Cognition 211 (2021) 104619

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Encoding and decoding of meaning through structured variability in intonational speech prosody

Xin Xie[a,1,*], Andrés Buxó-Lugo[b,1,*], Chigusa Kurumada[a]

[a] Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA
[b] Department of Psychology, University of Maryland, College Park, MD 20742, USA

ARTICLE INFO

Keywords:
Prosody
Meaning
Intonation
Adaptation
Language production
Language comprehension
Variability

ABSTRACT

Speech prosody plays an important role in communication of meaning. The cognitive and computational mechanisms supporting this communication remain to be understood, however. Prosodic cues vary across talkers and speaking conditions, creating ambiguity in the sound-to-meaning mapping. We hypothesize that listeners ameliorate this ambiguity in part by learning talker-specific statistics of prosodic cues. To test this hypothesis, we investigate the production and recognition of question vs. statement prosody in American English. Experiment 1 elicits productions of questions and statements from 65 talkers to examine the distributional statistics characterizing within- and cross-talker variability in these productions. We use Bayesian ideal observer models to assess the predicted consequences of cross-talker variability on listeners' recognition of prosody. We find that learning of talker-specific distributional statistics is predicted to facilitate recognition, above and beyond what can be achieved via commonly assumed normalizations of prosodic cues. Experiment 2 tests this prediction in a comprehension experiment. We expose different groups of listeners to different prosodic input statistics and assess listeners' recognition of questions and statements both prior to, and following, exposure. Prior to exposure, ideal observer-derived predictions based on Experiment 1 provide a good qualitative fit against listeners' recognition of prosodic contours in Experiment 2. Following exposure, listeners shift the categorization boundary between questions and statements in ways consistent with learning of talker-specific statistics.

1. Introduction

Prosody—the rhythm and cadence of speech—plays a critical role in the communication of meaning. Subtle differences in utterance-final intonation contours, for instance, change an utterance's meaning from a statement (e.g., *It's raining.* [falling intonation]) to a question (e.g., *It's raining?* [rising intonation]). There is a rich evidence base indicating that listeners recognize such meaning-distinguishing prosodic categories (Bolinger, 1989; Gussenhoven, 2002; (Ladd, D Robert, 2008); Pierrehumbert and Hirschberg, 1990) and integrate the meaning as an utterance unfolds (Cutler, 2015; Dahan, 2015; Ito and Speer, 2008; Weber et al., 2006). However, the cognitive and perceptual mechanisms supporting this recognition remain poorly understood.

One major source of difficulty stems from variability in the prosodic signal across talkers and contexts (Arvaniti, 2019; Brugos et al., 2006; Cangemi et al., 2015; Cangemi and Grice, 2016; Cole, 2015). Continuing on the case of statements vs. questions in American English, the exact form and level of the rise produced to signal a question meaning can vary across talkers as well as talker groups (e.g., age, gender, dialect) (Arvaniti and Garding, 2007; Clopper and Smiljanic, 2011). For example, due to difficulties in controlling their pitch, young children tend to produce a smaller degree of a rise than older children (Patel and Grigos, 2006). Also, rising intonation can be used to signal other, including social, meanings (e.g., 'uptalk', Warren, 2016). As a result of this talker variability, one person's production of a statement and another person's production of a question can be phonetically identical.

The present study explores how listeners may navigate this "lack of invariance" in the realization of prosody. Although talker variability in speech acoustics has been an issue central to speech perception research (e.g., Hillenbrand et al., 1995; Newman et al., 2001; Theodore et al., 2009), relevant accounts for how listeners may cope with the variability focus almost exclusively on segmental (as opposed to prosodic) speech

[*] Corresponding authors at: Department of Brain and Cognitive Sciences, Meliora Hall, University of Rochester, Rochester, NY 14627, and Department of Psychology, University of Maryland, Biology/Psychology Building, 4094 Campus Dr., College Park, MD 20742, United States.
E-mail addresses: xxie13@ur.rochester.edu (X. Xie), buxolugo@umd.edu (A. Buxó-Lugo).
[1] The first two authors contributed equally.

https://doi.org/10.1016/j.cognition.2021.104619
Received 1 August 2020; Received in revised form 25 November 2020; Accepted 27 January 2021

perception. Here we begin to test a hypothesis that listeners track and learn structured phonetic variation to optimize the mapping to prosodic categories *on a talker-contingent basis*. As a first step, we examine the structure in talker variability of declarative questions vs. statements (e.g., *It's raining?* vs. *It's raining.*). Below, we briefly review relevant background and outline three concrete steps we take in the present study.

1.1. *What is known about how comprehension copes with cross-talker variability?*

Listeners appear to cope with variability in the prosodic input in more than one way (Lehet and Holt, 2020). One mechanism is called *normalization* or *compensation*. Both F0 and changes in F0 (e.g., rising and falling) are known to be perceived relative to a given talker's F0 baseline (Bishop and Keating, 2012; Hirschberg et al., 2004; Mahrt et al., 2012; Tang et al., 2017). Similarly, rhythmical and durational information can be evaluated relative to the local speech rate (Baese-Berk et al., 2014; Diehl et al., 1980; Reinisch and Maximilian, 2015; Sawusch and Newman, 2000): the same physical duration can be perceived as "long" or "short" depending on the local speech rate. There is now considerable evidence that such normalization can occur within 20 to 50 ms of speech (Lee, 2009) and operate continuously as an utterance unfolds (also known as "distal prosody", Brown et al., 2011; Dilley and Pitt, 2010; Morrill et al., 2015).

Normalization is generally taken to be a low-level auditory process—a type of automatic signal transformation (Adank et al., 2015; Flynn and Foulkes, 2011; Lobanov, 1971; Monahan and Idsardi, 2010). Recognition of a sound category is thought to involve normalization either based on previous observations of the same category (*extrinsic normalization*) or based on other concurrently perceived features (*intrinsic normalization*, e.g., between the first and second formant based on pitch in vowel perception, Adank et al., 2004). Often implicit in this view is the assumption of an underlying, invariant, mapping between phonetic cues and abstract linguistic categories (e.g., phonemes). Normalization *strips away* and discards surface variability, enabling listeners to arrive at categories universal across talkers and contexts (for review and critique, see Weatherholtz & Jaeger, 2016).[2]

An alternative view—the one we seek to explore here—suggests that the human comprehension system *learns* and *stores* variability in the input (Kurumada et al., 2017; Roettger and Franke, 2019; Roettger and Rimland, 2020). Prominent theoretical frameworks that incorporate this idea include Bayesian (Clayards et al., 2008; Norris and McQueen, 2008), episodic (Goldinger, 1996a), or exemplar theories (e.g., Johnson, 1990; Nygaard et al., 1994; Pierrehumbert, 2001; for prosody: Schweitzer, 2012; Smith and Hawkins, 2012). While different in important details, these frameworks identify the significance of retaining idiosyncratic differences across talkers instead of discarding them. Listeners are expected to retain, rather than discard, knowledge about talker- or group-specific cue-category mappings and draw on this knowledge during recognition.

For instance, Warren (2017) showed that the interpretation of utterance-final rising pitch in New Zealand English (NZE) can depend on the (inferred) age of the talker. Both younger and older speakers of NZE use utterance final rising pitch for questions, but some younger speakers also use a variant of utterance-final rising pitch in statements ("uptalk", Warren, 2016). Under the type of account we investigate here, we expect NZE listeners to be less likely to interpret an utterance-final rising pitch as a question for young talkers, compared to old talkers. This is what Warren (2017) found. Critically, the local prosody (e.g., the mean pitch preceding the final rise) was held constant in the experiment, so that Warren's findings cannot be accounted for in terms of talker-independent normalization. The proposal that listeners derive expectations for talker-specific pronunciations further predicts that increasing exposure to a previously unfamiliar talker should facilitate faster and more accurate distinction of question vs. statement, in line with existing evidence (Saindon et al., 2017).

Results like these point to an intriguing possibility that listeners learn and store previously experienced input in a way that helps to overcome the cross-talker variability. How this is achieved, however, remains an open question. The proposal that we explore here holds that listeners learn to build talker-specific expectations about phonetic distributions (Chodroff and Wilson, 2017; Kleinschmidt and Jaeger, 2015). Listeners may then deploy or "swap out" these expectations whenever the talker swiches. Tests of this view have so far been limited to segmental speech perception—specifically, changes in how listeners categorize, say /b/ vs. /p/, following exposure to an unfamiliar talker (Clayards et al., 2008; Kleinschmidt and Jaeger, 2016; Theodore and Monto, 2019). These studies have found that distributional learning—changes in listeners' implicit beliefs about the mean and variance of speech categories—provides a good qualitative and quantitative model of human perceptual judgments. The goal of the present work is to provide an initial test of this hypothesis in the domain of prosody. To this end, we address three important gaps in the literature.

1.2. *Present study: talker-related variability in declarative question vs. statement prosody*

The first gap pertains to a basic question: (1) *What does the distributional structure of phonetic cues to prosody look like*? Although recent studies have begun to describe "talker-specific" realizations of prosodic categories (Brugos et al., 2006; Cangemi et al., 2015; Chodroff and Cole, 2019b), quantification of how and how much talkers vary in their productions has still been limited. In particular, a better understanding of the 'typical' distribution of phonetic cues to prosody is critical to tests of the distributional learning hypothesis. Influential accounts of speech perception share the assumption that listeners recognize new input based on previously experienced input (Goldinger, 1996b; Hay et al., 2006; Johnson, 2005a; Kraljic and Samuel, 2011; Norris et al., 2015; Pierrehumbert, 2003, *inter alios*). One way to conceptualize this is that listeners' expectations for input from a "typical" unfamiliar talker are incrementally updated based on subsequent input from that talker (Kleinschmidt and Jaeger, 2015, 2016). In short, without knowing more about a structure in phonetic distributions across talkers, it is impossible to understand the changes listeners might exhibit when they are exposed to an unfamiliar talker.

The second gap we seek to address concerns the question: (2) *Is talker-specific distributional learning expected to improve recognition above and beyond what can be achieved via normalizations?* Put differently, we ask if talker-specific learning is even expected to be *critical* for prosodic processing. There certainly are situations in which normalization alone cannot resolve cross-talker variability in the cue-category mapping. In extreme cases, listeners experience a full-scale category mismatch (e.g., yes-no questions in some British dialects regularly have a falling intonation, unlike in most variants of American English, Crystal, 1969; Grabe & Post, 2002; Grabe, 2002). This variation, like many other accent features, must be learned and stored. Yet, it is possible that such categorical differences are restricted to highly idiosyncratic cases, and that most variability encountered within a variant of English can be resolved through normalization of phonetic cues against its surrounding

[2] The term normalization is also sometimes used in a broader sense, referring to signal transformations that do not solely rely on the local signal (e.g., Johnson, 2005b). Those instances of 'normalization' might require storage, and thus learning, of the relevant information from previous observation. For example, pitch normalization relative to the talkers' age or gender might fall into this category (Bishop and Keating, 2012), as do any approaches that interpret cues relative to talker-specific expectations (e.g., C-CuRE, McMurray and Jongman, 2011). For any of these accounts, the question arises of how those "relative expectations" come to be stored in the first place. These accounts thus raise the same question we explore here.

context (e.g., baseline vocal pitch height, speech rate, vocal range). In that case, learning the mapping for each individual talker may not be critical.

The third gap concerns what human listeners actually do: (3) *Can listeners learn to adapt their implicit expectations about the cue distributions?* There are so far only a few studies that directly manipulate the prosodic input, or reliability with which prosodic cues support a particular meaning (Kurumada et al., 2017; Nakamura et al., 2019; Roettger and Franke, 2019). Kurumada et al. (2017), for example, used between subject conditions in a perceptual learning paradigm (e.g., Eisner & McQueen, 2005; Kraljic and Samuel, 2005; Norris et al., 2003; Liu & Jaeger, 2018). They demonstrated that *a priori* ambiguous tokens receive distinct interpretations according to the patterns seen in the input. However, none of these studies has made a direct link between phonetic distributions in production and listeners' recognition of prosodic categories. Laying out such link forms a core of the current investigation.

To address questions (1) and (2), we administer a large-scale production experiment to assess within- and cross-talker variability associated with productions of declarative question vs. statement prosody. We qualitatively and quantitatively assess the variability of phonetic cue distributions both within and across talkers. We then use a computational model ("ideal observers") to test how much recognition can *in principle* improve if listeners learn the talker-specific distributional statistics. We do so both with the raw, un-normalized, cues and while considering commonly assumed types of normalization. To address question (3), we conduct a comprehension experiment to examine: a) how cue distributions in the production data can predict listeners' recognition judgments at the outset of the exposure; and b) whether listeners continue to adapt their judgments in response to input from a particular talker.

At its essence, we submit, prosodic processing is inference over noisy and variable perceptual input. As such, listeners' judgments are likely impacted by their knowledge of an *expected* structure of phonetic cues and how the structure varies across contexts (e.g., talkers). Despite the increasing interests in prosodic variability, it has so far not been fully recognized just how far-reaching the consequences of the distributional knowledge are (but see Cangemi and Grice, 2016). Declarative questions vs. statements (e.g., *It's raining?* vs. *It's raining*) offer a suitable test-case for this exploration. Prosodic realizations of questions and subtle differences across different question types (e.g., declarative vs. interrogative) have been studied extensively (Geffen and Mintz, 2017; Grabe, 2002; Haan, 2001). Unlike many other prosodic categories, the question vs. statement contrasts in their prosodic features as well as in their meaning are relatively clear-cut and accessible to both talkers and listeners (Bartels, 1999; Bögels and Torreira, 2015; Couper-Kuhlen and Selting, 1996; Doherty et al., 2003; Doherty et al., 2004; Gussenhoven, 1999). In the general discussion, we paint a broader picture of how our approach might scale to other types of prosodic categories (e.g., pitch accents).

2. Experiment 1: characterizing the variability of prosodic production

To examine distributional statistics and their potential impacts on comprehension (Questions 1 and 2), we construct a database of declarative question vs. statement productions from 65 native speakers of American English. We phonetically annotate the F0 and duration of all syllables in those sentences to verify that declarative questions are distinguished from statements primarily in the F0 contour during the final syllable (i.e., −*ing*).

To address question (1), we visually examine the distribution of the F0 and duration, depending on whether the utterance is intended to be a question or a statement. We consider this comparison for both raw, un-normalized cues (F0 and syllable duration), and for normalized cues. To anticipate the results, we find that there is substantial overlap between the joint F0-duration distributions of question and declarative productions. This overlap persists under multiple commonly applied normalizations. We show that much of this overlap is avoided if the F0-duration distributions are visualized separately by talker. In other words, the extensive overlap across talkers results from cross-talker variability rather than from lack of consistency within each talker.

To address question (2), we employ Bayesian ideal observer models to quantify how critical learning talker-relevant information is for comprehension. Specifically, we compare a set of models trained in different cue spaces (e.g., un-normalized vs. normalized) and with different amounts of talker-information. We note that there is a wide range of normalization methods and information assumptions one can apply, making it impossible to test all possible models. Here, we present and compare six models that represent varying levels of talker-specificity (in the General Discussion, we outline ways in which one can extend the current approach).

2.1. Methods

2.1.1. Participants

65 (17 male, 48 female) undergraduate students of University of Rochester participated in exchange for monetary compensation. The number of participants was determined based on a similar large-scale production experiment conducted in the past (Buxó-Lugo et al., 2018). All participants were native speakers of American English and naïve to the purpose of the experiment.

