Dynamic listening: temporal expectations guide perception of phonetic detail

Jason A. Shaw (Western Sydney U. & Keio U.)
J.Shaw@westernsydney.edu.au

Introduction: Processing advantages found for familiar talkers have been sometimes interpreted as support for representations incorporating both phonological and indexical information in the lexicon (Goldinger, 1996), although the specific phonetic parameters that contribute to familiarity advantages are poorly understood (Nygaard, 2005 for overview). In this study, we investigate how phonetic parameters relating to time, i.e., gesture duration and gesture coordination, influence talker discrimination, a task requiring perception of talker-specific phonetic detail.

Experiment: Listeners, 36 native English speakers, heard tokens of the word raw or draw spoken in isolation by one of 10 talkers. On each trial, listeners first heard between one and five tokens from the same talker before hearing the test token, which was either from the same talker or a different talker. After the test token, the screen changed colour and listeners were asked to decide whether the last token they heard was produced by the same talker or by a new talker. Listeners heard a total of 240 trials consisting of 24 different talker pairings (described below) X five different levels of exposure to the first talker (one to five tokens) X two conditions (same talker vs. talker change).

Talker selection: The 10 talkers included in the experiment were selected based on two factors: (1) temporal typicality, the main parameter of interest, and (2) overall perceptual similarity, a control variable. To quantify temporal typicality, we fit stochastic time models of syllable structure, following Shaw and Gafos (2015), to data from each talker. These models use gesture durations measured from the data (which we estimated from acoustic segment durations) and language-specific coordination relations to predict the stability of syllable-referential temporal intervals, i.e., the left edge to anchor, center to anchor and right edge to anchor intervals of, e.g., Browman and Goldstein (1988), Hermes et al., 2013, and Shaw et al, 2009. The models were fit to talker data in two different ways. In the first model, gesture durations were specified according to the population distribution, estimated from 97 speakers English. In the second model, gesture durations from the population were replaced with talker-specific distributions. For some talkers, using talker-specific segment durations improved model fit. We refer to these as temporally distinct talkers. For other talkers, talker-specific parameters provided no improvement over the population distribution. We refer to these as temporally typical talkers. The experiment included five temporally typical talkers and five temporally distinct talkers. These 10 talkers were combined to form 24 talker pairs, consisting of an equal balance of typical and distinct talkers as exposure and test talkers. To control for overall perceptual similarity between tokens, we obtained similarity judgements for each combination of raw and draw included in the experiment. Judgements were made on a 7-point likert scale by 31 speakers of English (none of which were participants in the main experiment). The mean similarity ratings for exposure and test tokens was 4.43 (1.22) for temporally distinct test talkers and 4.42(1.44) for temporally typical test talkers.

Hypotheses: If listeners rapidly construct talker-specific representations, then increased exposure to the first talker (from one to five tokens) should aid talker change detection. If gesture duration is an indexical phonetic parameter, then it should be easier to detect a change when the new talker is temporally distinct. Alternatively, segment duration may not be an indexical parameter but temporal typicality may facilitate processing of other phonetic details relevant to discriminating talkers. On this hypothesis, it may be harder to detect a change in talker if the new talker is temporally distinct.

Results and Discussion: Results were analysed in terms of d prime. Figure 1 shows the effects of exposure and temporal typicality of exposure talker (Figure 1a) and test talker (Figure 1b). The effect of exposure was significant [$F(140,4) = 30.96, p <.001$], indicating that the ability to detect a change
in talker increases with exposure. Even a few additional tokens help. The temporal typicality of the test talker was not significant. This also indicates rapid adaptation to temporal variability. The temporal typicality of the test talker, on the other hand, was significant \([F(35,1) = 44.25, p < .001]\). It was harder to detect a talker change when the new talker was temporally distinct. This negative effect of temporal distinctiveness suggests that gesture duration is not an indexical parameter. Rather, timing that deviates from the population norm seems to disrupt listener ability to process phonetic detail relevant to the task. A significant interaction between temporal typicality of exposure and test talkers \([F(35,1) = 13.52, p < .01]\) further supports this interpretation. As shown in Figure 2, temporally distinct talkers are even harder to recognize following exposure of a temporally typical talker. On the basis of these results, we conclude that listeners dynamically adjust their temporal expectations for phonetic events and that perception of phonetic detail hinges in part on these temporal expectations.

**Figure 1a:** effect of exposure and temporal typicality of exposure talker

**Figure 1b:** effect of exposure and temporal typicality of test talker

**Figure 2:** interaction between temporal typicality of exposure and test talkers

**References**


