Are you paying attention? Investigating the extent of automatic speech adaptation

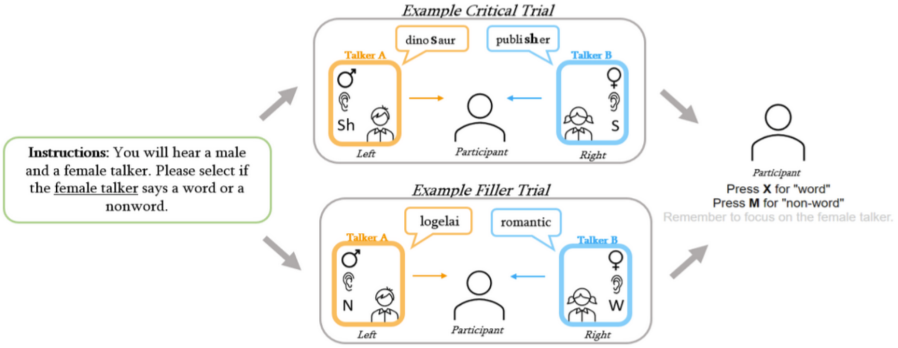
Pronunciation is variable: even talkers with similar language backgrounds realize the same phonological category differently (e.g., the amount of spectral energy use to differentiate /s/ from /ʃ/). Listeners' perception often flexibly accommodates this variation. Our ability to adapt seems to be surprisingly automatic: research on lexically-guided perceptual recalibration (PR) suggests that adaptation is not inhibited by distractions [1], lack of intention [2], or exposure to multiple talkers [3, 4]. Here we explore the extent of this automaticity. We investigated PR to simultaneous speech from two talkers, only one of which listeners attended to—an extreme version of the cocktail party problem. If the malleability of PR is contingent on the allocation of attentional resources, then listeners should adjust to a talker's phonetic pronunciations if and only if the listener is directing their attention towards that talker's verbal stream.

Experiment (n=60 participants). Following the typical structure of PR experiments, we first exposed listeners to speech from unfamiliar talkers (*Figure 2*), and then used a test phase to assess how this exposure affected the listeners' perception of those talkers. However, unlike typical PR experiments, each exposure trial consisted of isolated word recordings from *two different simulated talkers* (one female, one male)*, played at the same time* (one talker to the left ear and one talker to the right ear; talker-to-ear-assignment was balanced across trials). Participants were asked to always attend to the same talker (e.g., always the female talker). We manipulated whether that talker was ʃ-biased (producing /s/ in a way that sounded /ʃ/-like) or S-biased (producing /ʃ/ as /s/-like) between participants. Previous studies have found PR in similar paradigms [3,4] and in paradigms where the two words partially overlap (e.g, if the unattended talker's speech started with speech onset asynchrony (SOA) of 200ms after the attended talker's speech) [5]. In the test phase, we assessed listeners' PR of separate ASI-ASHI continua produced by the two talkers, presented in a random, non-blocked order.

Results. Participants' lexical decision accuracy during exposure was above chance (≥80% accuracy), though lower than in PR experiments where each recording contains speech from only one talker (~85-95% accuracy). Critically, mixed-effects logistic regression of participants' categorization responses during the test found no significant PR ($p>0.05$, *Figure 3)*. This finding complements previous experiments that found PR during longer SOAs, suggesting limits to the automaticity of PR despite the archival of lexical access (as demonstrated by the lexical decision accuracy during exposure). Furthermore, our findings may suggest that available attentional resources mediate our ability to internalize/integrate lower-level speech input, or that lower-level phonetic perception is potentially moderated by higher-level cognitive judgements.

References

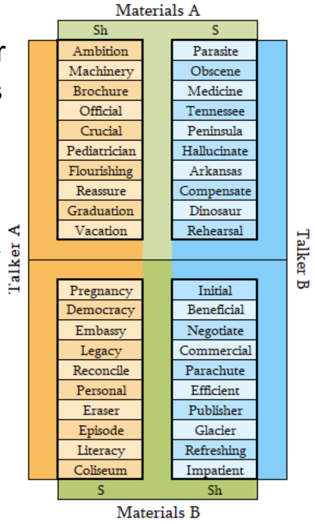
[1] Zhang, X., & Samuel, A. G. (2014). Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance*. https://doi.org/10.1037/a0033182. **[2]** McAuliffe, M., & Babel, M. (2016). Stimulus-directed attention attenuates lexically-guided perceptual learning. *JASA***.** **[3]** Cummings, S. N., & Theodore, R. M. (2022). Perceptual learning of multiple talkers: Determinants, characteristics, and limitations. *Attention, perception & psychophysics*. **[4]** Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. J*ournal of Memory and Language*. https://doi.org/10.3758/s13414-022-02556-6. **[5]** Samuel, A. G. (2016). Lexical representations are malleable for about one second: Evidence for the non-automaticity of perceptual recalibration. *Cognitive Psychology*, 88, 88–114. https://doi.org/10.1016/j.cogpsych.2016.06.007.



