Sublexical ARTifacts: Bottom-up Interference in a Lexical Category Search

William Clapp (wsclapp@stanford.edu)

Nora Dee (noradee@stanford.edu)

Meghan Sumner (sumner@stanford.edu)
Department of Linguistics, Margaret Jacks Hall, Building 460
Stanford, CA 94305

Abstract

How listeners adapt to unfamiliar talkers and accents is a central question in psycholinguistics. In this study, we explored how listeners dynamically shift mappings from acoustic information to mental representations after hearing a new talker via novel eye-tracking methods. We tested a prediction from Adaptive Resonance Theory (ART) that an anomaly in the signal (in this case, a change in talker) increases the influence of bottom-up relative to top-down information, creating an environment where sublexical competitors (e.g. *Arch* within *Archer*) would be more likely interfere with lexical access for the target. In two experiments (Exp. 1: General American English [GA] talkers; Exp. 2: GA and Spanish-accented [SP] talkers), this prediction was supported via analyses of accuracy, latency, and gaze. In Exp. 2, we found that the effect replicated but did not differ based on the accent of the talker. The data suggest new paths forward in speech adaptation research.

Keywords: psycholinguistics; speech perception; talker adaptation; language comprehension; phonetics; eye-tracking

Introduction

For the last half century, one of the most important questions in speech perception research has been: How are we able to understand speech without invariant acoustic cues that uniquely identify speech sounds (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967)? A highly productive thread of research has explored how even the same physical acoustic characteristics (VOT, center of gravity, formant ratios, etc.) can be interpreted differently based on context, talker, and experience (Kraljic & Samuel, 2006; Norris, McQueen, & Cutler, 2003; Xie, Jaeger, & Kurumada, 2023). This line of work has illuminated how category boundaries shift after exposure to informative exemplars, but we still know relatively little about the underlying cognitive dynamics of rapid adaptation to novel talkers, which we tend to experience outside the lab as a nearly instantaneous process without any lengthy exposure phase.

To explore these dynamics, we turn to the unified theory of mind Adaptive Resonance Theory (ART, Grossberg, 2013, 2021). ART makes explicit predictions about how categories are activated and updated in real time during perception and cognition. In this theory, recognition is achieved when input patterns (*bottom-up*) and category-level mental states adaptively tuned through prior experience (*top-down*) are sufficiently matched to form neural resonance, leading to a conscious percept. Crucially, this approach is able to achieve stable percepts without a need for context-independent representations at any particular linguistic level, such as phonemes, words, or allophones (Grossberg, Boardman, & Cohen, 1997). In recent decades, this model has been applied to speech perception and language understanding in part to overcome problems inherent to approaches that posit a perceptual substrate of discrete linguistic units (Goldinger & Azuma, 2003; Samuel, 2020).

One property of the ART architecture is that during conscious perception, coarser categories *mask* finer-grained categories (Kazerounian & Grossberg, 2014). For example, someone who has successfully understood a sentence may not be consciously aware of the words that compose it or the speech sounds that compose those words (at least not without attention explicitly directed towards those features). However, when no immediately available category node provides a sufficient match to the bottom-up input pattern to produce a resonant state, a *category search* is triggered. In this case, a wider net is cast to either find a better matching but less active category, or a new category is formed (Grossberg, 2021). When a category search is underway, this masking effect may be dampened, making finer-grained patterns in the input more consciously perceptible than they would otherwise be.

A testable hypothesis falls out of this architecture which may inform our understanding of how listeners are able to dynamically adapt to novel talkers by remapping acoustic patterns to mental categories. Specifically, when a listener hears a new voice or accent, the mismatch between top-down and bottom-up information may trigger a category search, reducing the masking effect, and leading to greater interference from bottom-up, sublexical information. In the present study, we tested this hypothesis by inducing a category search and measuring sublexical interference using eye-tracking in the visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Critical stimulus sentences contained multisyllabic target nouns which in turn contained separate, semantically unrelated monosyllabic nouns (e.g., [ARCH]er; bi[KEY]ni). [1] After being habituated to the voice of a primary talker (*anchor talker*), some critical trials included a *voice switch*, where the voice of an unfamiliar talker (*intru-*

[1] This approach is related but not identical to other traditions in psycholinguistics exploring phonological competition effects in online processing (e.g., Allopenna, Magnuson, & Tanenhaus, 1998; Gaskell & Marslen-Wilson, 1997).

