 26.46]
Voice-onset time	4.02 [3.76, 4.28]	30.45	< .001	55.58 [42.91, 71.98]
Shifted phone	−0.11 [−0.34, 0.13]	−0.89	.372	0.9 [0.71, 1.14]
Timing	−0.15 [−0.39, 0.08]	−1.28	.201	0.86 [0.68, 1.09]
Shifted Phone × Timing	−0.29 [−0.53, −0.05]	−2.4	.016	0.75 [0.59, 0.95]

Note: Bracketed values are 95% confidence intervals.

not significantly better than one with only a main effect of VOT, $\chi^2(3) = 6.23$, $p = .101$. Together, these modeling results confirm that the timing-specific adaptation effect was present in both halves of the test phase, although it was not as robust in the second half.

Next, as in Experiment 1, we directly compared the effect of shifted phone separately in the two timing conditions (text before and text after) to confirm that the effect was indeed present in the text-before condition but not in the text-after condition (first half of test phase only). For the text-before condition, the best-fitting model was one that included main effects of VOT and shifted phone. This model was a better fit than one that did not include the effect of shifted phone, $\chi^2(1) = 6.18$, $p = .013$ (BF = 3.83). In contrast, in the text-after condition, the best-fitting model was one that included only the main effect of VOT. A model with the additional main effect of shifted phone was not a better fit, $\chi^2(1) = 0.90$, $p = .344$ (BF = 0.92). Whereas the BF of 0.92 on its own was essentially ambiguous for or against the null model, the alternative model (i.e., one including an effect of shifted phone) actually contained a weak trend in the opposite direction of the original acoustic signal: The overall rate of "t" choices was negligibly higher for participants in the text-after condition with shifted /d/ than it was for those in the text-after condition with shifted /t/. These modeling results demonstrate that the adaptation effect was not simply greater in the text-before condition than in the text-after condition but that an adaptation effect was not statistically detectable in the text-after condition at all in our data.

Lastly, the additional steps we took to address the asymmetry between the shifted-/d/ and -/t/ conditions did not appear to succeed. When examining effect of timing condition separately in the two shifted-phone conditions, we found that the effect of timing was significant in the shifted-/d/ condition, $\chi^2(1) = 7.28$, $p = .007$, but not the shifted-/t/ condition, $\chi^2(1) = 0.66$, $p = .416$. Although this may have been caused by residual voicing cues (e.g., vowel length), the asymmetry did not interact with either of the primary theories (signal retention vs. AOC) under discussion.

Discussion

Experiment 2 replicated the primary findings from Experiment 1. Adaptation to the exposure phase was observed when participants received disambiguating information before the acoustic signal (text-before condition), but not after (text-after condition). These findings were robust to a display-timing change and the additional manipulation of pitch in tandem with VOT.

General Discussion

In two experiments, we observed that listeners can adapt to speaker-specific acoustic cues to phone perception (i.e., VOT) but only when disambiguating information is provided before rather than after they hear the ambiguous acoustic input. When disambiguating text appeared after the ambiguous speech, listeners could verify and accept either lexical alternative (depending on condition, either *time* or *dime*), but they could not use this disambiguating text to learn the particular VOT-to-phone mapping. Only when the order was reversed (text, then speech) could listeners both verify the intended word and adapt. This finding is consistent with the AOC account and inconsistent with the signal-retention account of speech processing. According to the AOC account, graded activation of linguistic categories (e.g., phones, words) persists over time, but the acoustic evidence that gave rise to this probabilistic information does not. Maintenance of probabilistic information about linguistic categories permits the accurate lexical verification during the exposure phase of the text-after condition but blocks the ability to adapt because the acoustic cues were not retained. Even the most course-grained representation of acoustic cues would have been sufficient for adaptation (i.e., adaptation would have been possible if the system had represented continuous VOT values as being either "high" or "low," given that VOT values during exposure were at opposing ends of the continuum), yet adaptation did not occur.

Such a finding is consistent with the demands of real-time language processing. Consider how little is lost by not retaining VOT information compared with how much is gained in performance by storing probabilistic activation over higher-level categories. Indeed, we know of no linguistic phenomenon that requires the computation of long-distance dependencies (over seconds) between acoustic cues and later-arriving linguistic input, but dependencies abound for linguistic categories, such as phonemes and words, over which phonological and syntactic systems traffic, respectively. This likely reflects a general property of perception and cognition over time: Lower-level representations may be fast-changing and ephemeral, mirroring the input, whereas intermediate and higher-level categories are more persistent, given their need for inference and integration.

