Investigating the nature of speech representations

Maryann Tan & T Florian Jaeger

Stockholm University

University of Rochester

Research questions

Adaptivity in response to talkers with unexpected pronunciations is central to robust speech perception. Yet, much remains unknown about:

1. **The expectations that listeners hold in the earliest moments of a new talker encounter.** (this poster)
2. How these expectations change as more information about the talker is revealed while perceiving the input. (see poster *1pSC22*)

Exp 1 (N = 24) & Exp 2 (N =122)

1. Listeners categorise minimal pair continua (dill-till, din-tin, dip-tip, dim-tim).
2. VOT items are uniformly distributed.
3. Perception data is compared against the predictions from production data under different theoretical assumptions.

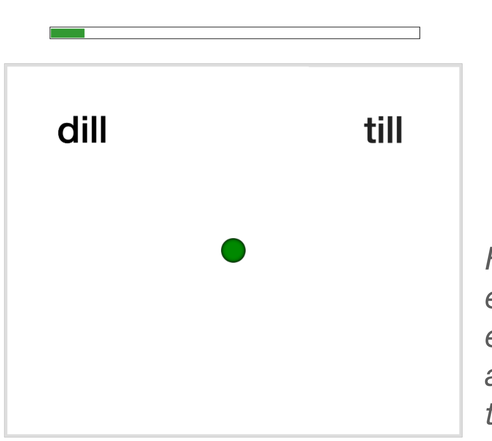

Figure 1: Procedure used in both experiments. Building on Clayards et al. (2008) A recording is played and participants click on the word they heard.

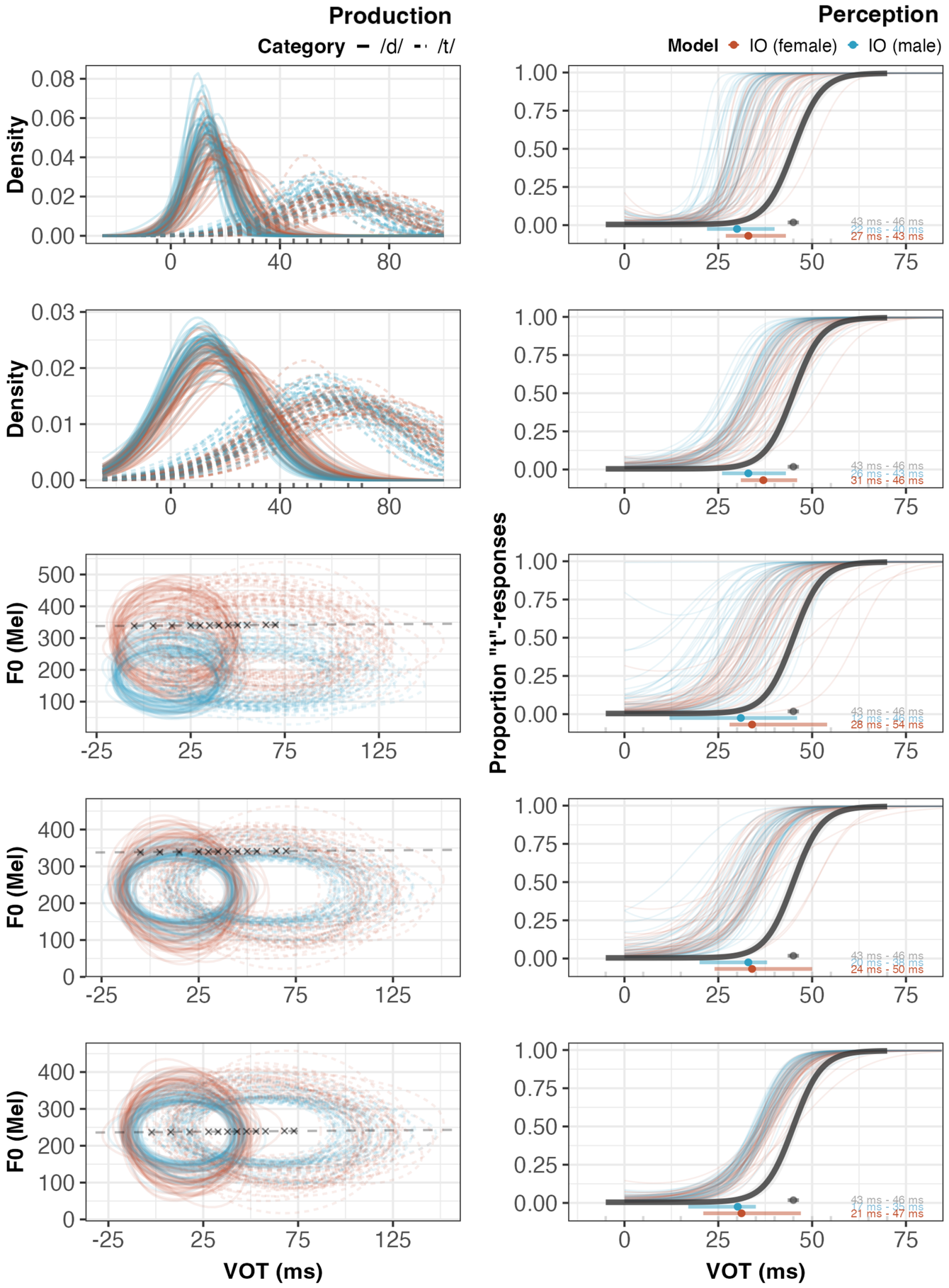

*Figure 2: **Left**: Distributions of phonetically annotated VOT and F0 cues of /d/-/t/ productions from 92 talkers of L1-US English (data from Chodroff & Wilson, 2018). **Right:** Fitted proportion of human"t"-responses (black line) against the predictions of 92 talker ideal observers (IOs), trained on the production data under five different assumptions about phonetic representations. **Right, row 1:** Raw VOT; **row 2:** VOT with perceptual noise; **row 3:** VOT-F0 with perceptual noise; **row 4:** talker-centred cues; **row 5:** VOT-F0 talker-centred and exposure-centred cues.*

