Distributional learning as a driver of robust speech processing

Xin Xie, Chigusa Kurumada (U. of Rochester) & Andrés Buxó-Lugo (U. of Maryland)

Many influential theories of language processing assume that listeners **learn and store previously experienced distributional statistics of the input** (Dell & Chang, 2014; Frank & Goodman, 2012; Futrell et al., 2020; Johnson, 2006; Levy, 2008; MacDonald, 2013; Maye et al., 2008; Pierrehumbert, 2001; Tanenhaus & Trueswell, 1995). This knowledge is considered critical for guiding listeners' expectations to achieve efficient language processing. Further, recent work suggests that learning distributions *specific to a talker* can be a key to accommodating the cross-talker variability ubiquitous in spoken language (Kleinschmidt & Jaeger, 2015; Theodore & Monto, 2019). However, approximations of the relevant long-term or talker-specific experiences of distributions often remain unattainable or unreliable because large-scale data of sufficient resolution (e.g., estimates of means and variances of cues to a particular linguistic category) are lacking. So far, most evidence that is taken to support distributional learning as a mechanism underlying speech processing has been based on a short-/mid-term exposure to researcher-curated distributional statistics (Clayards et al., 2008; but see McMurray & Jongman, 2011).

The current study addresses this critical gap in the domain of speech prosody. We, for the first time, combine production, modeling, and comprehension experiments to examine **whether listeners indeed store distributional statistics in productions and draw on them in comprehension.** We built a corpus of 65 talkers, each producing 24 questions vs. 24 statements in the form of "*It' X-ing*" (e.g., "It's changing?" vs, "It's changing") resulting in a total of 2974 tokens (after excluding speech errors). Recorded utterances were segmented into three sections 1) *it's,* 2) *X* (the stressed syllable), and 3)*-ing*. F0 and duration of each syllable were extracted (**Fig.1A, B**) and examined with respect to the structure of variability in the cue distributions (**Fig.1C**).

Experiment 1) Do long-term statistics predict listeners' categorization of a novel talker's speech? We trained two 65 classifiers (multivariate ideal-observers, extending Kleinschmidt, 2019), one for each talker, based on means and variances of the question vs. statement categories directly estimated from the corpus (**Fig.1D**). We then bundled these talker-specific models by the talkers' gender to create two "gender-specific" models, each simulating a prototypical female and a prototypical male talker. Additionally, we created a model without the knowledge of talker gender (the "gender-independent" model). We tested these models against human judgments (N = 240) on categorization of items from a 11-step continuum constructed based on recordings of two new talkers (1 male and 1 female). The *gender-specific* models significantly outperformed the gender-independent one (**Fig.2**), suggesting that **the long-term statistics estimated from male vs. female talkers' productions directly predict listeners' categorization of the prosodic input** (R^2 = .95). Listeners *do* seem to store gender-specific distributions and apply the knowledge in comprehension when first encountering a *novel* talker of a particular gender.

Experiment 2) Do listeners accommodate unexpected distributional statistics from a novel talker? The same human listeners from Experiment 1 were randomly assigned to three conditions: Q(uestion)-biasing, No-bias, S(tatement)-biasing. Those in the Q-biasing condition heard prototypical statements (step 1) and the ambiguous item (step 7 for the female and step 8 for the male talker) disambiguated as questions via feedback. Those in the S-biasing condition instead heard the prototypical Questions (step 11) and the ambiguous items as statements. In the No-bias condition, listeners received only prototypical questions and statements. Results show that the **listeners incrementally adjusted their responses to the ambiguous items throughout the 30 trials** (**Fig.3**, green lines), rapidly learning the underlying, talker-specific, distributions.

In sum, the current study is among the first to empirically demonstrate that speech processing does indeed leverage the implicit knowledge derived from long- and short-term learning of distributional statistics. Listeners process the variable linguistic input by applying distinct sets of expectations derived through prior experiences, which continue to be fine-tuned in response to new exposure.

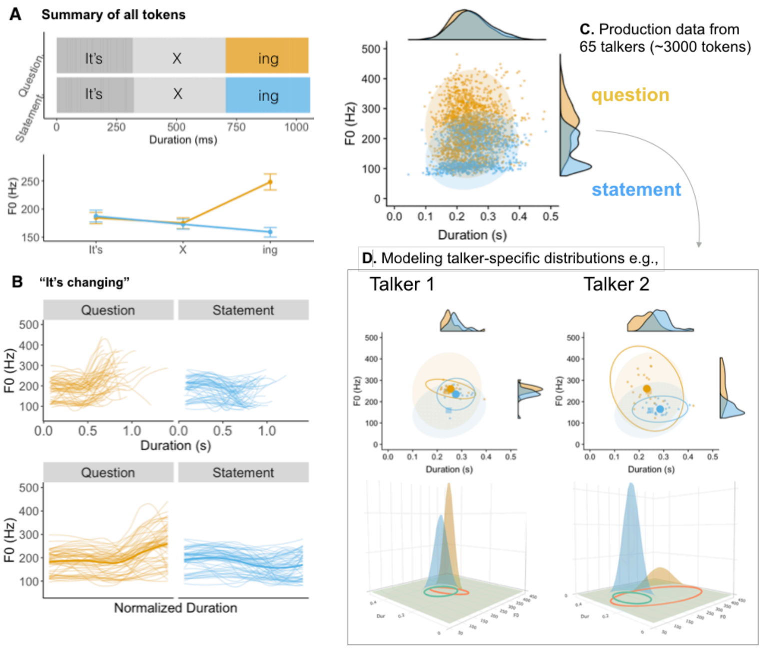

Figure 1.