Our experiments, though, can speak directly to intermediate speech representations only on the timescale of about 1 s and beyond. Indeed, neuroimaging studies (e.g., Toscano, Anderson, Fabiani, Gratton, & Garnsey, 2018; Toscano, McMurray, Dennhardt, & Luck, 2010) indicate that acoustic detail is present during early cortical processing for up to 200 ms. This suggests a more refined AOC account under which early perceptual representations are built on the basis of acoustic cues over the first few hundred milliseconds, with information passed on to higher-level categories beyond that. An alternative possibility is that although the fingerprint of acoustic cues can be detected during early cortical processing, this information is not available to the components of the cognitive system used for subsequent interpretation. Such a modular variant of the AOC account would provide a mechanism in support of previously identified limits on perceptual learning: Jesse and McQueen (2011) found that Dutch listeners adapt to speech when ambiguous targets appear word-medially ("bene[f/s]it") or word-finally ("regre[ss/ff]") but not word-initially ("[f/s]reedom"). They suggested a *timing hypothesis* that proposed that relevant lexical knowledge must be available before a listener hears the ambiguous sound to support adaptation. The AOC theory offers an explanation of why such a timing hypothesis would be true, namely that intermediate representations of speech consist of activated linguistic categories, not subphonemic or acoustic information. Future work is required to disentangle these two variant AOC accounts and related questions on a narrower timescale.

On a broader timescale, the AOC theory clarifies the interpretation of listeners' sensitivity to within-category acoustic variation. Past work showing that performance on memory tasks depends on acoustic clarity (Crowder & Morton, 1969; Frankish, 2008) or that sensitivity is maintained across syllables (Brown-Schmidt & Toscano, 2017; Falandays, Brown-Schmidt, & Toscano, 2020; McMurray et al., 2009) or integrated over a delay (Galle et al., 2019; Gwilliams, Linzen, Poeppel, & Marantz, 2018) did not address the internal contents of the representations that support such sensitivity. The present findings provide direct evidence in favor of the position that gradience is maintained through probabilistic uncertainty about potential categories. Similarly, although acoustic maintenance may appear to be supported by findings that unsupervised exposure or time-delayed subtitles may attenuate the processing difficulties associated with unfamiliar accents (e.g., Bradlow & Bent, 2008; Burchill, Liu, & Jaeger, 2018), such adaptation can also be accomplished under an AOC account through listeners' use of contextual information to predict upcoming words and adjust to the bottom-up mapping accordingly. Such a top-down mechanism finds support in recent electrophysiological evidence (Getz & Toscano, 2019). Likewise, infants' difficulty processing unfamiliar variants of their native languages (Cristia et al., 2012) is overcome when words are embedded within the context of highly familiar stories (van Heugten & Johnson, 2014). Thus, although there are experimental conditions that prevent adaptation from occurring (e.g., our text-after condition), being able to predict and activate upcoming linguistic material before the corresponding signal arrives (Jesse, 2021) compensates for the restrictions imposed by the immediacy of computation. Category representations provide the bridge that supports listeners' adaptation to variability despite computational and structural restrictions around the ephemeral signal.

Transparency

Action Editor: Sachiko Kinoshita
Editor: Patricia J. Bauer
Author Contributions
All of the authors developed the study concept and contributed to the study design. S. Caplan drafted the manuscript, and all of the authors were involved in revision. S. Caplan designed and edited the stimuli. A. Hafri programmed the experiments. All of the authors contributed to analysis decisions. S. Caplan and A. Hafri implemented the analyses. All of the authors approved the final manuscript for submission.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
Partial funding for the experiments was provided by an Integrated Language Sciences and Technology Small Grant from the Mind Center for Outreach, Research, and Education (MindCORE) initiative at the University of Pennsylvania.
Open Practices
All data, analysis code, and materials have been made publicly available via OSF and can be accessed at https://osf.io/wg6de/. The design and analysis plans for Experiments 1, 2, and S1 were preregistered at https://osf.io/

wg6de/ (as Experiments 1, 3, and 2, respectively). A divergence from the preregistration of Experiment 2 is discussed in the text. This article has received the badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

TC

ORCID iD

Spencer Caplan https://orcid.org/0000-0003-2733-8100

Acknowledgments

We thank Joseph Toscano for important discussions and commentary, along with Wednesday Bushong for clarifications of her study (Bushong & Jaeger, 2017). We also thank Mark Liberman, Mitch Marcus, and Charles Yang for helpful comments.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797620968787

Note

1. An initial version of this study introduced a confound between stimulus editing and phonological category (reported as Experiment S1 in the Supplemental Material).

References

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (Release 2). Philadelphia, PA: Linguistic Data Consortium.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*, Article 328. doi:10.3389/fpsyg.2013.00328

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592–597.