2.1.2. Materials

We constructed 24 sentences in the form: *It's X-ing* (e.g., *It's raining*). The verb stems used were all monosyllabic but varied in their syllable structures (see Appendix I). Diverse lexical contents were included to emulate the variability in natural language use stemming from the segmental features of words as well as their semantic (as well as situational) information that could affect production of the *It's X-ing* utterances.

2.1.3. Procedure

Each subject attended a recording session individually in a recording booth. Subjects were instructed to read the sentence on the screen first and then produce the sentence out loud as a question or a statement according to the punctuation displayed at the end of a sentence (e.g., *It's raining* {./?}). After producing each sentence, subjects clicked on a button on the screen to proceed to the next sentence in a self-paced manner. They produced all 24 sentences, once as a question and once as a statement, in unconstrained randomized order (48 trials in total).

2.1.4. Annotation

We excluded tokens with production errors such as false starts, insertion of filled and unfilled pauses mid-sentence, and production of a wrong verb (segmental change) (4.7% of all tokens). After the exclusions, 2974 recorded sentences were segmented into the following three syllables: 1) *it's*, 2) the stressed syllable (verb stem), and 3) *-ing*. The segmentation and annotations were conducted by the three experimenters: two took the first pass through the recordings based on the agreed upon criteria (Supplementary Information), and the third person verified the annotations. Any inconsistencies were resolved through discussions. We extracted F0 (mean F0 across each syllable) and duration of these three segments using Praat (Boersma and Weenink, 2012; Moulines and Charpentier, 1990). For ease of comparison to previous work on question and statement prosody, we show F0 in Hz, rather than on a log-transformed scale, such as Mel (Stevens et al., 1937) or Bark (Zwicker, 1961).

Replicating previous work (e.g., Bartels, 1999; Hedberg et al., 2017; Patel and Grigos, 2006), question and statement productions differed primarily in the F0 of the final syllable (i.e., −*ing*) (Fig. 1A). For this reason, the analyses of variability we present below will focus primarily

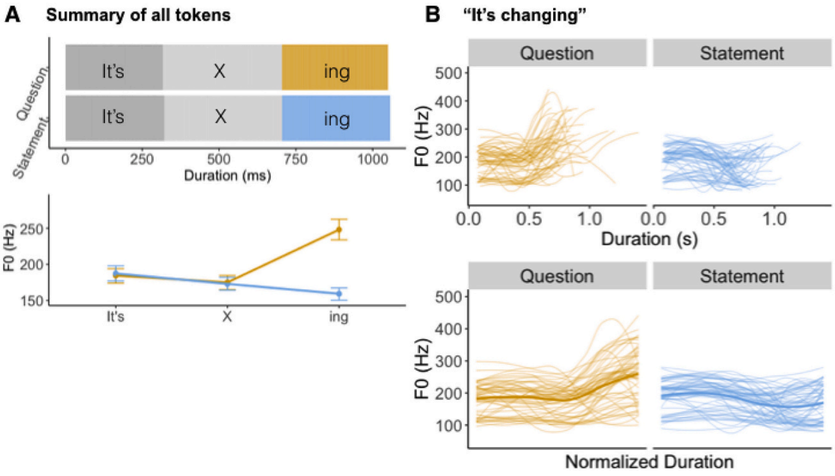

Fig. 1. Pitch contours for each sentence category (questions and statements) in Experiment 1. A: Mean syllable duration (top) and mean F0 values with bootstrapped 95% CIs (bottom). B: Productions of "It's changing" by the 65 talkers in the dataset. Top: Un-normalized, raw, cues plotted along the dimensions of duration (x-axis) and F0 (y-axis). Bottom: Cues after duration is equated for all the tokens. Bold lines indicate average contours across all tokens.

on the third, and final, syllable. It is important, however, to note that the difference in the overall *means* of the two categories conveys little information about the extent to which the *distributions* of the two categories overlap. It is in fact the latter—the extent to which the phonetic realizations of questions and statements overlap—that determines the perceptual ambiguity that listeners must resolve. To illustrate, Fig. 1B shows the pitch contours of all 65 talkers for one item type (*It's changing*). The heterogeneity and the cross-category overlap of the distributions provide initial support to the assumption that there is no discrete mapping between the prosodic categories and cues used to encode the categories (i.e., lack of invariance).

2.2. *Visualizing distributions of mean F0 and duration of the utterance-final syllable*

We first present the marginal (i.e., talker-independent) distributions of raw F0 and duration—i.e., ignoring both talker identity and potential normalization. We then zoom into characteristics of productions by individual talkers. This serves as an initial illustration of the substantial variability *across* talkers in the realization of prosodic categories. We then show that the variability of the marginal distributions is only minimally ameliorated under commonly assumed normalization processes. Together, these visualizations address the first goal of Experiment 1, to provide an intuitive characterization of the within- and cross-talker variability in prosodic realization for the case of declarative questions and statements.

2.2.1. *Marginal distribution of raw, un-normalized phonetic cues*

Fig. 2 shows the distributions of raw (un-normalized) F0 and duration measured on the final syllable. The visualization here follows the majority of previous work that pools the data from all participants without applying any form of normalization to the cues (as in e.g., Buxó-Lugo et al., 2016; Patel and Grigos, 2006; Prieto et al., 2010).[3]

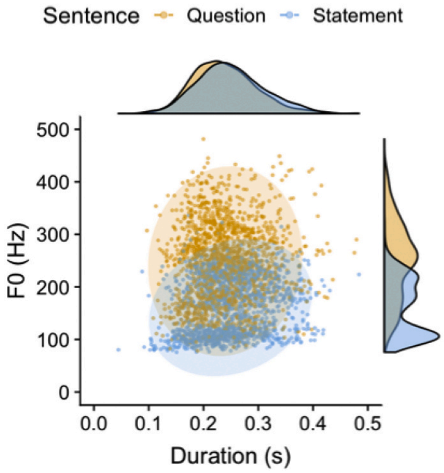

Fig. 2. Distribution of un-normalized utterance-final F0 and duration for each sentence category across all talkers. Points show individual tokens. Ellipses show bivariate Gaussian 95% CI of each category.

The data in Fig. 2 support three key observations about the marginal distribution of raw cues. First, in accordance with the previous work and our general intuition (Bartels, 1999; Bögels and Torreira, 2015; Bolinger, 1986; Couper-Kuhlen and Selting, 1996; Gussenhoven, 1999; Hedberg et al., 2017; Pierrehumbert and Hirschberg, 1990), questions and statements are overall better separated along F0 than syllable duration. In fact, Fig. 2 suggests no separation along duration at all. This contrasts to a previous finding with smaller sample sizes (12 talkers across 3 age groups) that questions are associated with longer duration of the final syllable (Patel and Grigos, 2006). We will examine in the next section if such a tendency was exhibited by some, if not all, talkers.

Second, as we anticipated based on Fig. 1B, there is substantial overlap between acoustic cue distributions of the question and statement categories, even along F0. Prior to considering normalization or talker-specific interpretation of cues, a large proportion of the question and statement productions would fall into the region that can feasibly be associated with either category, making these tokens potentially ambiguous to listeners.

Third, the variability of F0 seems to be larger for the question category—characterized by the lower densities and wider variances of distributions in Fig. 2—than for the statement category, at least when measured in Hz. Talkers are comparatively consistent in how to encode the statement category (A similar asymmetry was reported by Cangemi and Grice (2016)). One possible explanation for this is that statements are the "default" or unmarked pattern of production. Alternatively, or

[3] The distributions in Figure 2 appear to be bimodal along F0. This bimodality is in part attributable to by-gender differences in pitch, which we discuss in Section 3. We also found that a subset of female talkers exhibited lower baseline pitch compared to the other female talkers, further contributing to the bimodality. Differences across items (i.e., verbs) did not significantly contribute to this pattern.

additionally, it is possible that the productions elicited as questions reflect *multiple* types of questions that are characterized by distinct semantics and accompanied slightly different phonetic realizations (e.g., "assertive" vs. "inquisitive" declaratives with rising boundary tone, Jeong, 2018). That is, different participants might have called to mind slightly different types of declarative questions. A third possibility is that the larger variability of questions is an artifact of comparing categories in Hz, rather than a log-transformed scale like Mel or Bark. Although of importance to theories for question sentences in English, we do not explore it further since it is not central to the present purpose.

2.2.2. Talker-specific distribution of raw, un-normalized phonetic cues

To what extent is the category overlap seen in Fig. 2 due to variable realizations of prosody *between talkers*? Fig. 3 shows six sample talkers for a comparison (see Supplementary Information for acoustic distributions for all 65 talkers). This reveals a remarkable degree of variability between talkers. Talkers differ both in terms of category means and in terms of category shapes. This includes cross-talker differences in the magnitude of variability along F0 and duration, as well as the correlation between the two cues. Talkers also vary widely in how their productions of the two categories deviate from what can be *a priori* expected from the marginal distributions. That is, some talkers' distributions largely resemble the marginal distributions (e.g., Talkers (c), (e), and (f)) while others do so less (e.g., Talkers (a) and (b)).

Although F0 was the primary cue for most talkers, some talkers also used syllable duration (e.g., Fig. 3 (c) and (d)). This confirms findings that utterance-final syllable duration *can* be a cue to the question-statement contrast (Patel and Grigos, 2006, as mentioned above). It also serves as a reminder for caution: if only marginal distributions are analyzed (ignoring differences between talkers), studies, especially those with smaller sample sizes, might well come to seemingly conflicting conclusions about the realization of phonetic or prosodic contrasts.

In stark contrast to the substantial variability across talkers, the realization of questions and statements was rather consistent *within each talker*. This is reflected in the small degree of overlap between the two categories within each talker, compared to the marginal distributions. In quite a few talkers (e.g., Talkers (a), (d)-(f)), productions are in fact cleanly separated along F0 if the talker-specific realization of questions and statements is considered.

This suggests that the substantial overlap seen in the marginal distributions in Fig. 2 results in large part from differences *between talkers*, rather than category variability *within talkers*. This provides initial support for the hypothesis that there is systematic structure in the cross-talker variability. A failure to recognize this structure is therefore predicted to impede comprehension (Kleinschmidt, 2019; Kleinschmidt and Jaeger, 2015; McMurray and Jongman, 2011). This motivates the quantitative explorations of the *benefit* listeners would gain by learning such talker-specific structures and idiosyncrasies that we present below. Before turning to that exploration, we ask whether variability can be removed, or at least substantially reduced, by normalization.

2.2.3. Is normalization a solution to the lack of invariance?

To test if the variability can be removed by normalization, we transform both phonetic cues (i.e., F0 and duration) by crossing two different types of baselines used in previous work: for a given *utterance* (extrinsic normalization) and for a given *talker* (intrinsic normalization).

To normalize cues within an *utterance*, we subtract values of the second syllable from those of the final one to compute the difference between the two. This reflects the idea that the question vs. statement categories are recognized by perceiving the relative changes of phonetic cues within an utterance rather than their absolute values (Chodroff and Cole, 2019a; J. B. Pierrehumbert, 1979)—an assumption that is also central to work on utterance-internal calibration of prosodic processing (e.g., Brown et al., 2011; Dilley and Pitt, 2010; Morrill et al., 2014; Reinisch et al., 2011). As we implement it here, those *utterance-normalized* cues entail no talker information other than what occurs within the same utterance (but see, Baese-Berk et al., 2014; Maslowski et al., 2019). As such, this type of normalization is considered cognitively and computationally less taxing compared to those that require estimating a baseline across multiple utterances from the same talker.

To normalize cues within a *talker*, we first compute the mean of a particular cue (i.e., F0 or duration) across all tokens of the final syllable produced by the same talker (irrespective of their category affiliation). We then subtract the mean from the value of each token.[4] Considering a talker-specific baseline is proven effective for compensating cross-talker variability due to physiological differences (e.g., vocal tract length, Johnson, 2005a). Talker-normalized cues have been used, for example, in the C-CuRE model of speech perception, providing a good fit against listeners' perception of segmental categories (McMurray and Jongman, 2011). This type of normalization goes beyond taking a running average of acoustic-phonetic cues within an utterance. It can therefore be considered a form of learning, whereby listeners store information about past exposures to a given talker. Unlike the talker-specific models which we discuss below, however, models like C-CuRE do not consider category-specific differences between talkers.

Finally, to account for the possibility that both types of normalization are applied, we first implemented utterance normalization and then subtracted out talker's mean of the utterance-normalized cues values across the talkers' tokens from both categories (similar to what was used in Ryalls et al., 1994).

Fig. 4 compares the un-normalized (Fig. 4A, same as Fig. 2) and normalized marginal distributions (Fig. 4B-D). Even after normalization, question and statement realizations continue to exhibit substantial overlap. There is thus little evidence that utterance- or talker-normalized cues lead to a clear separation of the categories. (We quantify the degrees of separations between the categories in Section 2.3).[5] This shows the depth of the "lack of invariance" problem in prosodic category recognition. A majority of past research has assumed that the "rise" and "fall" of pitch, a critical predictor of the question vs. statement categories, can be rather straightforwardly extracted once a contextual baseline is considered. The data in fact suggest that no simple normalization may suffice.

Fig. 4E shows by-talker distributions of F0 (thin dotted lines) after both types of normalizations were applied. As seen in Fig. 3, each talker's distribution is tightly clustered around its mean. However, means and variances of these talker-specific distributions vary between talkers, contributing to the overlap between the marginal distributions of the two categories (solid lines). In short, commonly applied normalizations do *not* seem to effectively ameliorate cross-talker variability in the realization of questions and statement prosody.

2.3. Predicting expected gains in comprehension from production data

We now turn to question (2). We ask if talker-specific distributional learning is expected to improve recognition above and beyond what can be achieved via normalizations. To this end, we train different ideal observer models and use them to quantify the expected benefit of the knowledge of talker-specific distributions to recognition accuracies (see

[4] This method for talker-normalization is also known as "centering." It accounts for baseline differences across cue means but not variances. We have also considered z-score transformations of cues to account for variability of cue variances. We do not report the results of z-scoring here for the sake of simplicity and ease of comparison between our models and previously implemented models. In particular, we chose the current method of talker-based normalization because of its conceptual similarity to McMurray and Jongman's (2011) C-CuRE model.

[5] There is some evidence that the bimodality of the F0 distribution in the raw cue distributions (Figure 4A) is reduced by normalization (Figure 4B-D). However, this only mitigates, but does not remove, the category overlap.