A. Summary statistics of duration (top) and fundamental frequency (F0, bottom) in the intonation contours for "It's X-ing" utterences produced by 65 native English speakers.

B. F0 values of individual tokens of "It's changing" to illlustrate the magnitude of talker variability seen for each of the 24 item types.

C. Group-level variations of syllable mean F0 (y-axis) and duration (x-axis) in the ~3000 tokens collected;

D. Talker-specific ideal observer models of productions for two example talkers (Talker 1 and Talker 2).

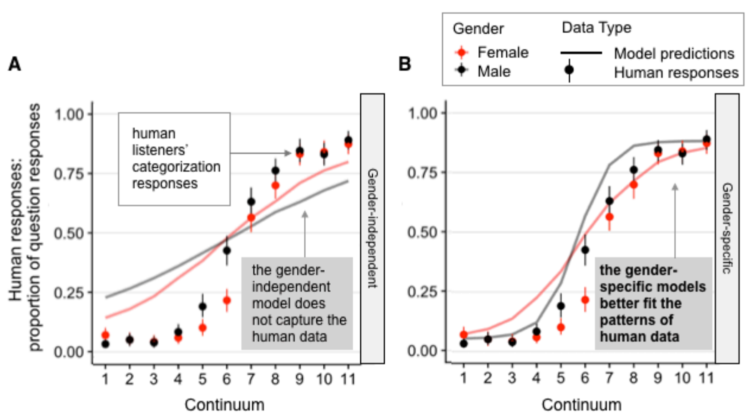

Figure 2.

Categorization functions predicted by ideal observers (lines) and actual categorization by human listeners (pointranges). (The points indicate the by-item means averaged across listeners. Error bars indicate bootstrapped 95% confidence intervals. The human data plotted are identical between the two panels.) **A** : gender-independent model, wherein the two lines represent predictions of one model for the female vs. male talker data. **B**: gender-specific models.

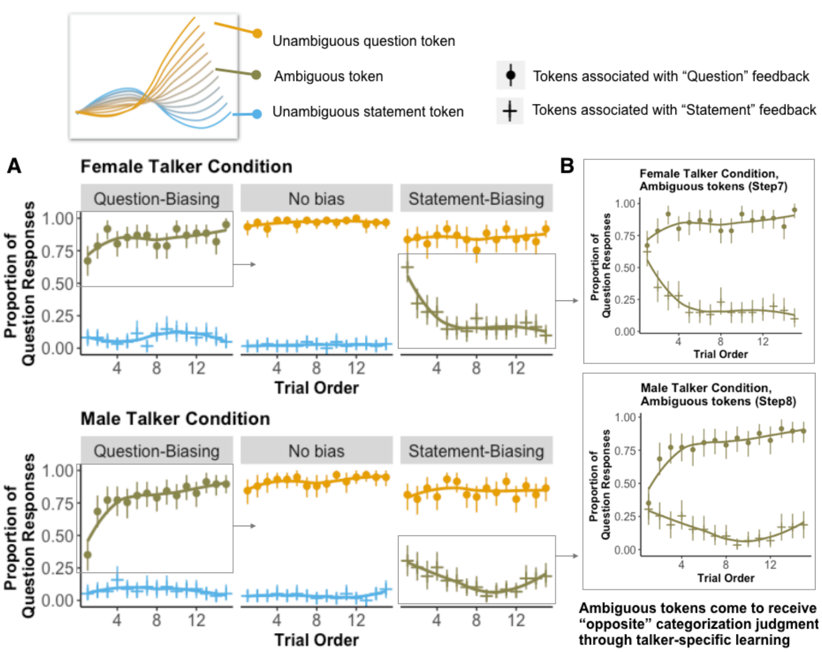

Figure 3.

A. Overall response patterns across the between-subject conditions. X-axis: The relative ordering of the 15 exposure tokens associated with the question vs. the statement feedback. Blue and yellow indicate unambiguous tokens (Step 1 and Step 11, respectively) and green represents the ambiguous items. Error bars indicate bootstrapped 95% confidence intervals.

B. Responses given to the prosodically ambiguous tokens in the female vs. the male talker conditions; the top line (circles) and the bottom line (crosses) represent the Question-Biasing and the Statement-Biasing conditions.