Bicknell, K., Jaeger, T. F., & Tanenhaus, M. K. (2016). Now or . . . later: Perceptual data are *not* immediately forgotten during language processing. *Behavioral and Brain Sciences*, *39*, Article e67. doi:10.1017/S0140525X15000734

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707–729.

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, *32*, 1211–1228.

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PLOS ONE*, *13*(8), Article e0199358. doi:10.1371/journal.pone.0199358

Bürkner, P.-C. (2017). brms: An R Package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). doi:10.18637/jss.v080.i01

Bushong, W., & Jaeger, T. F. (2017). Maintenance of perceptual information in speech perception. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 186–191). Austin, TX: Cognitive Science Society.

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral & Brain Sciences*, *39*, Article e62. doi:10.1017/S0140525X1500031X

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

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. *Journal of Memory and Language*, *30*, 234–250.

Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., & Floccia, C. (2012). Linguistic processing of accented speech across the lifespan. *Frontiers in Psychology*, *3*, Article 479. doi:10.3389/fpsyg.2012.00479

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics*, *5*, 365–373.

Darwin, C. J., & Baddeley, A. D. (1974). Acoustic memory and the perception of speech. *Cognitive Psychology*, *6*, 41–60.

Dmitrieva, O., Llanos, F., Shultz, A. A., & Francis, A. L. (2015). Phonological status, not voice onset time, determines the acoustic realization of onset *f*0 as a secondary voicing cue in Spanish and English. *Journal of Phonetics*, *49*, 77–95.

Drouin, J. R., & Theodore, R. M. (2018). Lexically guided perceptual learning is robust to task-based changes in listening strategy. *The Journal of the Acoustical Society of America*, *144*, 1089–1099.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review*, *112*, 912–931.

Falandays, J. B., Brown-Schmidt, S., & Toscano, J. C. (2020). Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language*, *112*, Article 104088. doi:10.1016/j.jml.2020.104088

Frankish, C. (2008). Precategorical acoustic storage and the perception of speech. *Journal of Memory and Language*, *58*, 815–836.

Galle, M. E., Klein-Packard, J., Schreiber, K., & McMurray, B. (2019). What are you waiting for? Real-time integration of cues for fricatives suggests encapsulated auditory memory. *Cognitive Science*, *43*, Article e12700. doi:10.1111/cogs.12700

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Ganong, W. F., III. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Getz, L. M., & Toscano, J. C. (2019). Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological Science*, *30*, 830–841.

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

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . . Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*, 829–842.

Gwilliams, L., Linzen, T., Poeppel, D., & Marantz, A. (2018). In spoken word recognition, the future predicts the past. *The Journal of Neuroscience*, *38*, 7585–7599.

Jesse, A. (2021). Sentence context guides phonetic retuning to speaker idiosyncrasies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*, 184–194. doi:10.1037/xlm0000805

Jesse, A., & McQueen, J. M. (2011). Positional effects in the lexical retuning of speech perception. *Psychonomic Bulletin & Review*, *18*, 943–950. doi:10.3758/s13423-011-0129-2

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*, 262–268.

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

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1–71.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29–63. doi:10.1016/0010-0285(78)90018-X

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*, 65–91.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech*, *49*, 101–112.

Munson, C. M. (2011). *Perceptual learning in speech reveals pathways of processing* [Doctoral dissertation]. University of Iowa, Iowa City. Retrieved from https://ir.uiowa.edu/cgi/viewcontent.cgi?article=2727&context=etd

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*, 204–238.

Postle, B. R. (2015). The cognitive neuroscience of visual short-term memory. *Current Opinion in Behavioral Sciences*, *1*, 40–46.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 539–555.

Sagi, D. (2011). Perceptual learning in vision research. *Vision Research*, *51*, 1552–1566. doi:10.1016/j.visres.2010.10.019

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*, 1207–1218.

Schuler, K. D., Kodner, J., & Caplan, S. (2020). Abstractions are good for brains and machines: A commentary on Ambridge (2020). *First Language*, *40*, 631–635. doi:10.1177/0142723720906233

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts* (Vol. 9). Cambridge, MA: Harvard University Press.

Toscano, J. C., Anderson, N. D., Fabiani, M., Gratton, G., & Garnsey, S. M. (2018). The time-course of cortical responses to speech revealed by fast optical imaging. *Brain and Language*, *184*, 32–42.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*, 1532–1540.

van Heugten, M., & Johnson, E. K. (2014). Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General*, *143*, 340–350. doi:10.1037/a0032192

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*, 2064–2072.

Zellou, G., & Dahan, D. (2019). Listeners maintain phonological uncertainty over time and across words: The case of vowel nasality in English. *Journal of Phonetics*, *76*, Article 100910. doi:10.1016/j.wocn.2019.06.001