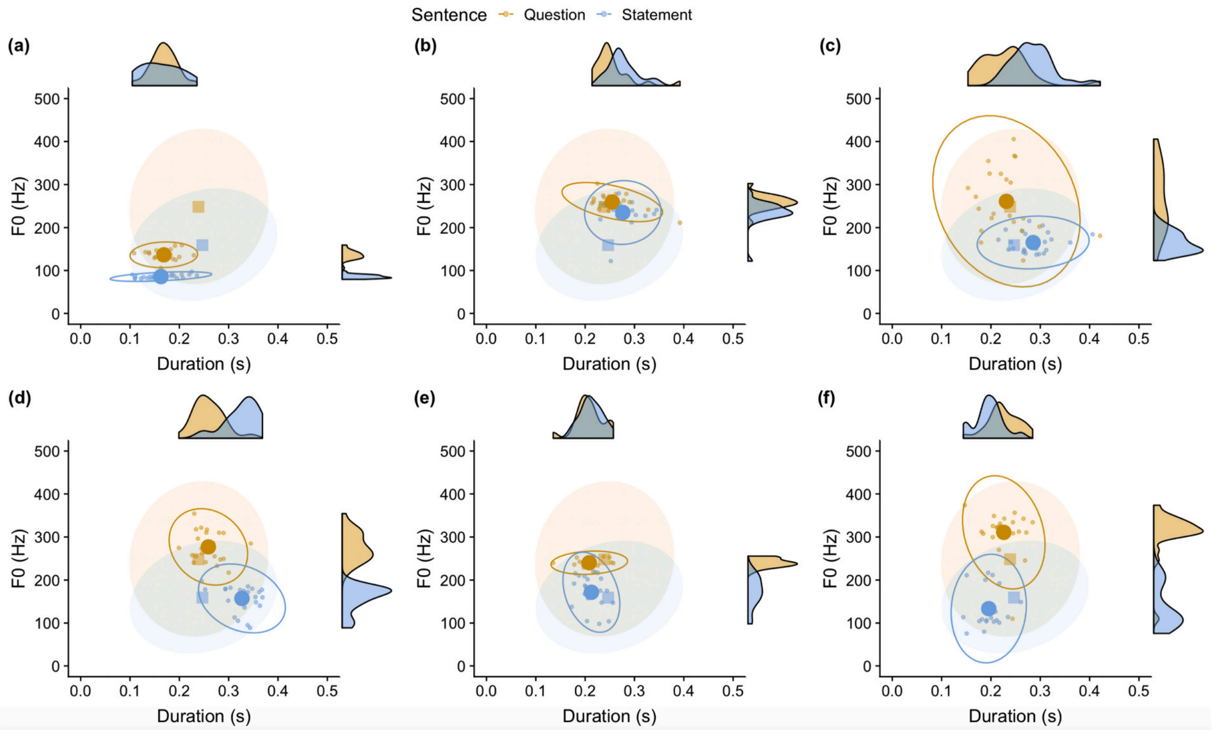

Fig. 3. Distribution of un-normalized utterance-final F0 and duration for 6 example talkers (a) –(f) from Experiment 1. Small points show individual tokens produced by the talker. Ellipses (solid lines) indicate bivariate Gaussian 95% CI of that talker's categories. Filled ellipses in the background show bivariate Gaussian 95% CI of marginal distributions of each category from Fig. 2.

recent work on segmental perception, Kleinschmidt, 2019; Kleinschmidt et al., 2018).[6] Before we turn to our design, we briefly provide some background on this approach.

Ideal observer models assume that listeners infer the posterior probabilities of the intended category (here: questions or statements) from the observed acoustic input based on their implicit knowledge of the relevant statistics. Bayes' rule states that the posterior probability that a token *i* is a member of a category *c*, $p(c|\,token_i)$, is proportional to the *likelihood* of observing token *i* given the category *c*, $p(token_i|\,c)$, and the *prior probability* of hearing a category *c* in this context, $p(c)$. The posterior probability of a statement is 1 minus the posterior probability of a question, so the posterior probabilities of the two categories always add to 1 (Clayards et al., 2008; Kleinschmidt and Jaeger, 2015; Kronrod et al., 2016; Rohde and Kurumada, 2018).

To predict recognition judgments, it is further necessary to specify a decision rule that allows listeners to make a choice between the two categories based on the posterior probability of each category (for introduction, see Friedman and Massaro, 1998). We follow previous work on speech perception (Dahan et al., 2001; Feldman et al., 2009; Luce and Pisoni, 1998) and assume Luce's choice rule (Luce, 1963), which assumes that observers respond by *proportionally sampling* from the posterior distribution of the alternative categories (i.e., the question category and the statement category) and that other irrelevant alternatives play no role in the decision. Under this decision rule, the probability of recognizing a token *i* as a question is simply:

$$p(resp = question|token_i) = \frac{p(token_i\mid question)*p(\,question)}{p(token_i\mid question)*p(\,question) + p(token_i\mid statement)*p(statement)}$$

Critically, the likelihood term depends on the distribution of the cues. It follows that the posterior probabilities depend on the distribution of cues for both the question and statement categories. Fig. 5 visually illustrates how ideal observers link production data (i.e., the distribution of cues, shown in the left panels) to predictions about recognition (i.e., the ideal categorization function, shown in the right panel). The top panel of Fig. 5 makes this link for the case in which only one cue is considered (here: F0). The bottom panel extends the exact same reasoning to the case in which two cues (here: F0 and syllable duration) are considered. In both cases, the left panels show the *simulated* marginal cue distributions of the two categories (i.e., questions and statements).[7] The right panels show the resulting ideal categorization function if the prior probability of questions and statements is the same ("a uniform prior"). Non-uniform priors—for example, because contextual factors bias towards a question or statement interpretation—would simply shift the categorization function to the left or right, without changing its shape or slope.[8]

For the present purpose, we considered eight different ideal observers, fully crossing: whether the cues are utterance-normalized or not (2 levels); whether the cues are talker-normalized or not (2 levels), and the use of talker-specific vs. marginal, *talker-independent* distributions

[6] For the present purpose, exemplar-based models (Foulkes and Hay, 2015; K Johnson, 2005; J. B. Pierrehumbert, 2001) or commonly used measures of category separability (Newman et al., 2001; Xie & Jaeger, 2020) would present alternative approaches that should yield qualitatively identical results.

[7] We follow previous work on segmental speech perception and assume Gaussian categories (e.g., Bejjanki et al., 2011; Clayards et al., 2008a; Feldman et al., 2009; Kronrod et al., 2016; McMurray et al., 2009; Norris and McQueen, 2008). This is not a critical assumption (for discussion, see Kleinschmidt and Jaeger, 2015).

[8] We will revisit this assumption of uniform prior as we compare model predictions with human judgment results in Experiment 2.

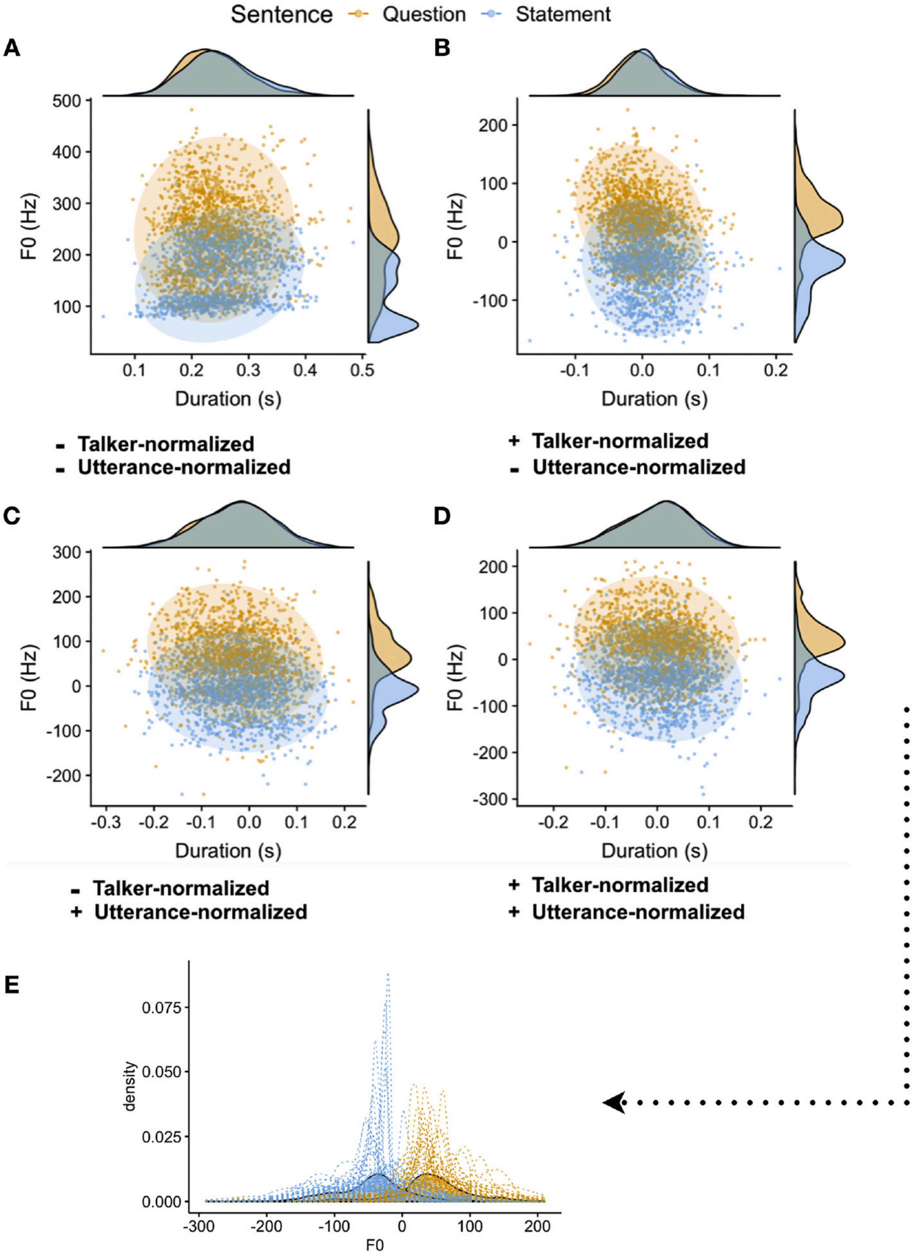

Fig. 4. Panel A-D: Distribution of utterance-final F0 and duration for each sentence category across all talkers, depending on whether utterance- and/or talker-normalization is applied. Points show individual tokens. Ellipses show bivariate Gaussian 95% CI of each category. Panel E: The solid density distributions are identical to those shown for F0 in Panel D; the dotted lines show the density distributions from individual talkers.

for the question and statement categories (2 levels). As talker-specific distributions for the question and statement categories entail talker-normalization (but not vice versa), this resulted in six unique ideal observers, shown in Fig. 6. Four models were trained on the *talker-independent* distributions. Each of these models was fit to phonetic cues normalized in one of the four ways discussed above (Fig. 4). The other two models were trained on the *talker-specific* distributions of utterance-normalized or un-normalized cues. A better performance of the talker-specific models (indicated as white circles in Fig. 6) compared to the talker-independent models (indicated as gray circles in Fig. 6) will suggest that there is *in principle* benefit for listeners to track talker-specific cue distributions.

2.3.1. *Methods*

Each of the six ideal observers represents the question and statement categories in terms of their two-dimensional means (for F0 and duration) and two-by-two-dimensional variance-covariance matrices (with the diagonal containing the variances along F0 and duration, and the off-diagonal containing the category's covariance between F0 and duration).

Training an ideal observer thus requires fitting the mean and covariance matrix. Once trained, the ideal observer can then be evaluated against test data. Training and testing procedures of the ideal observer models are summarized in Fig. 7. To avoid over-fitting, we use five-fold cross-validation to split the productions from Experiment 1 into training data and test data. On each fold, 80% of the data are used for training and the remaining 20% of the data are used for test. This means that out of 24 tokens per category per talker, 19–20 tokens (80%) are assigned as the training data and 4–5 tokens (20%) are test data. Across the five-folds, each production is used four times for training and once for test (for further details, see Supplementary Information).

For each fold of each talker, the test data are held constant between

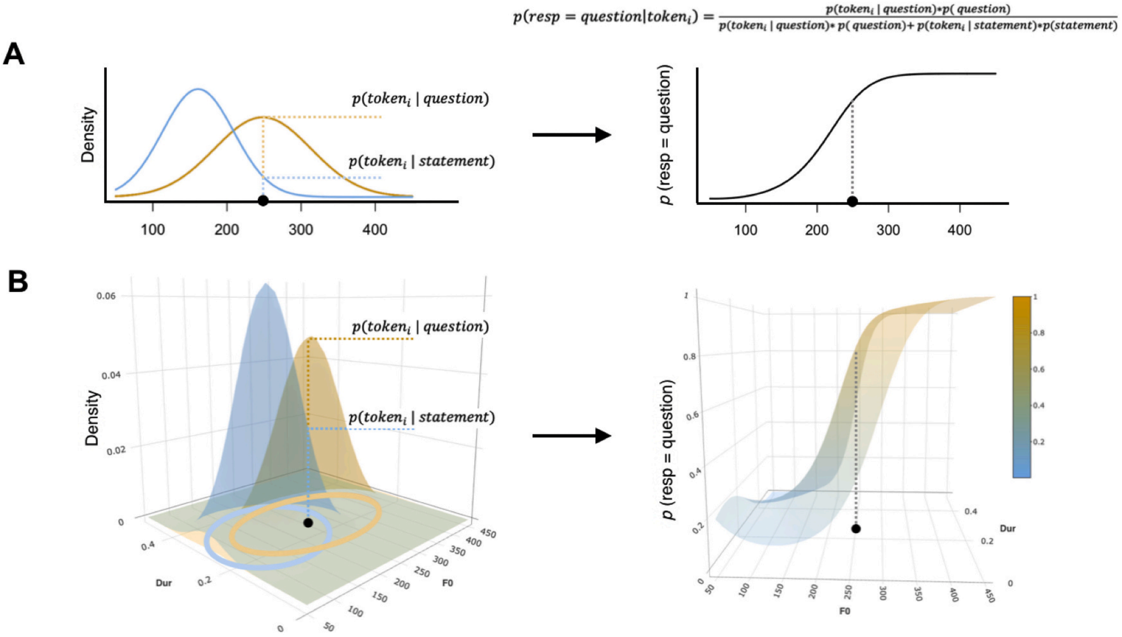

Fig. 5. Illustration of how response curves change as a function of the phonetic cue distributions. Panel A illustrates the case of a single cue. Panel B illustrates a case with two cues. The left panels represent acoustic cue distributions (based on raw cues) for questions and statements (color-coded). The right panels represent ideal observers' response curves under the cue distribution pattern. The dotted lines (color-coded) in the left panel mark the likelihood of token i under each category distribution; the black dotted line in the right panel marks the posterior probability of a question response for token i.

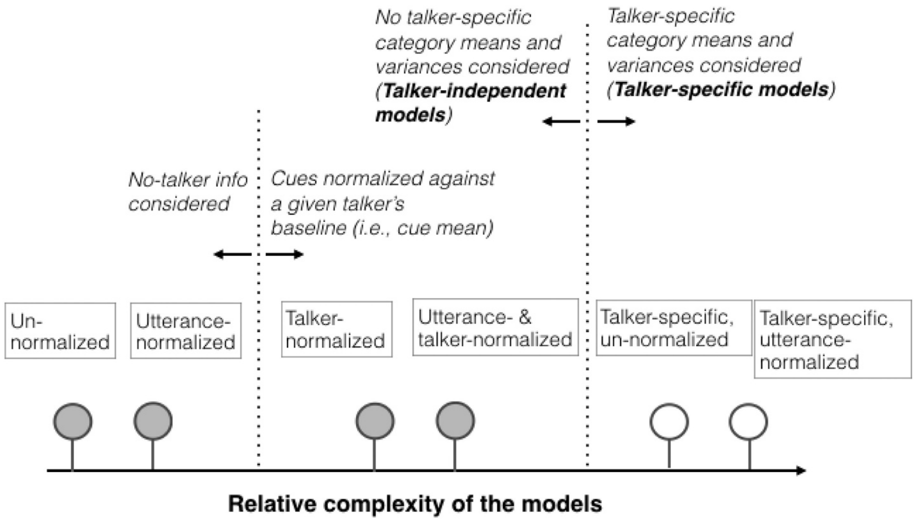
No talker-specific category means and variances considered (Talker-independent models)
Talker-specific category means and variances considered (Talker-specific models)
No-talker info considered
Cues normalized against a given talker's baseline (i.e., cue mean)
Un-normalized
Utterance-normalized
Talker-normalized
Utterance- & talker-normalized
Talker-specific, un-normalized
Talker-specific, utterance-normalized
Relative complexity of the models

Fig. 6. Six ideal observers (four talker-independent models and two talker-specific models) and their relative levels of complexity in terms of the amount of information considered and stored.

the talker-independent and the talker-specific models. The training data for the talker-independent models consist of the training folds from all talkers and are therefore identical for all test talkers (see Fig. 7). Notably, the number of training tokens in the talker-independent distribution far exceed those in a by-talker distribution. This creates a bias in favor of the *talker-independent* model—and thus *against* our hypothesis. But this also reflects the situation that human learners face: there is always more data to estimate the talker-independent distribution than there is to estimate talker-specific distributions.