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

sive talker) was spliced into the sentence at the onset of target.

We hypothesized that when a listener's top-down expectations were violated in this way, a category search would be triggered, reducing the masking effect, and leading to increased interference from sublexical patterns in the speech signal. We predicted that this effect would be amplified by the visual presence of the sublexical competitor (e.g., *Arch* for *Archer*). We also predicted that this pattern would be amplified as the phonetic distance between the anchor talker and the intrusive talker increased, for example when one of these talkers spoke foreign-accented English. To test these hypotheses, we conducted two experiments using different talkers. In the first experiment, all talkers were General American English speakers (GA), and in the second experiment, we introduced L2 Spanish-accented English speakers (SP).

Experiment 1

Methods

Participants Data were collected from 148 participants (Gender: Female, $N = 68$; Male, $N = 77$; Non-binary, $N = 3$; Age: $mean = 37.43$, $sd = 11.76$) who were recruited via Prolific.co and compensated $7.00 for their time (mean duration = 28 m). Participants were pre-screened based on responses to Prolific's questionnaire: They identified as American and listed United States as their current country of residence, listed English as a first language, reported having no language disorders or hearing difficulties, reported either not needing vision correction or using contact lenses, and had an approval rate of 99-100%. Results were removed from analysis for participants with accuracy rates of less than 90% (pre-exclusion mean accuracy rate = 95.16%). This led to the exclusion of 9 participants. Some participants were also directed away from the study after completing the calibration phase and provided with partial compensation. This was the case when either the sample rate was recorded as being below 5 Hz or the validation accuracy was below 65%. The mean validation accuracy on trials included for analysis was 88.19%.

Stimuli Auditory stimuli consisted of spoken sentences where the last word was an imageable noun. There were 100 sentences, including 84 fillers and 16 critical sentences. Filler sentences were selected from the Revised Speech Perception In Noise (R-SPIN) sentence list (Bilger, 1984). Of those, half were semantically predictable (e.g., *His pants were held up by a belt.*) and half were unpredictable (e.g., *I'm glad you heard about the plant.*). Critical sentences were constructed to resemble unpredictable sentences from the R-SPIN list. Crucially, the final noun in critical trials was di- or trisyllabic and contained a separate, semantically unrelated, imageable, monosyllabic word. For example, targets *Archer* and *Bikini* contain competitors *Arch* and *Key*, respectively. The competitor always occurred in the penultimate syllable of the target and always received primary stress.

Stimuli were recorded by 5 talkers, all of whom were speakers of General American English (GA). One female talker was designated the *Anchor* talker. The other four were designated *Intrusive* talkers (2 female, 2 male). The Anchor talker read all 100 stimulus items, and the Intrusive talkers read only the critical sentences. Each talker read through three randomized lists of the sentences. Audio was captured in a sound-attenuated booth through an Electro-Voice RE320 dynamic microphone at a sample rate of 48 kHz. Audio was normalized to the same mean intensity. Critical stimuli were created by marking the onset of the target word in each audio file, removing the anchor talker's production of the target word, and splicing in the intrusive talker's production. For the control trials, where the whole sentence was produced by the Anchor talker, a separate production of the target was spliced in to ensure that effects were the result of the voice swap rather than a simple stream anomaly. All audio manipulation was conducted in Praat (Boersma & Weenink, 2024).

Image stimuli were taken from the MultiPic database (Duñabeitia et al., 2022). Several nouns which were not available in the database were rendered by a local artist who recreated the simple visual style of the originals. Supplemented images were normed to ensure that they conjured the intended word labels and that they were not visually distinguishable from the MultiPic images.

Design & Procedure Each participant completed four experimental blocks, each consisting of 25 trials. Of these 25 trials, four were critical, with two including an intrusive talker and two being produced only by the anchor talker. Each intrusive talker was heard two times throughout the whole experiment. Critical trials were relatively sparse among fillers because the design was contingent on participants being surprised by the intrusive voice. Thus, more critical trials may have led to quicker habituation to the manipulation. Among critical trials, half of the image sets included a competitor (e.g., *Arch* for *Archer*), and half did not. All image sets contained either two or three semantically and phonologically unrelated distractors so that the full set contained four images. The four types of critical trial are displayed in Fig. 1.