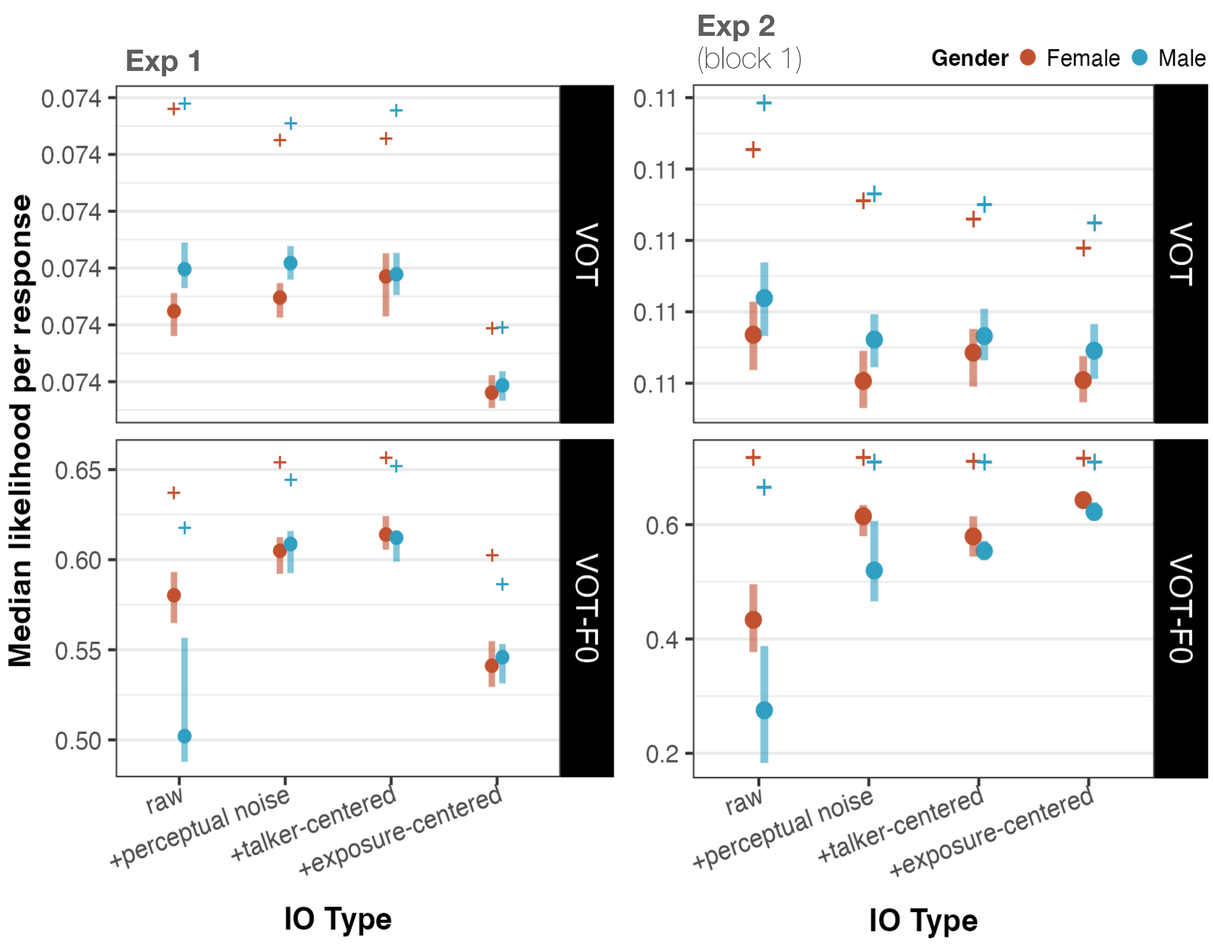

Figure 3: Comparing the fit of each IO type to human responses in Exp 1(left) and Exp 2 (right). "+" indicate likelihood per response under the best-fitting talker-specific model from 1000 bootstrapped samples. Point intervals show median likelihood across all 92 talker-specific models, and the 95% bootstrapped CI.

Take-home points

1. ***Prior to informative exposure*** to an unfamiliar talker, listeners draw on previously experienced speech input that i) integrates multiple cues, ii) is perturbed by perceptual noise, and iii) normalised by talker.
2. Poor fit of exposure-centred model could be due to i) different registers of exposure stimuli vis-a-vis production data (e.g. hyper-articulation); ii) mismatches in vowel context between stimuli and database iii) lack of normalisation of speech rate in the analysis iv) failure to model normalisation as an *incremental inference process*.