2.3.2. Results

Fig. 8 shows the performance of each of the six ideal observer models (implemented using R package phondisttools; Kleinschmidt, 2019). The performance of all six ideal observer models during test was significantly above chance (50%). This indicates that the production database is sufficiently large to reliably estimate the relevant distributions (recall that test data never overlapped with the training data).

We employed mixed-effect logistic regression using the *glmer* function from the *lme4* package in R (lme4 1.1–21; Bates, Mächler, Bolker, & Walker, 2015; R version 3.5.0; R Core Team, 2020) to compare the predicted accuracy across the six ideal observers (Jaeger, 2008). Specifically, we analyzed performance of the four talker-independent and two talker-specific models as a 2 × 3 design, crossing utterance-normalized cues (contrast-coded, yes or no) and talker-specificity (sliding difference-coded with two orthogonal contrasts comparing no talker-normalized cues < talker-normalized cues < talker-specific cue distributions). The analysis further included the full random effect structure by test talker: random by-talker intercepts and by-talker slopes

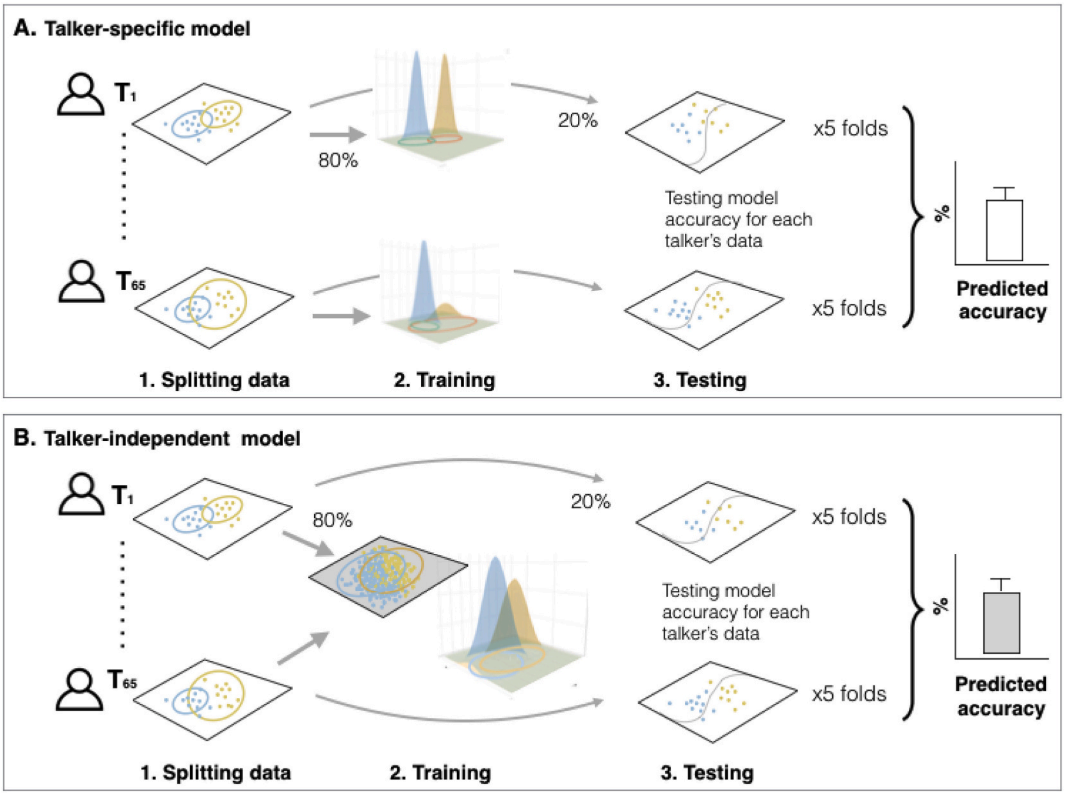

Fig. 7. Procedure for training and testing ideal observers. Shown is one-fold of the five-fold cross-validation. Panel A: For the talker-specific models, 80% of the individual talker's data was used to train a model and 20% was to test the model. Panel B: For talker-independent models, 80% of each talker's data was aggregated to train a single talker-independent model. Data used for training and testing never overlapped.

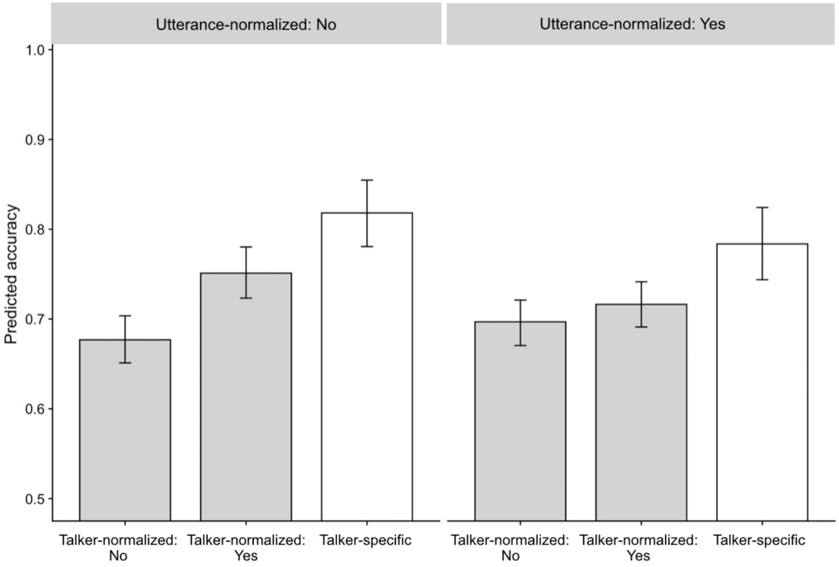

Fig. 8. Predicted accuracy of the six different ideal observer models. Error bars show bootstrapped 95% CIs of predicted by-talker accuracy (an average of the five folds was computed for each talker, CIs are over these by-talker means). Note that the y-axis starts at chance performance (0.5). *p*-values show simple effects from a mixed-effects logistic regression (see text for details).

for the two fixed factors, as well as their interaction.[9]

Fig. 8 shows the predicted accuracies of the six models. Analyses supported three observations. First, there was no main effect of utterance-normalized cues ($\widehat{\beta} = 0.019$; $z = 0.66$; $p = .51$). Second, talker-independent models with talker-normalized cues performed significantly better than those without ($\widehat{\beta} = 0.51$; $z = 7.87$; $p < .0001$): interpreting F0 and durational cue values *relative to what is expected for a given talker* yields better categorizations. Third, the talker-specific

[9] *P*-values are computed from the t-distribution using degrees of freedom on the Satterthwaites approximation as implemented in the lmerTest package version lmerTest_3.0–1, (Kuznetsova et al., 2017).

models performed even better than the talker-independent models with talker-normalized cues ($\widehat{\beta} = 0.28$; $z = 2.68$; $p = .0007$). Interactions indicated that the difference between the two talker-independent models was significantly larger when utterance-based normalization was not applied ($\widehat{\beta} = -0.33$; $z = -5.50$; $p < .0001$). The difference between talker-specific models and talker-independent models with talker-normalized cues was not affected by utterance-based normalization ($\widehat{\beta} = -0.07$; $z = -1.09$; $p = .28$).

A question of importance was whether the advantage of talker-specific models indeed originated from the talker-specificity in the prosodic cue distributions rather than from other differences in ways that these models were trained and tested. To answer this question, our follow-up analyses examined two possible sources of discrepancies between the talker-independent versus talker-specific models (for details, see Appendix III). First, as shown in Fig. 7, a talker-specific model consisted of 65 models, each trained on a given talker's cue distributions. To eliminate the possibility that this internal complexity itself was sufficient to yield a better performance, we trained a new set of talker-independent models by randomly sub-sampling so the number of training tokens would match that in a talker-specific model. That is, we constructed 65 talker-independent models, each trained on an arbitrary grouping of training tokens from various talkers. Not surprisingly, these models performed significantly worse than talker-specific models (see Fig. A1 Panel A in Appendix II), suggesting that the higher predicted accuracy of the talker-specific models (Fig. 8) was not attributable to the number (and complexity) of the models considered.

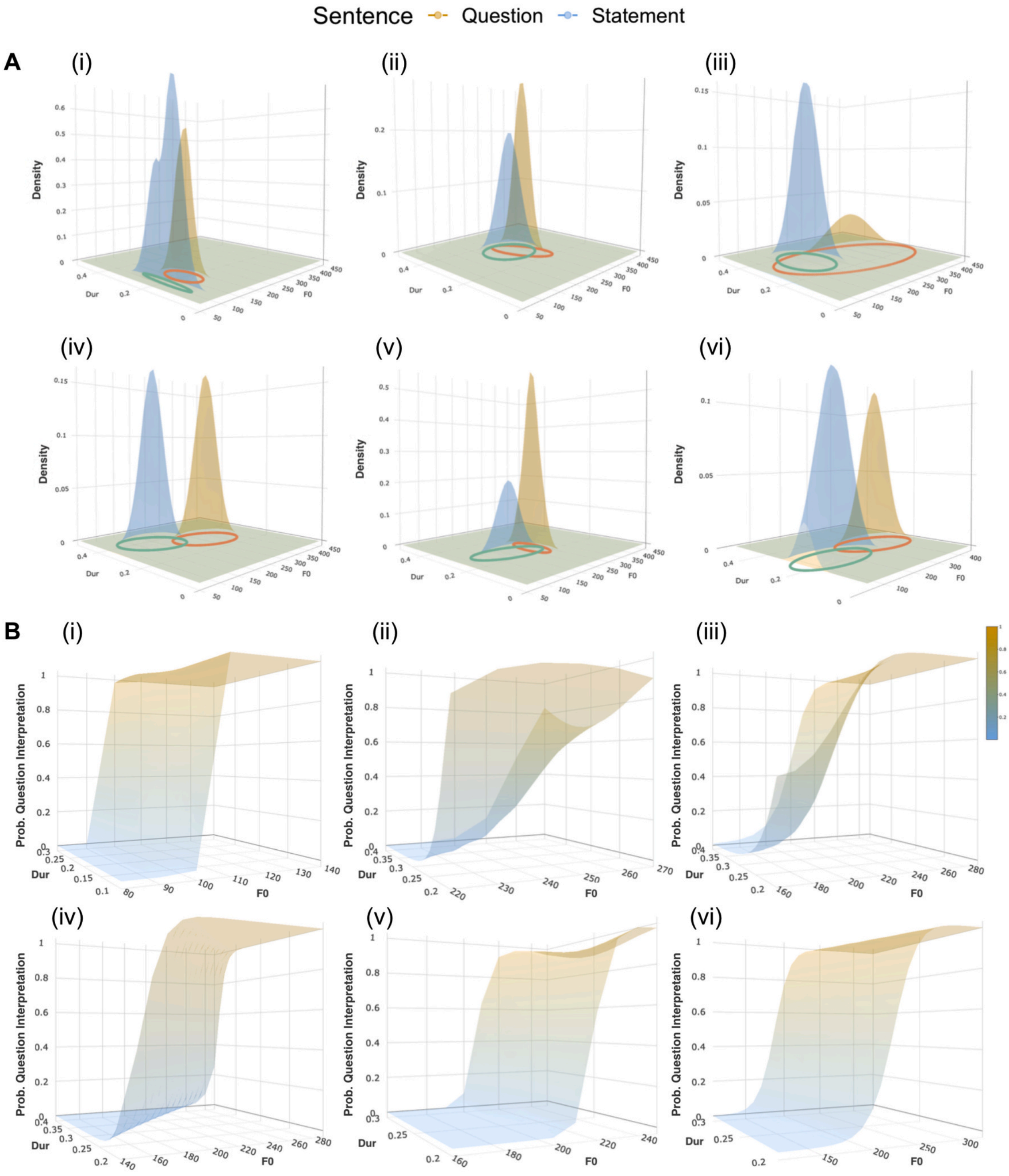

Fig. 9. Panel A: Visualization of talker-specific ideal observers (using raw, un-normalized cues) for the same six talkers presented in Fig. 3. Ellipses show 95% of the probability mass of the bivariate Gaussian categories. Panel B: Categorization functions for each of the six talker-specific ideal observers. Y-axis indicates the posterior probability of a question interpretation.

Second, we examined whether the benefit associated with a talker-specific model could be seen if training and test tokens came from different talkers. To this end, we tested the same talker-specific models as described above on data randomly sampled across talkers. Conceptually, these models simulate a situation where listeners encode prosodic cue distributions separately for each talker but indiscriminately apply this information upon hearing tokens from unfamiliar talkers. Again, these models yielded substantially poorer performances relative to cases where they were tested on tokens from the matching (i.e., training) talkers (Fig. A1, Panel B in Appendix II). Taken together, these results lend additional support to our conclusion that the higher predicted accuracies of the talker-specific models, compared to the talker-independent ones, stem specifically from the underlying cross-talker differences in prosodic cue distributions.

Additionally, the better performance of the talker-specific models (i.e., the white bars in each panel of Fig. 8) over the talker-independent models with talker-normalized cues (i.e., the middle gray bar in each panel of Fig. 8) suggests that talkers differed in their distributional statistics of the two prosodic *categories*—beyond differences in the overall mean of the cues (which talker-normalization removes). This is supported by visual inspections of Panel A of Fig. 9, which shows the talker-specific models for the same six individual talkers as in Fig. 3. The six panels show the bivariate Gaussian distributions of the talker-specific models, projected back into the original (un-normalized) cue space. As we already saw above, these six talkers differ in their category means, variances (which are positively related to the length of the ellipses along the x- and y-axes) and co-variances (the orientation of the ellipsis). These differences result in different predicted categorization functions across the six talkers (Fig. 9, Panel B). Deriving talker-specific categorization functions, compared to applying the same function to all talks, will thus *in principle* facilitate recognition of prosodic categories in face of substantial cross-talker variability.

2.4. Summary and discussion

To answer questions (1) and (2), Experiment 1 presented a large-scale, phonetically annotated database of declarative question and statement productions. These data are shared via OSF (https://osf.io/kr7y6/). With 65 talkers and 24 minimal pair productions per talker (for a total of 2974 productions), it is possible to reliably estimate the means and covariance matrices of question and statement categories, not only across talkers (for which there is more data) but also at a talker-specific level.