Instructions: You will hear a male and a female talker. Please select if the female talker says a word or a nonword.

Press **X** for "word"
Press **M** for "non-word"
Remember to focus on the female talker.

Figure 1: **Procedure during exposure phases**. Participants were instructed to listen to recordings that played single word utterances from two talkers. All exposure trials consisted of a recording from the female talker and a recording from the male talker, played at the same time as a stereo sound. Participants were asked to always attend to either the male (♂) or the female (♀) talker (counterbalanced across participants). Critical trials (top) consisted of one ʃ-word and one S-word (see *Figure 2*). Filler trials (bottom) consisted of a word and a non-word. The participant had to respond whether the *attended* talker produced a word or a nonword.


Talker A
Talker B

Materials A

Sh	S
Ambition	Parasite
Machinery	Obscene
Brochure	Medicine
Official	Tennessee
Crucial	Peninsula
Pediatrician	Hallucinate
Flourishing	Arkansas
Reassure	Compensate
Graduation	Dinosaur
Vacation	Rehearsal

Materials B

S	Sh
Pregnancy	Initial
Democracy	Beneficial
Embassy	Negotiate
Legacy	Commercial
Reconcile	Parachute
Personal	Efficient
Eraser	Publisher
Episode	Glacier
Literacy	Refreshing
Coliseum	Impatient

Figure 2: **Materials.** Critical words spoken by Talkers A and B were grouped into pairs of ʃ-words and S-words. For half of the pairs, Talker A produced the ʃ-word and Talker B produced the S-word. For the other half, this assignment was reversed. Eight variants of these pair-to-talker assignments were created by crossing various nuisance factors: (a) whether Talker A was ʃ-biased and Talker B was S-biased or vice versa (talker-to-bias assignment); (b) whether Talker A was the female talker and Talker B was the male talker or vice versa (talker-to-gender assignment); (c) whether critical words of Talker A were heard in the left ear and critical words by Talker B in the right ear or vice versa (talker-to-ear assignment; filler trials counterbalanced this assignment within participants).

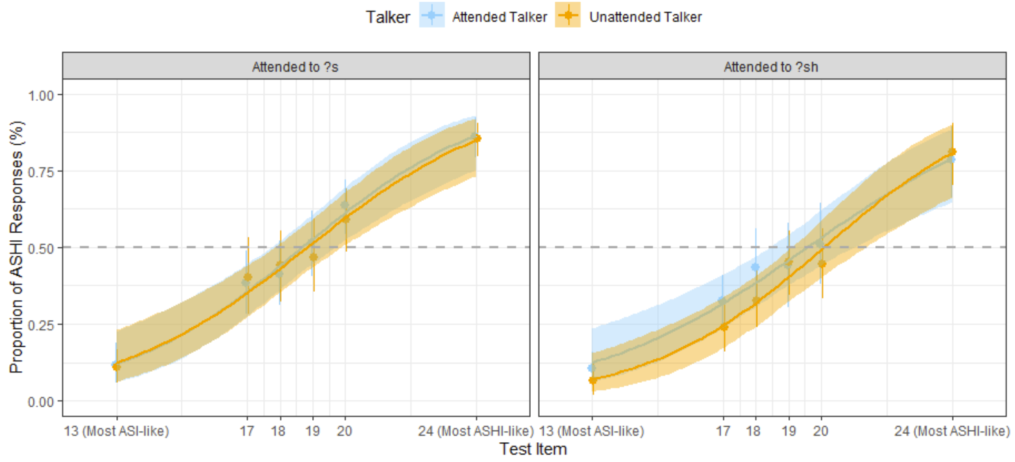


Figure 3: **Results.** The average proportion of ASHI responses during test for the attended talker (blue) and unattended talker (orange), depending on whether the talker produced ʃ-biased words during exposure or S-biased words. Point ranges show proportion of responses (first averaged by participant) with bootstrapped 95% CIs. Lines show best fit of a mixed-effect logistic regression with 95% CIs.