Participants completed the experiment remotely via their own web browsers. The experiment procedure was coded in JavaScript using the jsPsych library and the WebGazer eye-tracking plugin (De Leeuw, Gilbert, & Luchterhandt, 2023; Papoutsaki et al., 2016). After providing consent, they proceeded to a calibration and validation stage. Participants who did not pass this stage were directed back to Prolific and compensated for their time. Participants who did pass read instructions for the main experiment and then proceeded to the first block. In each trial, participants saw a visual world with a fixation cross at the center and four images across four quadrants of the screen. Targets (and competitors when present) were placed randomly in one of the four quadrants on each trial. The visual world was displayed for 2500 ms before the sentence began playing. Participants were instructed to look at the image that represented the last word in the sentence and click on it. After a selection was made, there was a 500 ms ITI and then the next set of images appeared automatically.

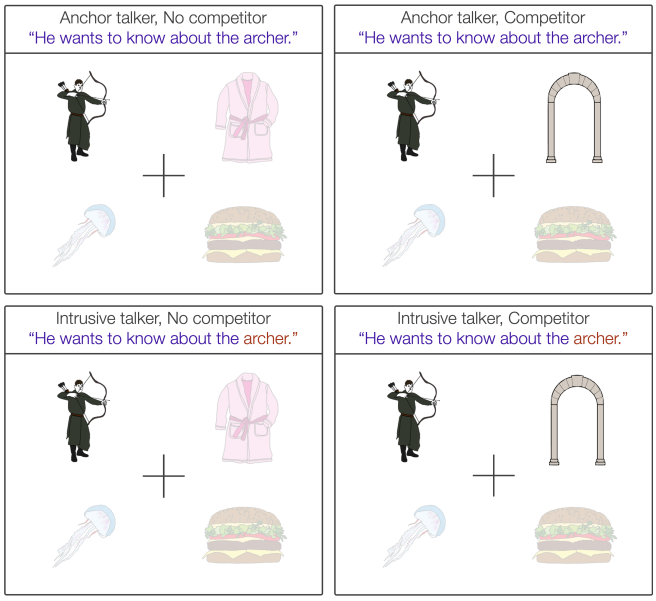

Figure 1: Four different critical trial types, where the voice of the *anchor* talker is represented by blue text and the voice of the *intrusive* talker is represented by red text. In all cases, the *Archer* is the target (top-left quadrant). In trials where a competitor is present, the *Arch* is the competitor (top-right quadrant). Distractors are displayed at a lower opacity here only for demonstrative purposes. Positions of targets and competitors were randomized.

After each block, participants completed a recalibration phase to ensure that the eye-tracker was still measuring gaze accurately. Afterward, participants continued to the next block. After the final block, participants completed a brief demographic questionnaire and then were directed back to Prolific.

Analysis Mixed-effects models of accuracy and latency were fitted in R using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) with p-values attained via Satterthwaite's method using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017). Random-effects structures were determined following Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017). Accuracy was analyzed with a generalized binomial model, where correct responses were coded as 1 and incorrect responses were coded as 0. The model included two sum-coded binary categorical predictors and their interaction: *Talker* (*Anchor* vs. *Intrusive*) and *CompetitorCondition* (*Competitor* vs. *NoCompetitor*).

RTs were analyzed only on correct responses and were removed from analysis if they were more than 2.5 SDs from the mean, leading to data loss of 2.46%. The model of latency was fitted to natural-log-transformed response time measured in ms from the onset of the target noun (logRT). LogRT was analyzed only for correct responses. The model of logRT included the same categorical predictors as the model of accuracy, but also included a control for the duration of the target (rescaled 0-1 and mean-centered).

In the analysis of gaze, a participant was considered to be looking at a given area of interest (AOI) if their gaze was measured as falling within the boundaries of a particular image, and coded as *background* if their gaze did not fall within the boundaries of any image. Looks to AOIs were converted to empirical logits (Barr, 2008) and modeled with a generalized additive mixed model (GAMM) using the *mgcv* and *itsadug* packages in R (Van Rij, Wieling, & Baayen, 2015; Wood, 2000). The model used looks-to-target in empirical logits as the dependent variable and each combination of *Talker* and *CompetitorCondition* as an independent variable (i.e., *AnchorComp*, *AnchorNoComp*, *IntrusiveComp*, and *IntrusiveNoComp*. The model included a smooth term for condition by time in ms with target onset coded as 0, as well as tensor smooths for trial number and target duration. Random intercepts were included for participants and random smooths for participants by time. The model included an autocorrelation parameter, estimated from a simpler model of the data.