If only considering marginal cue distributions, there was significant overlap between the question and statement categories along utterance-final F0 and duration. This overlap was primarily caused by variability between talkers, rather than variability within talkers. Ideal observer analyses confirmed this: among the models considered here, those with an increased amount of talker information resulted in higher predicted accuracies. Utterance normalization improved categorization accuracy only when by-talker means or talker-specific distributions were *not* considered (see Fig. 8). This parallels findings from segmental speech perception that listeners seem to compute cues relative to expectations for a given talker (e.g., C-CuRE model for fricatives, McMurray and Jongman, 2011).

Our results suggested further that listeners can in theory do better than just normalizing cues relative to the talker's overall mean. Unlike the C-CuRE model, the talker-specific model considered not only the talker's overall mean, but also the category-specific means and co-variances, and achieves the highest categorization accuracy. These talker-specific models showed higher accuracies than the other models in categorizing unseen data.

It is possible that considering other cues (e.g., alignment of a pitch peak, Cangemi and Grice, 2016), measurements and transformations (e.g., using log-transformed F0) and/or more sophisticated utterance-based normalization would lead to different results. In fact, previous studies suggest that pitch movements in an area distant from the utterance final tone (e.g., a pre-nuclear region) impact categorizations of questions vs. statements (Chodroff and Cole, 2019a; Petrone and D'Imperio, 2011; Studdert-Kennedy et al., 1973). This type of effect of an earlier part of an utterance was only approximated in the current utterance-based normalization. Future tests of alternative normalization methods are facilitated by open access to our data.

At the very least, our results show that storing past experiences with individual talkers is *one* way to substantially improve comprehension accuracy. Based on the predicted benefit of storing talker-specific cue distributions, we now ask whether human listeners can indeed draw on their knowledge of talker-specific cue distributions in their recognition of prosodic categories.

3. Experiment 2: Adapting to talker-specific prosody in comprehension

Experiment 2 addresses question (3) outlined in the introduction: *Can listeners learn to adapt their implicit expectations about the distributional statistics?* The paradigm we employ resembles perceptual recalibration experiments on segmental speech perception (e.g., Kraljic and Samuel, 2005; Norris et al., 2003). The design is shown in Fig. 10. Participants are exposed to declarative question and statement productions—similar to those from Experiment 1—from an unfamiliar talker. Like in perceptual recalibration experiments, the exposure input is effectively labeled so that the intended interpretation (question or statement) is clear to participants. Both before and after exposure, we assess participants' categorization function for unlabeled input from the same talker. Between participants, we manipulate the prosodic realization of questions and statements during exposure. By comparing listeners' categorization function on the same input before and after exposure, we can assess whether input from an unfamiliar talker changes how listeners interpret input from that talker.

We conduct two versions of this experiment—one with a female talker and one with a male talker. This serves two purposes. It provides an internal replication of the effect of exposure, testing the robustness of any effects we find. It also allows us to ask if listeners' initial processing of prosodic input from an unfamiliar talker draws on their previous experience with talkers of a given gender. We use the data from Experiment 1 to construct two *gender-specific* ideal observer models (one female and one male) that we compare against the talker-independent models developed in Experiment 1. Under the hypothesis that listeners learn and store talker- and group-specific prosodic realizations, we expected the gender-specific models to better predict participants' pre-test responses than talker-independent models do.

3.1. Participants

360 participants were recruited through an online crowdsourcing platform Amazon Mechanical Turk (https://www.mturk.com/). The data for the two talker conditions was collected in two separate sessions, with data for the female talker condition elicited first. No participant took part in both conditions. Two participants in the female talker condition and six in the male talker condition were excluded for providing either only questions or only statement responses throughout pre- and post-test. This left 352 participants for analysis. Participants were self-identified native speakers of American English and received monetary compensation for their participation.

3.2. Materials

Due to privacy considerations, the experimental protocol used for Experiment 1 did not allow us to distribute the recorded tokens as

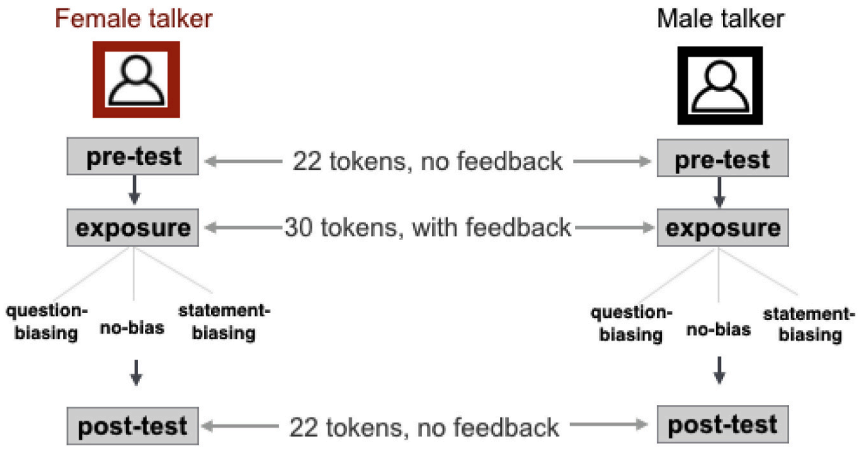

Fig. 10. Design of Experiment 2. Pre- and post-test: Participants provide categorization judgments on tokens sampled uniformly along a 11-step continuum ranging from clear statement tokens to clear question tokens (in a randomized order). Exposure: Participants are assigned to one of the three exposure conditions that differ in how questions and statements are realized. Exposure always includes feedback about the intended interpretation of each token.

stimuli for an internet-based survey.[10] We thus obtained new recordings from a female and a male native speaker of American English. This has the additional, critical, advantage that we can obtain multiple recordings of each stimulus sentence, mitigating data loss due to pitch track errors (e.g., pitch doubling and halving), thereby facilitating the reliable estimation of F0 and syllable durations.

Each talker produced two tokens each of six pairs of question and statement sentences (i.e., *It's* {*cooking, booting, cooling, losing, moving, muting*}). Fig. 11A shows these items against the un-normalized marginal distributions from Experiment 1. Just as in other talkers examined in Experiment 1, the two talkers' question vs. statement categories were relatively consistent internally and clearly separated within each talker. Of note is that the two talkers differed substantially from each other in their realization of questions and statements: the statement category produced by the female talker roughly occupied a similar phonetic region as the question category by the male talker. The input from these two talkers would thus result in systematic ambiguity *if no normalization or talker-specificity processing was applied.*

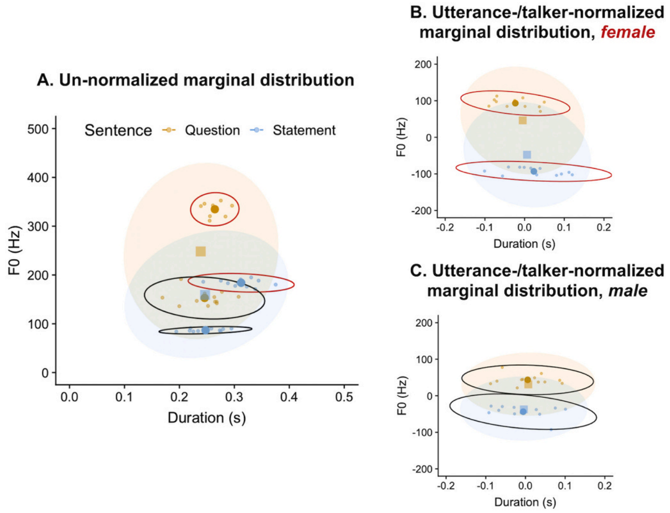

Fig. 11. Phonetic cue distributions of the two talkers recorded for Experiment 2. Panel A: The two talkers' distributions of the question and the statement categories against the un-normalized marginal distributions from Experiment 1. Panels B and C: Utterance- and talker-normalized talker-specific distributions against gender-matched talkers from Experiment 1. Filled ellipses show bivariate Gaussian 95% CI of all tokens of the 65 talkers in Experiment 1; solid ellipses show those of the female and male talkers in Experiment 2, respectively. Squares and large circles indicate the category means of the marginal distributions and talker-specific distributions, respectively. Smaller dots indicate individual tokens of each sentence type by each talker.

Figs. 11B and C show the two talkers' productions against the distributions of *gender-matched* talkers from Experiment 1. Phonetic cues are both utterance- and talker-normalized to compensate for baseline differences among each group. The proximity of the talker-specific category means (i.e., the orange and the blue circles) to the marginal category means (i.e., the orange and the blue squares) indicate that their production patterns can be deemed *typical* given the general patterns observed in each gender group. If listeners can draw on their implicit knowledge of how female and male talkers produce questions vs. statements (and if listeners can normalize phonetic cues), they should be able to reliably categorize the tokens from these talkers prior to extensive exposure.

Acoustic resynthesis of exposure and test stimuli followed the technique described in Kurumada et al. (2017). The recordings (two tokens per item) were first segmented into three regions corresponding to three syllables (i.e., *It's* | *X-* | *ing*) as in Experiment 1. To achieve best results in resynthesis, we paid attention to both a turning point in the F0 contour within the "X-ing" portion as well as the segmental information to delineate the last two syllables. The F0 of each region was sampled at 20 equi-spaced time points, and measures from each time point were aggregated across items to derive mean F0 contours for statement and question contours, following Isaacs and Watson (2010). Similarly, the durations of each region were averaged across items by contour type. Mean F0 contours and durations were then derived by interpolating between values within each region and manipulating F0 and duration of each recording to match the interpolated values using the pitch-synchronous overlap-and-add algorithm implemented in Praat (Boersma and Weenink, 2012; Moulines and Charpentier, 1990).

The five types of items with the /u/ vowel (i.e., *booting, cooling, losing, moving, muting*) were selected as exposure items. 12 step continua for these items were created based on F0 and durational values taken from recordings of *It's moving* to ensure that the 12 steps of all of the

[10] Production data (audio files) can be shared upon request.

exposure items had consistent F0 and durational values.[11] *Cooking* (with the /ʊ/ vowel) was selected as the test item. Continua were constructed based on recordings of *It's cooking* (Fig. 12). Consequently, the exact acoustic values for the exposure and the test items were distinct from each other, which allowed us to test if listeners' learned expectations about cue-category mapping can generalize beyond memory traces of previously heard (exposure) tokens.

As shown in Fig. 12, F0 serves as the primary cue to questions vs. statements for both the female and male talker.[12] Of interest, however, is the role of duration: while questions and statements differed along the duration cue for the female talker, this pattern was much less pronounced for the male talker. The pattern reflects the distributions in original productions of the two talkers (Fig. 11B and C).

We then normed the stimuli with 90 participants using Amazon Mechanical Turk. Each participant heard 24 tokens sampled uniformly from the continuum and provided 2AFC judgments (i.e., Is this a statement or a question?) without feedback. This norming study is reported in Appendix III. Since responses to Step 0 and Step 1 received nearly identical responses, Experiment 2 only uses Steps 1–11. Steps 6 and 7 were the most ambiguous for the female and the male talker, respectively (receiving a "question" response 46.7% and 55.3% of the time, respectively).

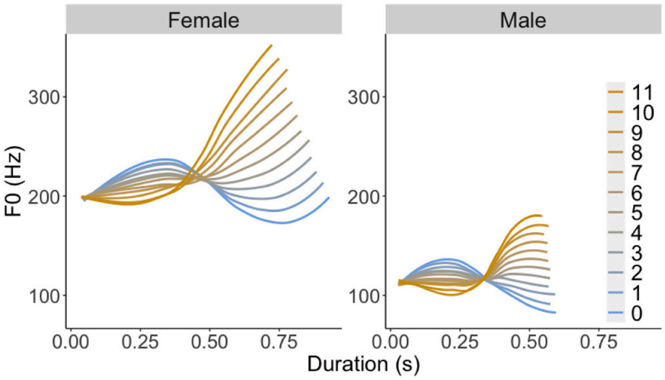

Fig. 12. Intonational contours for the test item "It's cooking" at Steps 0–11 for both the female and the male talker. Similar continua were created for each of the exposure items. The bottom and top lines at the right edge of the continua are Step 0 (statement) and Step 11 (question), respectively. Since Steps 0 and 1 resulted in identical responses during norming, Experiment 2 employs only Steps 1–11.

3.3. Procedure

Participants were assigned to one of the six between-participant conditions (female vs. male talkers x the three exposure conditions, see Fig. 10). In the pre-test, all participants heard 22 tokens of *It's cooking* sampled uniformly along the 11-step continuum (in a randomized order). They were asked to categorize them as a question or a statement by clicking either the "statement" or the "question" button (i.e., 2AFC). No feedback was provided.

During exposure, participants were assigned to one of the three conditions: *no bias*, *statement-biasing*, or *question-biasing* exposure (see Fig. 13). Exposure always consisted of 30 tokens (i.e., *It's* {*booting, cooling, losing, moving, muting*}, six tokens of each). The task was the same as during the pre-test, except that participants received feedback after each categorization response ("You were {correct/wrong}. It was a {statement/question}.", accompanied by two different types of ring tones). As such, the current task deviated from representative perceptual learning experiments in two aspects. First, the feedback was explicit rather than implicit.[13] Second, the feedback labeled the acoustic input with some delay, *after* participants listened to a stimulus and selected a response. This mode of feedback was chosen to provide clear category information for the prosodic input supporting the abstract meanings (i.e., a question vs. a statement). We anticipated that a possible effect of these changes, if any, would bias *against* the outcomes predicted by the hypothesis we investigate: delayed label information can impede learning of distributional statistics because ideal adaptation requires listeners to maintain information about the perceptual input (Burchill et al., 2018).

In the no bias condition, 15 exposure tokens with statement feedback were sampled from Step 1, and another 15 with question feedback from Step 11 (Fig. 13, middle panel). The statement- and the question-biasing conditions differed from each other in the feedback associated with the tokens that had been normed to be most ambiguous (i.e., Step 6 for the female talker, Step 7 for the male talker). In the question-biasing condition, participants heard 15 of the ambiguous tokens with question feedback, and another 15 from Step 1 with statement feedback. Finally, in the statement-biasing condition, participants heard 15 of the ambiguous items with statement feedback, and another 15 from Step 11 with question feedback.

During the post-test, participants provided responses to the same 22 tokens of *It's cooking* from the pre-test with no feedback. The bottom panels of Fig. 13 present predicted patterns of shifts in the categorization functions, which will evidence talker-specific adaptation of cue-category mappings. Eliciting responses to productions from the three *variants* of each talker, rather than using distinct talkers for the distinct exposure conditions, we rule out the possibility that post-test differences are driven solely by the talker's inferred gender, age, or other idiosyncratic (acoustic) properties; They can be attributable to the input experienced during the exposure phase.

[11] "It's moving" was chosen as a base item because of the continuing resonant sound of the sonorant (/m/) and the voiced fricative (/v/) made it easier to extract F0 values with generally fewer track losses compared to other items. Equating the vowel across the training items (i.e., the tense /u/ vowel) was necessary for creating natural sounding stimuli, where we superimposed the prosodic contour while preserving segmental information.

[12] The female talkers' production of Step 0 exhibits a slight rise at the end, which may be perceived as a so-called rise-fall-rise contour (as notated as L*+ H L-H% in Pierrehumbert and Steele, 1987). The same type of rise was not present in the male talker's production of Step 0. A rise-fall-rise contour has traditionally been associated with meanings such as incredulity (when associated with a wide pitch range) or uncertainty (when associated with a narrow pitch range) (Hirschberg and Ward, 1992). Depending on a context, it can convey a contrast between the uttered content and a contextually assumed content (Constant, 2012; de Marneffe and Tonhauser, 2019; Kurumada et al., 2014). It can also be a signature of a socio-indexical feature or speech style as known as uptalk (Warren, 2016). As can be seen in Figure 14, the stimuli tokens from Steps 1 and 2 in the female talker conditions reliably elicited a "statement" response and were clearly distinguished from the base token for the question category (i.e., Step 11). We therefore determined that the slight terminal rise in the female talkers' productions was part of the talker-specific realizations of the question vs, statement categories, which listeners must navigate as they categorize different stimuli tokens. However, it can safely be considered as orthogonal to the main effect of adaptation we investigate here.

