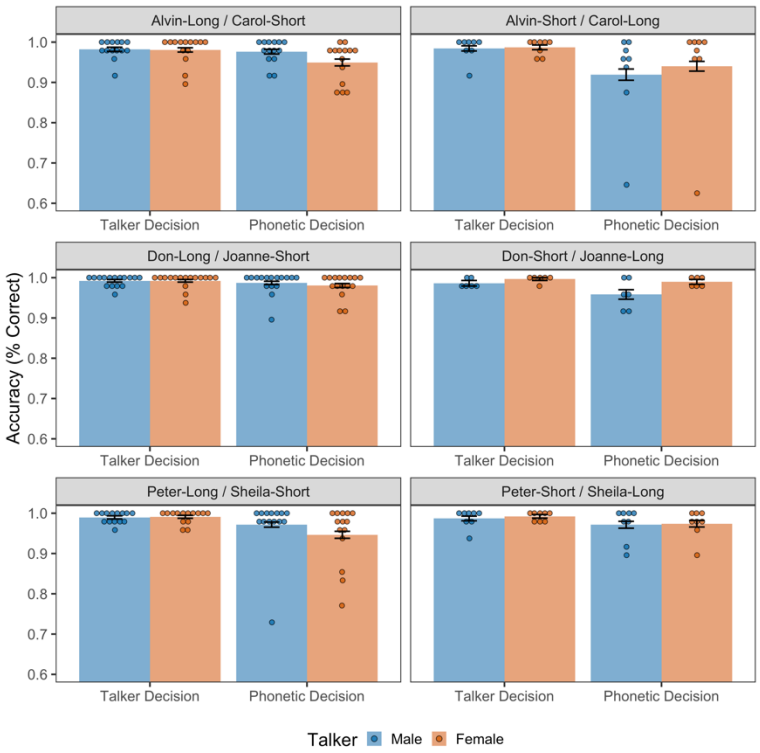
Supplementary Materials

As reported in the main text, Experiment 1 had a larger age range (20-64) compared to Experiment 2 (19-35). To ensure that this larger age range did not drive the results of Experiment 1, we report here the results of an analysis of the Experiment 1 data that only included participants aged 20-35 (N = 22). In brief, we observed the same pattern of results as in the main experiment.

Performance on the training task is visualized in Figure S1A. In analyzing the training data, we separately assessed listeners' ability to identify the talker (regardless of whether they were correct in identifying which word was said) as well as their ability to determine which word was said (regardless of whether they were correct in identifying the talker). Listeners were highly accurate in the talker decision and the phonetic decision, regardless of the talker or their typical VOT (mean accuracy >91% in all cases).

Because a short-VOT variant might be more easily confused with a voiced stimulus (compared to a long-VOT variant), we statistically assessed the influence of the talker's typical VOT (i.e., whether the talker produced voiceless stops with a short or long VOT) on the training task; separate models were conducted for talker identification performance and phonetic identification performance. The statistical modeling approach was the same as in the main text. The talker's typical VOT did not influence performance on the talker identification component of the task, $\chi^2(1) = 0.66$, $p = 0.415$, but did have a significant effect on phonetic identification, $\chi^2(1) = 16.90$, $p < 0.001$; this latter effect was driven by slightly less accurate responses when the talker had a short VOT (mean: 0.96, SD: 0.20) compared to when the talker had a long VOT (mean: 0.97, SD: 0.16). That is, short-VOT variants were more likely to be confused with voiced tokens, but long-VOT variants were mislabeled relatively less often.

A



B

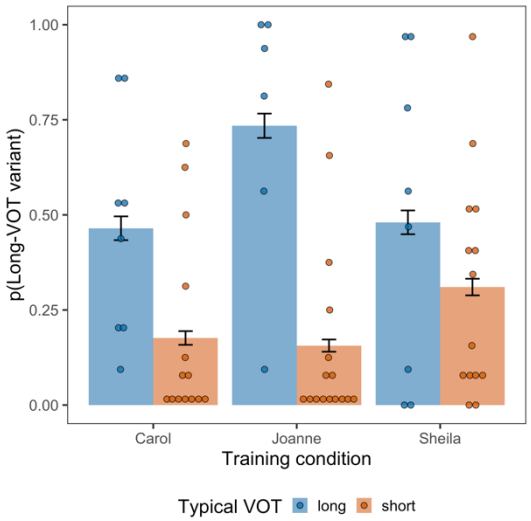


Figure S1. Experiment 1 results, including only participants aged 35 and younger. (A) Performance on the training task, separately considering whether listeners were accurate in identifying who was talking ("Talker Decision") and which word they said ("Phonetic Decision"). Accuracy values are shown on the y-axis. Each row shows performance on a different block. In plots on the left, the female talker produced voiceless stop consonants with a short VOT, and in plots on the right, she produced these consonants with a long VOT. Dots represent individual subject data. Error bars indicate standard error of the mean. (B) Results from the test phase of Experiment 1, showing the probability that a listener selected the long-VOT variant (y-axis) as a function of the talker (x-axis) and whether the talker produced voiceless stops with long (blue bars) or short (orange bars) VOTs during training. Dots represent individual subject data. Error bars indicate standard error of the mean.

Mean overall accuracy in the test phase was 70.7% (SD: 15.8%), and results from the test phase are plotted in Figure S1B. Visually, it is clear that participants were more likely to select the long-VOT variant as the more typical one when the talker had previously produced long-VOT variants during training. To evaluate this statistically, test data were submitted to a linear mixed effects regression that assessed how fixed factors of Talker (Carol, Joanne, Sheila; sum-coded) and Typical VOT (long/short; sum-coded) influenced whether participants selected the long-VOT variant. To select our random effect structure, we began with the maximal random effect structure that converged (Barr et al., 2013) and used a backward-stepping procedure to identify whether we could use a simpler model structure without significantly compromising model fit (Matuschek et al., 2017). In this way, we selected a random effects structure with random by-subject slopes for Typical VOT as well as random by-subject intercepts. We observed a significant effect of Typical VOT, $\chi^2(1) = 12.71$, $p < 0.001$, driven by more long-VOT responses if the talker's characteristic VOT was long (mean: 0.54, SD: 0.50) than if it was short (mean: 0.21, SD: 0.41). No other effects were significant ($p > 0.19$).