In the GAMM modeling framework, the parametric coefficients indicate whether there were more looks to the target overall within a given condition. Our reporting focuses on the smooth terms, which show how these differences evolve over time. A significant smooth indicates that the time-course of gaze reliably departs from a flat trajectory. The most important significance tests come from a direct comparison of smooths for individual conditions. We report here the time windows where binary pairs of smooths differed from one another significantly. Time windows where smooths differed significantly were evaluated directly from the fitted GAMM. These are regions where the 95% confidence band surrounding the difference between two smooths did not include 0. In the difference plots (Fig. 3, right; Fig. 5, right; Fig. 6), these windows are marked with red bands on the x-axis.

Results

Accuracy & Latency Participants were highly accurate in selecting the correct image across conditions, suggesting that they were generally able to parse stimulus sentences correctly even in the presence of an intrusive talker. However, participants were more accurate when the anchor talker produced the whole sentence than when an intrusive talker produced the target noun ($\beta = 0.29, SE = 0.14, z = 2.02, p < 0.05$; Fig. 2, left). Participants were also faster to answer correctly when targets were produced by anchor talkers than when they were produced by intrusive talkers ($\beta = -0.020, SE = 0.0049, t = -4.041, p < 0.001$; Fig. 2, right).

Participants were also more accurate when no competitor was present in the visual world than when a competitor was present ($\beta = 0.30, SE = 0.14, z = 2.01, p < 0.05$). However, correct responses were only marginally faster when no competitor was present than when it was ($p = 0.07$). Taken together, these patterns indicate that performance was substantially inhibited both by the presence of an intrusive talker and by the presence of a sublexical competitor. However, the lack

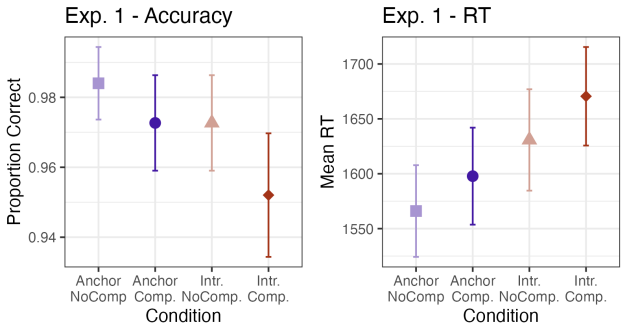

Figure 2: *Left:* Proportion correct responses in Experiment 1 with different conditions on the x-axis (each combination of *Anchor talker/Intrusive talker* and *Competitor/No competitor*). Dots represent means and error bars reflect 95% CIs calculated on raw data. *Right:* Raw response times measured in ms for correct responses. Means and errors were calculated on raw data.

of an interaction between these factors indicates that there is no compounded influence for accuracy or processing speed.

Gaze Output of the GAMM analysis of gaze is shown in Fig. 3. All four smooth terms for condition were significant, including for *AnchorComp* ($edf = 6.59, F = 9.28, p < 0.001$), *AnchorNoComp* ($edf = 7.20, F = 9.07, p < 0.001$), *IntrusiveComp* ($edf = 7.22, F = 11.6, p < 0.001$), *IntrusiveNoComp* ($edf = 6.20, F = 6.11, p < 0.001$). The parametric term was significant only for *IntrusiveComp*, suggesting that there were fewer overall looks to the target than average when the target was produced by the intrusive talker and the competitor was present ($\beta = -0.081, SE = 0.036, t = -2.28, p < 0.05$).

In the direct comparison of trials where a competitor was present, there were more looks to the target when the word was produced by the *anchor* talker than when it was produced by the *intrusive* talker from approximately 500–1050 ms after the target onset (Fig. 3, top). This suggests that the competitor was more influential when there was than when there was not a voice switch and supports the hypothesis. A related finding was observed in the comparison of trials where the target was produced by the *intrusive* talker. Participants were more likely to look at the target from 475–775 ms after the target onset if no competitor was present than if the competitor was present (Fig. 3, middle). Interestingly, there was no equivalent effect in the comparison of gaze on trials where the target was produced by the *anchor* talker with and without competitors. These findings supports our hypothesis in showing that a voice switch was more influential in the presence of a competitor than when no competitor was present.