[13] The most common perceptual recalibration paradigms for segmental speech perception have typically used lexical (Kraljic and Samuel, 2005; Norris et al., 2003) or visual labeling (Vroomen et al., 2007) that is delivered concurrently with the acoustic input. For instance, in Norris et al., (2003), a sound that was meant to be midway between /f/ and /s/ was included in a lexical carrier such as "witlo?". This context disambiguated the ambiguous sound as an /f/ for a Dutch listener because "witlof" is a word while "witlos" is not. Currently, to our knowledge, it remains an open question if the mode of feedback (e.g., implicit vs. explicit) and its timing would significantly affect the speed and amount of perceptual adaptation to phoneme or prosodic variability. We acknowledge that this needs to be further examined in future studies.

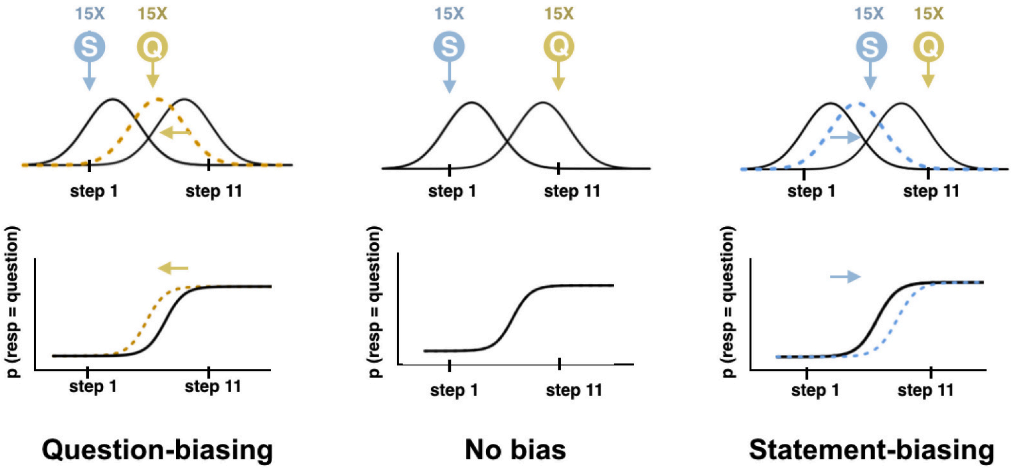

Fig. 13. The three exposure conditions of Experiment 2. Top: Placement of the 30 exposure tokens (circles) along the prosodic continuum from statement to question, depending on the exposure condition. Bottom: Qualitative predictions of the direction of shifts in the categorization functions.

3.4. Analysis approach and results

We analyze the data in two steps. First, we focus exclusively on pre-test responses. We fit ideal observer models trained on the production data from Experiment 1 to explore what implicit knowledge listeners may be applying during the initial encounter with an unfamiliar talker, prior to labeled exposure from that talker. Second, we compare pre-test and post-test responses. For both the female and the male talker, we test whether listeners' categorization functions shifted in the predicted directions across the three exposure conditions (shown in Fig. 13).

3.4.1. What information do participants draw on during pre-test?

Before we introduce the ideal observer analyses, we summarize participants' pre-test responses. Fig. 14A shows participants' pre-test responses for each talker, collapsing over the three exposure conditions (i.e., each line is based on approximately 180 participants). Replicating the norming study (see Appendix III), listeners' judgments shifted gradually from a statement to a question along the 11 step continua. The *category boundary*—the point along the continua that is ambiguous between a statement vs. a question—differed somewhat between the two talkers.

Although the similarity between the categorization functions for the female and male talker is expected given the norming, how such similarity can be achieved is worth exploring. Recall that the acoustic realizations of the 11 steps differ substantially between the two talkers (Fig. 11 and Fig. 12). That is, the items that listeners perceived to be similarly ambiguous (e.g., Steps 6 and 7) or unambiguous (e.g., Steps 1 and 11) for the female and the male talker occupied different areas in the un-normalized acoustic space (Fig. 14B). Pre-test categorization of the male talker's productions had to, and did, depend almost exclusively on F0, with a very narrow range along this cue dimension (the black line in Fig. 14B). This contrasts with the female talker's productions, for which listeners could rely on both cues; the tokens span widely along the dimension of duration as well as that of F0. Even prior to further analysis, this suggests that listeners must be mapping the categories onto phonetic cues in a manner malleable to such cross-talker variability.

To what extent do normalizations of cues get the listeners to achieve this perceptual constancy between the two talkers? To examine this, we compared how well different ideal observer models predict participants' responses during pre-test. As in Experiment 1, we contrast the predictions of talker-independent and talker-specific models, both under different assumptions about cue normalization. The talker-independent models are exactly the same as explained in Experiment 1, based on the marginal distributions across the 65 talkers. To test predictions of the talker-specific models, one additional step is necessary: since we are assessing listeners' recognition judgments *prior to* an exposure to a given talker's productions, no talker-specific model is yet available. The best they can do is to draw on models of talkers that are *similar,* and hence relevant, to the current talker, for instance, previously encountered talkers of the same gender.[14]

Here, we merely assume that listeners can recognize acoustic properties of speech correlated with gender (for a review, see Foulkes and Hay, 2015) and draw on predictions of talker-specific models aggregated over multiple talkers within each gender group. We thus constructed *gender-specific* models for female and male talkers by averaging means, variances and covariances of female and male talkers from Experiment 1, respectively. This approximates a "prototypical" female or male talker that listeners may have in mind based on their past experiences. If the gender-specific models predict human categorization better than the talker-independent models, it would suggest that listeners learn and store talker-specific, or at least gender-specific, cue distributions.

The models' recognition judgments are derived on the eleven test tokens from the female and male talker (Fig. 15). Note that the talker-independent models, as implemented here, have relatively wide distributions encompassing distributions of all talkers in the corpus (Fig. 15A, the identical model was used for the female and male talkers' test tokens). The gender-specific models, in contrast, have generally narrower distributions because each of them represents an *average* of taker-specific models of a given gender (Fig. 15B). This is particularly evident in the male talker model, predicting a sharper categorization curve along the dimension of F0 as compared to the other models. These differences have important consequences in their fit to the human data, which we detail below.

Fig. 16 shows the correlation between the models' predicted posterior probabilities for the question category (x-axis) and listeners' categorization responses during pre-test (y-axis). All ideal observer models exhibit positive correlations with listeners' categorizations. In particular, models with talker-normalized cues, as opposed to un-normalized or utterance-normalized cues, fit human judgments well ($R^2 > 0.9$). Considered in conjunction with the results of Experiment 1, talker-normalized cues were expected to facilitate accurate categorization, and listeners do indeed seem to draw on talker-normalized, rather than un-normalized, cues. Further, the gender-specific models (Fig. 16, bottom row) consistently outperform the talker-independent models regardless of normalization. The gender-specific model over talker-

[14] What exactly constitutes "similar" is a question beyond the scope of the present work (for relevant discussion, see Johnson, 2006; Witteman et al., 2013; Xie and Myers, 2017).

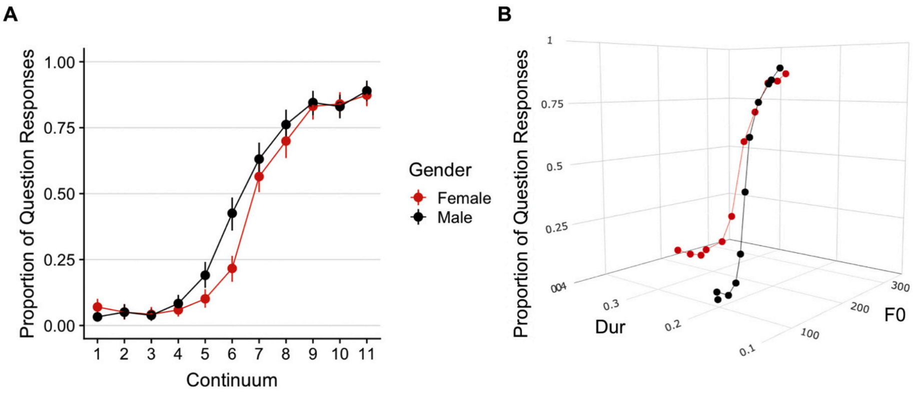

Fig. 14. Panel A: pre-test responses for the female and male talkers as a function of continuum steps. Panel B: Same pre-test responses plotted in (un-normalized) F0-Duration acoustic space.

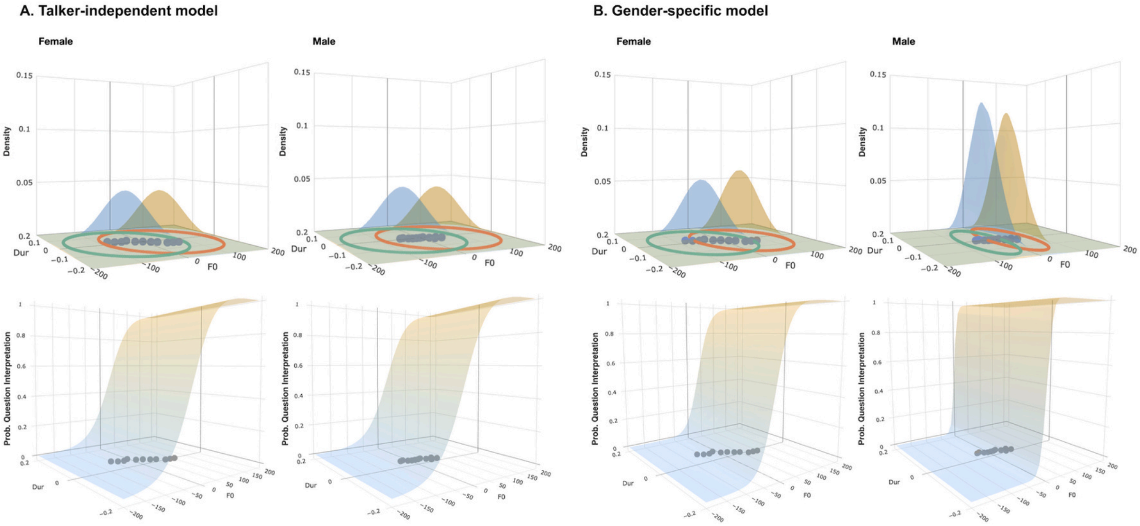

Fig. 15. Panel A: modeling results from the talker-independent model (identical for the female and male talkers); Panel B: modeling results from the gender-specific models. In each panel, estimated cue density functions (top) and predicted categorization (bottom). Utterance- and talker-normalized talker-specific distributions are used.

normalized cues (Fig. 16, bottom right panel) achieves the highest correlation with listeners' actual categorization responses ($R^2 = 0.95$). This goodness-of-fit is remarkable given that the ideal observers have 0 degrees of freedom—they are based on data from different talkers from Experiment 1 and yet predict how listeners interpret the input from the two unfamiliar talkers in Experiment 2.

What is also notable in Fig. 16 is that the ideal observers are *not* good *linear* predictors of listeners' responses. There are at least two reasons why this non-linearity is expected, both originating in simplifying assumptions of the ideal observers we have used so far. First, the ideal observers we use here make the simplifying assumption of uniform prior probabilities of questions and statements (see Experiment 1). This assumption is clearly wrong: statements are in general observed more frequently than questions (Boakye et al., 2009), and this asymmetry is likely even more pronounced for utterances with declarative syntax, as used in all stimuli for Experiment 2. Since listeners' responses are expected to be sensitive to this information, we would expect listeners' categorization function to be shifted towards the question category (rightward along the 11-step continuum) relative to the ideal observer.[15]

Second, unlike ideal observers, human listeners exhibit attentional lapses. On trials with such lapses, listeners do not have access to the acoustic input and thus can only rely on their belief about the prior probability of questions vs. statements or outright guess. The existence of lapses means that listeners' categorization functions do not converge on 0 or 1. This pattern is clearly visible in Fig. 14: on the left end of the continuum, listeners converged on 5.2% question responses; on the right end of the continuum they converged on 88.2% question responses. This suggests a lapse rate of 17.0% and a response bias on those lapse trials of 31% towards question responses.

By integrating these estimates, we can better evaluate ideal observers' predictions against human judgments. The updated predictions for the talker-independent and the gender-specific models are shown in Fig. 17, plotted against human responses. Specifically, we zoom in on the predictions for both un-normalized cues and utterance- and talker-normalized cues, as the best and the worst models of human judgments.

Considering these patterns in combination with the correlation results (Fig. 16), we can make two observations. First, models with normalized cues (Fig. 17; right column) better capture the pattern that listeners provided more question responses for the male talker than for the female talker. The talker-independent model with un-normalized cues (Fig. 17; top left panel) erroneously categorized most tokens from the male talker as statements due to their low F0 values. This was

[15] This shift would be identical across all ideal observers and would not affect the slope of the categorization function. It thus cannot explain the correlation results.

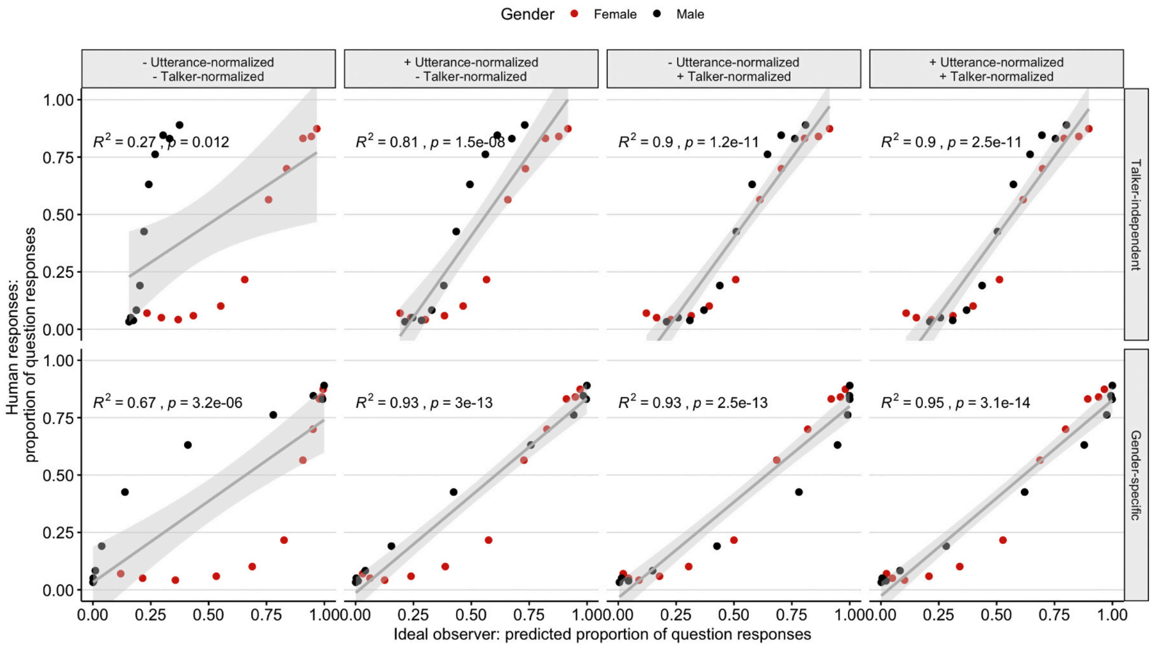

Fig. 16. Correlations between the ideal observer-predicted posterior probabilities for the question category (x-axis) and listeners' categorization during pre-test (y-axis). Top: talker independent models. Bottom: gender-specific models, derived by averaging over talker-specific models of each gender. Columns correspond to the different ways of normalizing (or not) utterance-final F0 and duration.