The comparison of trajectories for *anchor* and *intrusive* trials with no competitor also revealed a notable pattern. From 1350–2300 ms after target onset, there were more looks to the target in the *intrusive* than in the *anchor* condition (Fig. 3,

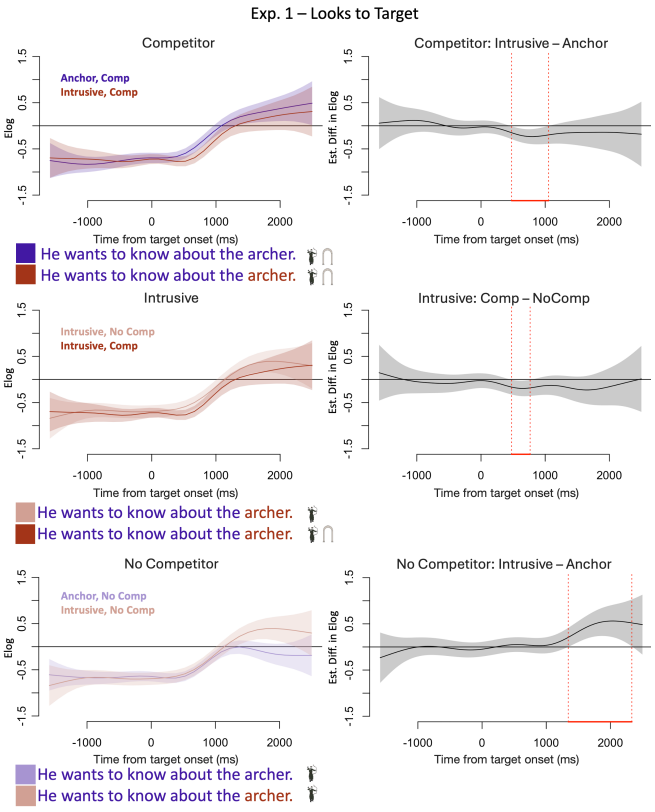

Figure 3: *Left:* For Exp. 1, looks to target measured in empirical logits (Elog), separated by trial type. Dark blue: Anchor talker, competitor present; Dark red: Intrusive talker, competitor present; Light blue: Anchor talker, no competitor present; Light red: Intrusive talker; no competitor present. On the x-axis, 0 reflects the onset of the target word. *Right:* Differences between each pair of curves shown on the left. Regions between dotted lines and marked with red on the x-axis reflect significant divergence between the two trajectories.

bottom). This pattern may reflect a sustained activation of the target after the category search that is induced by the presence of the *intrusive* talker. Meanwhile the target activation decays more quickly in the presence of the *anchor* talker, given that the process of lexical retrieval was relatively unremarkable and less effortful. In other words, this difference may represent listeners retracing the acoustic signal in working memory after an anomalous production.

Experiment 2

In Exp. 2, we asked how the effects observed in Exp. 1 would be influenced when phonological patterns differ more strongly between anchor and intrusive talkers. To test this, we conducted an experiment following largely the same design as Exp. 1 but with the addition of new Spanish-accented L2 talkers (SP).

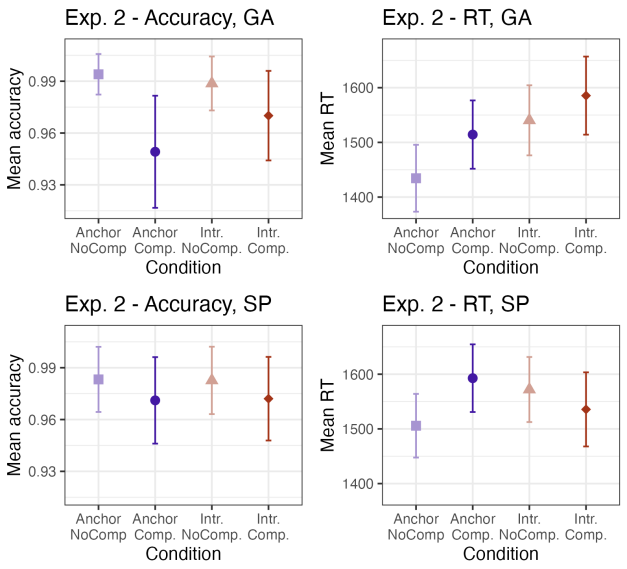

Figure 4: *Top left:* Accuracy on trials in Exp. 2 where a GA taker was the anchor talker. *Top right:* RT on trials where a GA talker was the anchor talker. *Bottom left:* Accuracy on trials in Exp. 2 where a SP taker was the anchor talker. *Bottom right:* RT on trials where a SP talker was the anchor talker.