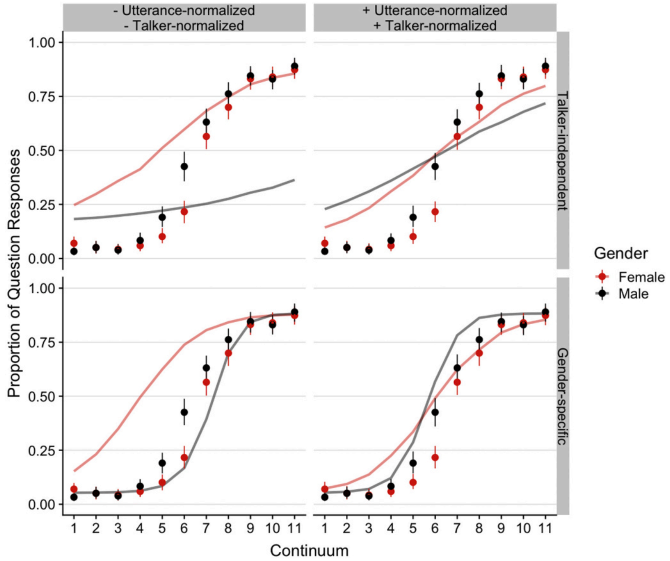

Fig. 17. Categorization functions predicted by ideal observers (lines) and actual categorization by listeners (pointranges). Top: talker independent models (un-normalized or both utterance- and talker-normalized). Bottom: gender-specific models (un-normalized or both utterance- and talker-normalized).

alleviated in the gender-specific model (Fig. 17, bottom left panel) due to the knowledge of the distinct baselines for the phonetic cues between males and females. Nonetheless, this model still fails to predict the relative ordering of the category boundaries between the male and the female conditions.

Second, the gender-specific models (Fig. 17, bottom row) better fit the steepness of the human categorization functions than the talker-specific models (Fig. 17, top row). One major reason for this is that, as mentioned above, the gender-specific models have smaller estimates of category variances than the talker-independent models (recall Fig. 15). The "narrower" variance estimates generally mean a relatively small amount of overlap between the underlying categories, which predicts a steeper categorization function (for related discussion, see Clayards et al., 2008; Theodore and Monto, 2019). The good fit of the gender-specific models to the human data may suggest that human listeners have the knowledge of gender-specific category variances as well as category means.[16]

In summary, the analyses of the pre-test responses supported our key predictions derived in Experiment 1. That is, human listeners process the prosodic input relative to a given talker's baseline, which helps them overcome the variability in raw cue spaces (e.g., between a male and a female talker). This resonates with results from segmental speech perception that phonetic cues are interpreted relative to talker-normalized expectations (McMurray and Jongman, 2011). Going beyond normalization, the present results further suggest that listeners begin processing the input by drawing on their knowledge of phonetic cue distributions experienced with talkers similar to the current (unfamiliar) talker.

3.4.2. How do listeners' categorizations change from pre- to post-test?

Next, we analyze changes in listeners' categorization function from pre- to post-test. If listeners learn talker-specific distributional statistics of prosodic categories, their post-test categorization boundaries should change to include more question responses after question-biasing exposure, and fewer question responses after statement-biasing exposure. Post-test categorization functions for the no bias condition should fall between the categorization boundaries of the statement and question-biasing conditions. To ensure that listeners in the question-biasing and the statement-biasing condition did indeed learn to categorize ambiguous items according to the feedback given during the

[16] There are other potential differences between the model predictions and human categorization: e.g., listeners likely employ more cues than the two cues considered here; and the model's assumption of Gaussian categories might be wrong (cf. exemplar or episodic approaches, which avoid this assumption, (Foulkes and Hay, 2015; Goldinger, 1996; Hawkins, 2003; Johnson, 2005; Pierrehumbert, 2001).

exposure phase, we have analyzed their responses to the 30 exposure tokens (Appendix IV). Participants seem to have latched onto the association between the prosodic pattern in the ambiguous tokens and their intended category affiliations from the very outset of the exposure phase. The learning seems to have continued over multiple instances of the ambiguous input, and participants generally reached a stable pattern of categorization judgments by the time they completed about 50% of the exposure trials.

Responses in the pre- and post-exposure tests are plotted in Fig. 18. As predicted, question-biasing exposure shifted the boundary between questions and statements leftwards from pre- to post-test, expanding the question category and leading to more question responses. Statement-biasing exposure did the opposite. Little to no difference between pre- and post-test was observed for no bias exposure. These differences between the exposure conditions were significant, as confirmed by mixed-effects logistic regression (see Table 1). The analysis included exposure condition (sliding difference-coded with two orthogonal contrasts comparing statement-biasing < no bias < question-biasing), talker (sum-coded, female = 1 vs. male = −1), continuum (steps 1–11, mean-centered, coded as a continuous variable), test block (pre vs post-test, sum-coded), and the full factorial interactions between exposure condition, talker, and test block as fixed effects. The analysis included the maximum random effect structure justified by the data (by-participant intercepts and slopes for Block and Continuum as well as their interaction).

We found main effects of talker ($\widehat{\beta} = -0.54$, $p < .0001$), test block ($\widehat{\beta} = 0.013$, $p = .023$), exposure condition (question-biasing vs. no bias, $\widehat{\beta} = 0.58$, $p = .004$) and continuum ($\widehat{\beta} = 1.55$, $p < .0001$). The main effect of talker as well as the two-way interaction between test block and talker ($\widehat{\beta} = -0.18$, $p < .001$) suggest that there was an overall bias for the listeners to provide a question response to the male talker's productions as opposed to the female talker's productions, especially after the exposure. This replicates the pattern we saw in our analysis of the pre-test data and what we showed to be predicted by ideal observers.

Critically, the predicted interaction term between test block and exposure condition was significant (no bias vs. statement-biasing: $\widehat{\beta} = 0.41$, $p = .0002$; question-biasing vs. no bias: $\widehat{\beta} = 0.66$, $p < .0001$): participants provided distinct categorization judgments reflecting the distributional patterns of the input and feedback given in the exposure phase. The absence of the higher order interactions including talker indicates that the amount of adaptation induced by the exposure input did not differ significantly across the two talker conditions.

One possible confound may be that participants across the three exposure conditions could be differentially biased to provide a question response.[17] Because participants were randomly assigned to the three conditions *after* they had provided their pre-test responses, no such difference was *a priori* expected. However, it is possible that participants in one condition happened to be more likely than others to provide a question (vs. a statement) response to ambiguous tokens. If there had been such a baseline difference, simple regression towards the mean in their patterns of responses would have resulted in the shifts of the categorization functions. A simple effect analysis, however, supported neither a significant difference between the exposure conditions in the pre-test block, nor an interaction between exposure conditions and talkers in the pre-test block. Therefore, we concluded that the participants' pre-test responses did *not* differ significantly across the three exposure conditions prior to the exposure.

3.5. Discussion

Despite the notable differences in the un-normalized cue distributions across the two exposure talkers, listeners effectively adapted to each talker's baseline. Adopting gender-specific distributional knowledge improved correlations between the model predictions and human categorization responses even for normalized cues (from $R^2 = 0.9$ to $R^2 = 0.95$). This might seem like a small benefit, especially since the talker-normalized, talker-independent model already achieves a reasonable correlation. However, evaluating the model predictions along the 11 step continua (Fig. 17), we found that the gender-specific models better capture the qualitive, in addition to the quantitative, patterns of the human categorization judgments.

We also found that labeled exposure to just 30 question and statement tokens was sufficient to shift listeners' categorization boundaries for subsequent input from the previously unfamiliar talker. In particular, between the pre- and post-tests, the recognition judgments almost completely reversed for tokens that were originally close to category boundary (similar to perceptual recalibration in segmental speech perception, e.g., Kraljic & Samuel, 2006; Norris et al., 2003). For instance, in the post-test, Step 7 in the female talker condition elicited the question response over 70% of the time in the question-biasing condition but only 28% of the time in the statement-biasing condition. These fast and robust shifts lend support to the feasibility of our central proposal: learning of talker-specific distributions allows listeners to navigate between-talker variability.

4. General discussion

The mapping between phonetic cues and prosodic categories is often variable and fluid across talkers (Arvaniti, 2019; Arvaniti and Garding, 2007; Cangemi et al., 2015; Cangemi and Grice, 2016; Clopper and Smiljanic, 2011). A major question that these observations raise is how listeners arrive at the mapping *intended for a given token of input.* Due to the variability, acoustic-phonetic properties of the input themselves can get them only halfway. We hypothesized that listeners may draw on their implicit knowledge of *structure* of the variability. In particular, the series of experiments presented here focused on how acoustic cue distributions vary across talkers and whether listeners can benefit from learning these talker-specific distributions.

We began the exploration with collecting large-scale production data. Adopting principles of ideal observer analyses (Kleinschmidt, 2019; Kleinschmidt and Jaeger, 2015), we quantified the amount of information listeners can gain from learning talker-specific distributions of phonetic cues over the question vs. statement categories. Put simply, if everybody's production is equivocal after normalizing their baseline differences, no additional information could be learned. Data revealed that talkers vary substantially in category means and variances in addition to their baselines (e.g., cue means), which suggests that talker-variability in prosody carries additional information for accurate category recognition. The feasibility of such learning was demonstrated in a comprehension experiment, where listeners adapted their categorization boundaries in response to patterns in the exposure input.

4.1. Does talker-specific information always help?

The results of the experiments present a proof of concept that learning talker-specific distributional statistics affords listeners the means to navigate variability in prosodic productions. At first glance, this may seem a foregone conclusion. Provided sufficient (representative) training data, *any* additional knowledge specific to a given talker's speech may be expected to improve categorization accuracy. This, however, is not necessarily the case.

Numerically, the talker-specific models reported in 2.3.2 had higher or equal accuracy than the talker-independent models *for all 65 talkers*. However, the degree to which a talker-specific model outperformed a

[17] This post-hoc analysis was conducted in response to an anonymous reviewers' observation that the pre-test responses in Figure 18 seem to suggest some *a priori* biases distinct across the exposure conditions.

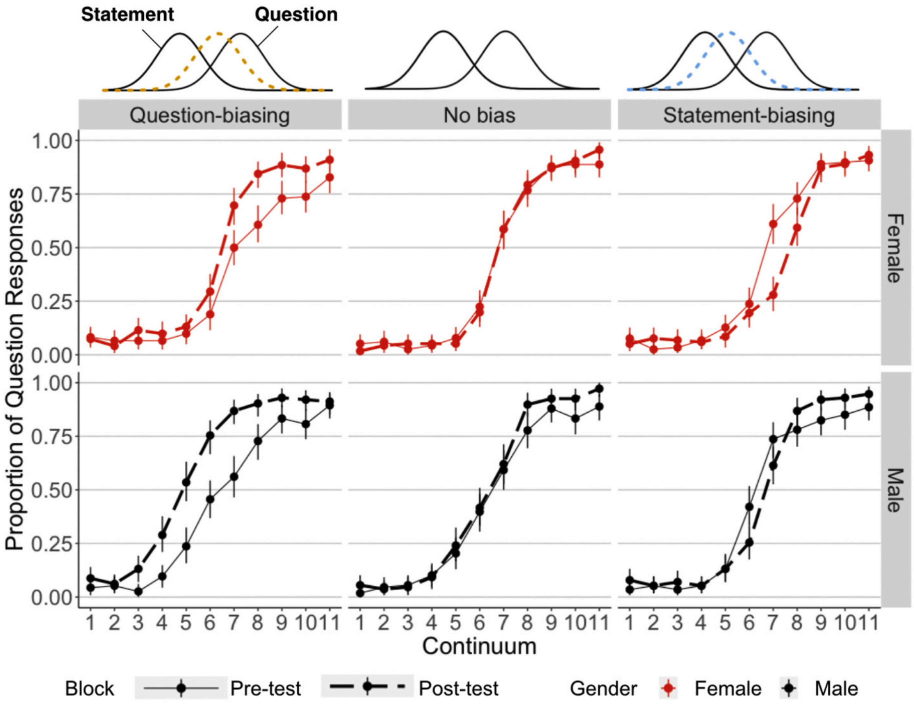

Fig. 18. Proportions of question responses during pre- and post-test by exposure condition (question-basing, no bias, and statement-biasing) in Experiment 2. Error bars indicate bootstrapped 95% CIs. Above the results, we show the qualitative predictions based on the hypothesis that listeners learn the talker-specific realization of prosodic categories during exposure.

Table 1
Summary of mixed-effects logistic regression with test block (pre- vs post-test, sum coded with post-test as 1 and pre-test as −1), exposure condition (sliding difference-coded comparing no bias vs. statement-biasing and question-biasing vs. no bias), their interactions, and continuum (centered) as fixed effects. Significant effects ($p < .05$) are bolded.

	Estimate $\widehat{\beta}$	S.E. $(\widehat{\beta})$	p value
Intercept	**−0.933**	**0.088**	**< 0.0001**
Talker (Female vs Male)	**−0.543**	**0.082**	**< 0.0001**
Block (Post vs Pre)	**0.134**	**0.059**	**0.023**
No bias (vs. Statement-biasing)	0.268	0.200	0.181
Question-biasing (vs. No bias)	**0.575**	**0.201**	**0.004**
Continuum	**1.545**	**0.061**	**< 0.0001**
Talker * Block	**−0.176**	**0.053**	**< 0.001**
Talker * No bias (vs. Statement-biasing)	0.054	0.200	0.785
Talker * Question-biasing (vs. No bias)	−0.342	0.201	0.088
Block * No bias (vs. Statement-biasing)	**0.408**	**0.132**	**0.002**
Block * Question-biasing (vs. No bias)	**0.656**	**0.133**	**< 0.0001**
Talker * Block * No bias (vs. Statement-biasing)	0.129	0.131	0.325
Talker * Block * Question-biasing (vs. No bias)	−0.143	0.131	0.276

talker-independent model differed across talkers. Considering the cases where talker-specific information did *not* majorly improve categorization provides important insight into the current discussion. Here we discuss three representative cases.

First, talker-specific models did not outperform talker-independent models when a given talker's category means were similar to the marginal means after appropriate normalization is applied. In these cases, model predictions based on the marginal distributions were close enough to distinguish the categories in a given talker. In contrast, talker-specific models fit the data better than marginal models when means of talker-specific distributions were either farther away from each other or closer together than what would be expected from marginal distributions. For instance, the categories produced by talkers like Talker (a) and (b) in Fig. 3 were closer to each other compared to those by (c) and (d) (see also bivariate Gaussian distributions of the corresponding talker-specific models in Fig. 9). Without the talker-specific knowledge, many of the tokens produced by Talkers (a) and (b) would not be reliably distinguished as a question or a statement.