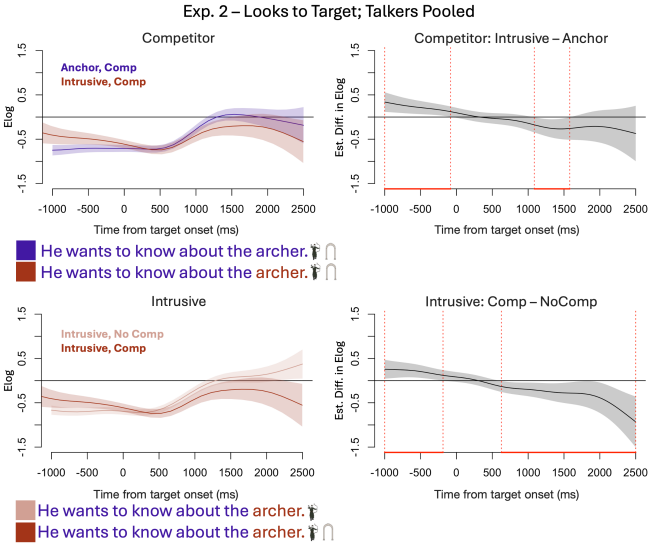

Figure 5: *Left:* For Exp. 2, looks to target measured in empirical logits (Elog), separated by trial type and pooled across accent. *Right:* Differences between each pair of curves shown on the left. Regions between dotted lines and marked with red on the x-axis reflect significant divergence between the two trajectories.

Methods

Participants Data were collected from 92 participants on Prolific using the same pre-screening filters as in Exp. 1 (Gender: Female, $N = 48$; Male, $N = 43$; Non-binary, $N = 1$; Age: *mean* $= 37.42$, *sd* $= 10.87$). Participants were again excluded if their accuracy rates were under 90%, leading to the exclusion of 5 participants.

Stimuli Exp. 2 used the same stimulus sentences as Exp. 1. Three new talkers were recorded. One of these was an additional female GA talker. Two others were female talkers whose native language was Puerto Rican Spanish but who lived in a predominantly English-speaking environment in the mainland United States at the time of recording and spoke fluent Spanish-accented English. These recording sessions followed the same procedure as in Exp. 1 and recordings were processed in the same way. Image stimuli were the same as in Exp. 1.

Design & Procedure The design and procedure of Exp. 2 were identical to Exp. 1 except that each participant was placed in a between-subjects *Accent* condition, reflecting the accent of the Anchor talker. In the GA condition ($N = 43$), one of the two GA talkers was randomly selected as the anchor talker, and the two SP talkers acted as intrusive talkers. In the SP condition ($N = 44$), one of the two SP talkers was randomly selected as the anchor talker, and the two GA talkers acted as the intrusive talkers. The design and procedure were otherwise identical to Exp. 1.

Analysis The analysis of accuracy and latency data were the same as in Exp. 1 except that both models contained an additional sum-coded, binary, categorical predictor: *Accent* (*GA* vs. *SP*), reflecting the accent of the anchor talker. For the analysis of RT, observations lying 2.5 SDs beyond the mean were again removed, leading to data loss of 2.35%. Gaze data was again analyzed using with a GAMM. Model specifications were the same as Exp. 1 with the addition of the *Accent* variable (*GA* vs. *SP*).

Results

Accuracy and Latency In Exp. 2, participants were not more accurate responding to targets produced by the anchor talker than by the intrusive talker ($p > 0.1$), although they were faster to do so ($\beta = -0.016, SE = 0.0057, t = -2.83, p < 0.01$). Participants were more accurate in selecting the target when there was no competitor than when there was a competitor ($\beta = 0.57, SE = 0.22, z = 2.61, p < 0.01$), and were faster as well ($\beta = -0.012, SE = 0.0057, t = -2.10, p < 0.05$). There was no effect of accent for either accuracy or latency.

Gaze Seven of the eight smooth terms for condition were significant, including for *AnchorNoCompGA* ($edf = 5.54, F = 8.53, p < 0.001$), *IntrusiveNoCompGA* ($edf = 5.71, F = 6.83, p < 0.001$), *AnchorCompGA* ($edf = 6.52, F = 3.38, p < 0.001$), *IntrusiveCompGA* ($edf = 6.04, F = 5.34, p < 0.001$), *IntrusiveNoCompSP* ($edf = 6.15, F = 8.50, p < 0.001$), *AnchorCompSP* ($edf = 6.84, F = 8.02, p < 0.001$), *IntrusiveCompSP* ($edf = 5.87, F =$

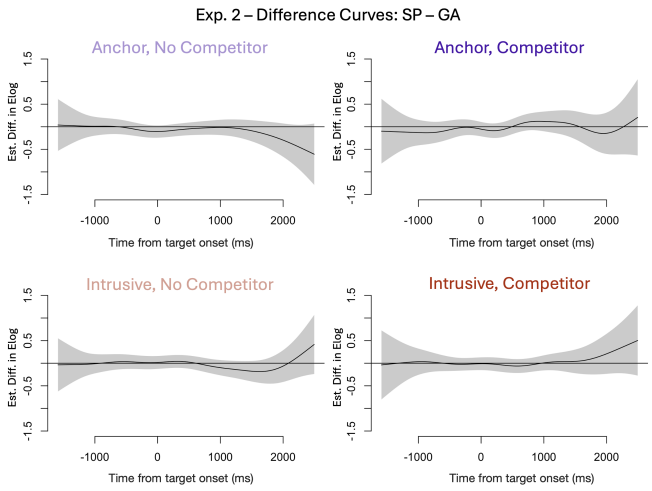

Figure 6: For Exp. 2, difference curves between each *Accent* condition, within each combination of *Talker* and *Competitor* conditions. The *Accent* of the anchor talker did not lead to significant differences in looks to the target in any of these conditions.

$6.93, p < 0.001$). The effect of *AnchorNoCompSP* was marginal ($edf = 4.94, F = 2.00, p = 0.056$).

For brevity, we only discuss several of the most important patterns in the trajectories here. With both *Accent* conditions pooled, there were more looks to the target when a competitor was present on trials produced by the *anchor* talker than on those with an *intrusive* talker from 1050–1550 ms after the target onset (Fig. 5, top). This replicates one of the central findings from Exp. 1. Comparing trials with an *intrusive* talker, with and without a competitor, there were more looks to the target when no competitor was present than when the competitor was present from 600 ms to 2500 ms (Fig. 5, bottom).[2] This again indicates that the voice switch had a larger impact on lexical retrieval in the presence of a competitor.

We also analyzed gaze by *Accent*, but did not find differences based on the accent of the *anchor/intrusive* talker. There were not significant differences between trajectories when the anchor talker was GA or SP (all $p > 0.05$). That indicates that while the central effects held in Exp. 2, participants behaved no differently as a result of the accents of the anchor and intrusive talkers. Comparisons of gaze trajectories across*Accent* conditions are shown in Fig. 6.

Discussion

In this study, we explored the hypotheses that when a listener's top-down expectations were violated via an unexpected voice switch, a category search would be triggered, reducing masking, and leading to increased interference from sublexical patterns in the speech signal. We predicted that this effect would be amplified by the visual presence of the sublexical competitor and that it would be further amplified when anchor and intrusive talkers spoke with different accents.

In Exp. 1, where all talkers spoke GA English, we found that participants were less accurate and slower to respond correctly for intrusive talkers than anchor talkers and less accurate when competitors were present than when they were not. We also found that intrusive talkers and visual competitors significantly altered trajectories of looks to targets, with fewer looks to the target in a critical window beginning approximately 500 ms after target onset, both *1)* for intrusive relative to anchor sentences when a competitor was present, and *2)* when a competitor was present than when it was not for targets produced by an intrusive talker.

The results for Exp. 2, where talkers spoke either GA or SP English, results were somewhat less clear. In our analysis of accuracy and latency, there were no effects of accent, and participants were faster but not more accurate in responding to targets produced by anchor than intrusive talkers. They were both faster and more accurate when no competitor was present in the visual world than when a competitor was present. In the gaze data, the most important patterns from Exp. 1 were replicated, including more looks to the target for anchor than intrusive trials when a competitor was present, and more looks to the target when no competitor was present than when it was on intrusive trials. We did not find significant differences between the accent conditions.

In sum, these results constitute support for our first hypothesis but not our second hypothesis. The most central prediction, that a voice switch would lead to increased interference from sublexical information, was supported in both experiments. However, we also predicted that these effects would be amplified when anchor and intrusive talkers spoke with different accents. While the central effects from Exp. 1 were replicated in Exp. 2, we did not find an influence of the talkers' accents or the direction of the voice switch on the results.

Perhaps most importantly, this study provides proof of concept for a new approach in studying the rapid and dynamic adaptation that allows listeners to understand an astounding range of talkers with little apparent effort. It also provides support for ART as a promising framework for future research, as it makes testable predictions about human cognitive behaviors. The present study had several limitations, mostly due to the large amount of noise inherent in remotely collected eye-tracking data. However, it has also made several avenues for future research clear, including testing similar questions in a more controlled laboratory environment, manipulating the phonological relationship between anchor and intrusive talkers more explicitly, exploring how social dynamics such as gender influence interference, testing how the effect surfaces when the accent but not the talker changes (e.g. by using stimuli from bidialectal talkers), and manipulating structural properties of targets, such as syllable structure and stress patterns.

[2]Note that in both analyses visualized in Fig. 5, some significant effects precede the target onset. This is an artifact of relatively sparse data toward the beginning of the analysis window, which approximates the mean sentence onset time ($-1,062$ ms).

Acknowledgments

We are grateful to members of the Stanford Phonetics Lab and audiences at the 6th California Meeting on Psycholinguistics (CAMP) for helpful comments on this work. Partial funding was provided by the Stanford Vice Provost for Undergraduate Education (VPUE) via an Undergraduate Research Small Grant.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(4), 419–439. doi: 10.1006/jmla.1997.2558

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474. doi: 10.1016/j.jml.2007.09.002

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bilger, R. C. (1984). Speech recognition test development. In E. Elkins (Ed.), *Speech Recognition by the Hearing Impaired* (pp. 2–15). Rockville, MD: American Speech-Language-Hearing Association.

Boersma, P., & Weenink, D. (2024). *Praat: doing phonetics by computer.* Retrieved from `http://www.praat.org/`

De Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jsPsych: Enabling an Open-Source CollaborativeEcosystem of Behavioral Experiments. *Journal of Open Source Software*, *8*(85), 5351. doi: 10.21105/joss.05351

Duñabeitia, J. A., Baciero, A., Antoniou, K., Antoniou, M., Ataman, E., Baus, C., ... Pliatsikas, C. (2022). The Multilingual Picture Database. *Scientific Data*, *9*(1), 431. doi: 10.1038/s41597-022-01552-7

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes*, *12*(5-6), 613–656. doi: 10.1080/016909697386646

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics*, *31*(3-4), 305–320. doi: 10.1016/S0095-4470(03)00030-5

Grossberg, S. (2013). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks*, *37*, 1–47. doi: 10.1016/j.neunet.2012.09.017

Grossberg, S. (2021). *Conscious mind, resonant brain: how each brain makes a mind.* New York, NY: Oxford University Press. (OCLC: on1237861748)

Grossberg, S., Boardman, I., & Cohen, M. (1997). Neural dynamics of variable-rate speech categorization. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(2), 481–503. doi: 10.1037/0096-1523.23.2.481

Kazerounian, S., & Grossberg, S. (2014). Real-time learning of predictive recognition categories that chunk sequences of items stored in working memory. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.01053

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonoic Bulletin & Review*, *13*(2), 262–268.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). doi: 10.18637/jss.v082.i13

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. doi: 10.1037/h0020279

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. doi: 10.1016/j.jml.2017.01.001

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

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable Webcam Eye Tracking Using User Interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 3839–3845).

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*, 104070. doi: 10.1016/j.jml.2019.104070

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of Visual and Linguistic Information in Spoken Language Comprehension. *Science*, *268*(5217), 1632–1634. doi: 10.1126/science.7777863

Van Rij, J., Wieling, M., & Baayen, R. H. (2015). *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs.* doi: 10.32614/CRAN.package.itsadug

Wood, S. (2000). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation.* doi: 10.32614/CRAN.package.mgcv

Xie, X., Jaeger, T. F., & Kurumada, C. (2023). What we do (not) know about the mechanisms underlying adaptive speech perception: A computational framework and review. *Cortex*, *166*, 377–424. doi: 10.1016/j.cortex.2023.05.003