Second, talker-specific models did not significantly deviate from marginal models when the structure of the underlying categories was more or less symmetrical (e.g., Talker (d) in Fig. 3 and Fig. 9). Put differently, talker-specific models performed better than their corresponding talker-independent models when the categories had distinct variances, especially along the dimension of F0. This was in part due to the nature of the question vs. statement categories, where the contrast was encoded primarily through the utterance final rise vs. fall of F0. When the two categories were similar in variance along F0, the category boundary could be extrapolated straightforwardly from the talker's mean F0. In this scenario, a talker-specific model and a talker-independent model with talker-normalized cues would be largely indistinguishable.

Finally, even when productions do not exhibit either of the aforementioned properties, benefits of talker-specific models were limited when either (or both) category had a large variance. These are talkers (e.g., Talker (f) in Fig. 3) whose productions are internally less consistent and variable compared to others in terms of how the two categories are encoded via F0 and/or duration. In these cases, even with talker-specific distributional knowledge, accuracies of categorization are bound to be limited.

The finding that talker-specific models overall outperformed talker-independent models was therefore informative about the *data*, not a necessary consequence of how the models were trained. It follows that, empirically, the distributional structures of the prosodic categories examined here varied across talkers in a manner such that learning of these structures was possible and overall beneficial for listeners. This validation was important for many reasons. Chief among them was to motivate the investigation of talker-specific learning of F0 and duration as a solution to the lack of invariance between questions and statements. Past research on talker-specific perception/comprehension of prosody presupposed cross-talker variability and endorsed learning as a solution (Kurumada et al., 2017; Nakamura et al., 2019; Roettger and Franke, 2019; Roettger and Rimland, 2020). The current study is the first to empirically show that a structure of talker-variability over the two prosodic categories is in fact worth learning.

The current ideal observer approach thus extends existing approaches to experience-based prosodic comprehension (Cangemi et al., 2015; Hawkins, 2003; J. B. Pierrehumbert, 2003; Saindon et al., 2017; Schweitzer, 2012; Smith and Hawkins, 2012). Beyond showing that learning talker-specificity *generally* benefits recognition, the current approach can quantitatively predict the amount of benefit. Further, the same logic can be used to estimate talker-level effects. For example, we can identify individual talkers for whom talker-specific learning likely leads to larger or smaller benefits. Evaluating these predictions against human judgments at the level of individual talkers, however, requires considerations of a wide range of variables currently unknown (e.g., cues other than F0 or duration that impact perception and interpretation, listeners' cue weighting strategies), and hence awaits future research.

4.2. Evidence of distributional learning?

The comprehension experiment in Experiment 2 extends the so-called perceptual learning paradigm. While this has been used widely to examine malleability of linguistic categorization of the perceptual input, the exact nature of learning is still under contention (Xie et al., 2016). Here we interpret the incorporation of *ambiguous* tokens into a recalibrated category as evidence of implicit learning of acoustic cue distributions. We consider this learning to be happening at the *prosodic,* rather than phonetic, level because exposure and test tokens never contained identical lexical or segmental information and hence are distinct in their exact acoustic cue values. The categorization boundary shift would not be expected if listeners were simply storing memory traces of phonetic cues and feedback (i.e., category labels) received during exposure.

However, there are at least two classes of alternative accounts for the observed patterns of data. One is that, instead of engaging with distributional learning, listeners may be simply learning to answer "question" (or "statement") when they hear a prosodically ambiguous token. In other words, they simply classify test tokens either as ambiguous or unambiguous, and respond to all ambiguous test tokens with the same category label given to ambiguous exposure tokens. This is not particularly plausible because listeners, at least in some conditions (e.g., the question-biasing condition with the male talker), shifted their judgments even for items close to the end points of the continua. However, the current data does not decisively reject this alternative account. A more stringent test for the distributional-learning hypothesis should therefore ask how much category-internal *structure* listeners can detect and learn. For instance, if listeners do indeed learn and store a distributional structure of a prosodic category, they might respond differently to prototypical vs. deviant category members (Xie et al., 2016). Further, listeners should also be attuned not only to shifts of category means but also to their variances (Clayards et al., 2008; Theodore and Monto, 2019).

Another possibility is that listeners may have used the exposure input to adjust their expected F0 range for a given talker. In the statement-biasing condition, for example, ambiguous tokens labeled as statements serve as the "bottom-end" of a recalibrated pitch range for a talker. Subsequently, all the post-test tokens are evaluated against this talker-normalized F0 range, which increases statement responses for intermediate steps (e.g., Steps 5–8). (Similar narrowing can happen in the opposite direction in the question-biasing condition.) Although still talker-specific, this type of normalization may not involve learning of distributions for each category. To gain further insights, future studies should manipulate an overall acoustic cue mean (e.g., a given talker's mean F0) and category means independently.

Additionally, it is possible that listeners combine normalization and distributional-learning and/or concurrently apply them for different purposes. In recent work, Lehet and Holt (2020) demonstrated that listeners do quickly learn distributional statistics of durational values informative about vowel categories in an artificial accent (extending Liu and Holt (2015)). Phonetic duration of these vowels, however, exerted a consistent influence on categorization of a subsequent consonant, an effect often associated with normalization. Lehet and Holt concluded that normalization and distributional learning can happen simultaneously to best navigate variability present at multiple levels of processing hierarchy.

We, too, have postulated that normalization and distributional learning are *not* mutually exclusive and that learning may or may not operate over normalized cues. Moreover, relative importance or usefulness of these two mechanisms will likely change across different cues and categories as well as across varying training regimes (e.g., amount of exposure and task difficulties). The proposed analysis framework – predicting comprehension from the distributional information in production – provides tools to probe when, in principle, distributional learning will be useful above and beyond what can be achieved via normalizations (Kleinschmidt, 2019). Before closing, we now sketch out how this framework may shed new light on the long-standing puzzle of form-meaning mapping in the phonological knowledge of speech prosody.

4.3. Implications for intonational speech prosody

The results of this study speak to a pertinent question: how do listeners maintain prosodic invariance under changes of perceptual detail (Arvaniti, 2019; Dilley, 2007; Grice et al., 2017; Gussenhoven, 1999; Ladd, D Robert, 2008; Ladd and Morton, 1997; Liberman and Pierrehumbert, 1984)? While F0, the primary phonetic property of intonational prosody, is continuous and variable across and within talkers, listeners can reliably recognize prosodic categories (for reviews see Dahan, 2015; Ward, 2019). At the same time, listeners are demonstratively sensitive to category-internal, phonetic variability to extract gradient meaning. For instance, relative height of a pitch peak and its temporal alignment with a stressed syllable can independently serve as a cue to degrees of intonational emphasis (Ladd and Morton, 1997), which in turn signal the relative significance of the message (Gussenhoven, 1984). A major theoretical challenge has thus been the need to account for the discreteness and gradience of the form-meaning mapping in a coherent framework (Arvaniti, 2019; Grice et al., 2017).

At its core, this is a question about how the comprehension system balances flexibility with the stability that is necessary for communication of meaning across contexts. The framework we explored with ideal observers allows us to begin answering this question through the lens of "inference under uncertainty" (Kleinschmidt and Jaeger, 2015). As discussed in Section 2.3, this framework (among many others) assumes that listeners do *not* directly map the acoustic input (e.g., F0) onto a stored, meaning-bearing, representation. Rather, they *infer* such a mapping. That is, they "combine what they know about how speech is generated in order to recover (or infer) the most likely explanation for the speech input they hear" (Kleinschmidt, 2019, p. 44). Across talkers, phonetic cues for each possible "explanation" (i.e., prosodic categories such as boundary tones and pitch accents) can change. Listeners therefore need to inferentially arrive at an explanation that is *most likely given relevant previous experiences*.

To take a concrete example, let us consider productions of question vs. statement prosody by adults and young children. Patel and Grigos (2006) reported that four-year-olds, compared to older children and adults, rely more on syllable duration rather than F0 to signal the question vs. statement contrast (Fig. 19). Their productions of questions are in fact often mistaken as statements by adult listeners who are not used to the speech pattern (Patel and Brayton, 2009). The analysis framework proposed in the current study makes the prediction that

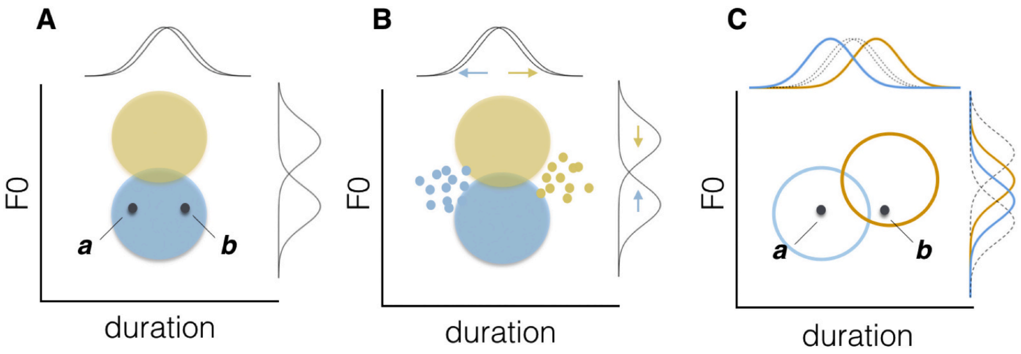

Fig. 19. Distributions of phonetic cues and density functions associated with question vs. statement categories. Panel A: marginal distributions expected from adult native speakers' productions; Panel B: adaptation in response to increased input from young children; Panel C: hypothesized data from 4-year-olds based on the observation in Patel and Grigos (2006).

adult listeners begin with their general expectations derived from marginal distributions (Fig. 19A). With increasing input from young children (Fig. 19B), expectations for underlying distributions may "adapt" to the patterns as the means and variances shift over time (Fig. 19C). Consequently, tokens that would initially be heard as part of the statement category (Tokens *a* and *b*) will be more reliably mapped onto the statement and the question categories, respectively. A simple, intuitive prediction is that parents and other caretakers of young children, compared to others, can deploy distributional representations more attuned to young children's productions. This will make it easier for them to distinguish (otherwise ambiguous) questions and statements from those children.

Although we are not aware of any direct, quantitative test of this specific hypothesis, existing evidence of short-term, talker-specific adjustments of prosodic perception (Baese-Berk et al., 2014; Patel and Schroeder, 2007; Saindon et al., 2017) as well as effects of native language on second language prosody categorization (Liu and Rodriguez, 2012) provide indirect support for the idea that prior experiences systematically govern cue-category mapping inferred in a context. The paradigm used in Experiment 2 can be extended to directly test whether and how distributional statistics associated with talkers (or talker groups) can predict distinct outcomes of inferences.

In summary, learning to represent structure of phonetic cues in the input allows listeners to maintain stable phonological categories despite their variable phonetic realizations. This dovetails nicely with a now-classic observation that prosodic contrast is "categorically interpreted but not categorically perceived" (Ladd and Morton, 1997). In other words, a categorical contrast (e.g., question vs. statement, focus vs. non-focus, stressed vs. unstressed) can emerge as an outcome of inference over the gradient input, guided by implicit knowledge of underlying distributional statistics. Stored details of prosodic input, and their gradient differences, can then be used to process gradience in meaning, as well as emotive and social meaning in the speech code. The ideal observer framework can be extended to model such a mapping between phonetic cues and gradient meaning (instead of a binary choice probability as we have examined here). Computational detail for this extension, however, remains to be investigated in future work.

4.4. Limitations and questions for future studies

A clear limitation of the current data stems from the highly restricted number and types of utterances used in the experiments. Indeed, the production data in Experiment 1 had only one construction ("It's X-ing") and the training and test items in Experiment 2 had the identical structure. The current stimuli therefore offer only a limited amount of information about an extent to which the effects of talker-specific learning are generalizable across different lexical items and sentence structures. An anonymous reviewer also pointed out that the circumscribed method we used to collect the production data in Experiment 1 (e.g., monologue, repetitive production of a single construction, no feedback) was likely to reduce overall variability in production and potentially promoting entrenchment of intonation contours across items. If true, that means that the current set of production data would underestimate the amount of within-talker variability compared to *true* underlying distributions to be observed in a more naturalistic form of language use. Accurately estimating distributional statistics that constrain the listener's prior knowledge and learning is by no means trivial both theoretically and methodologically (Kleinschmidt, 2019). We plan to address this question in our future experiments by systematically examining relevant parameters (e.g., types and tokens of sentences, discourse and communicative contexts, presence/absence of interlocutors) and their impacts on the structure of production variability.

Likewise, the comprehension experiment used between-participant design, where the exposure input and the pre−/post-exposure tests were produced by a single talker. It is therefore unclear if the learning was indeed conditioned on the talker or on the general task environment. Future studies must expand the scope of inquiry by covering a broader range of linguistic constructions produced by multiple talkers under different situations (e.g., reading vs. speaking, casual vs. formal speech, monologue vs. dialogue) to elucidate how variability can be ascribed to possible contextual sources.

One important question for future computational research concerns how to integrate the notion of a "talker" in a model. In the current study, we considered three different possibilities: talker-independent models with no notion of a talker identity, gender-specific models with the notion of a prototypical talker of a given gender, and strictly talker-specific models. As we discussed in 3.5.1, talker-independent models are bound to have wider category variances than gender- or talker-specific models as we implemented here. These models, therefore, can systematically overestimate the category variability when fit to tokens from one talker, inflating the relative benefit of talker-specific models. How we model a talker, in fact, relates to a deeper theoretical question about memory representations used in human speech perception. Do listeners model each talker's productions anew and store them separately? Or, alternatively, do they aggregate their experiences over multiple talkers? If so, do they estimate means and variances for a "prototypical" talker (for relevant discussions, see Kleinschmidt and Jaeger, 2015, 2016)? Future work should thus delve deeper into how human cognitive architecture may strike a balance between storing details of individual talkers' productions and compressing information to achieve efficient memory representations (e.g., for a recent review, see Bates and Jacobs, 2020).

Finally, the computational approach used here (an ideal-observer) is a normative model, and as such it makes certain assumptions that are not directly applicable to modeling human listeners. For instance, it has unlimited memory resources to accurately represent distributional information in the data. In addition, our models and other existing approaches (e.g., McMurray and Jongman, 2011) make a simplifying assumption that the talker-specific information is known and available to listeners. In reality, the information needs to be gradually extracted from utterances of varying length and contents. It is left for future work to implement modeling and behavioral testing approaches that can consider cumulative changes of distributional knowledge over time (Kleinschmidt and Jaeger, 2015, 2016; Theodore and Monto, 2019).

5. Conclusion

We began the current investigation with a puzzle: how can listeners reliably interpret meaning from the prosodic signal when stable mappings between the raw acoustic signal and prosodic categories appear to be lacking? There is now broad agreement that expectations based on implicit knowledge about categories' distributions of cues are critical to speech perception (Clayards et al., 2008; Feldman et al., 2009; Johnson, 2005b; Norris et al., 2015; Pierrehumbert, 2001; Pisoni and Luce, 1987; Smith and Hawkins, 2012). But this leaves open whether, and how, these expectations are contingent on the talker. Research on segmental speech perception has found that listeners seem to draw on expectations *relative to the present context,* including talkers (Foulkes and Hay, 2015; McMurray and Jongman, 2011; Nygaard and Pisoni, 1998). Similar findings have begun to emerge for supra-segmental speech perception, including intonational speech prosody (Cangemi et al., 2015; Kurumada et al., 2017; Nakamura et al., 2019; Roettger and Franke, 2019; Roettger and Rimland, 2020).

The present findings extended this insight. Only when cues were normalized relative to the talker's cue distribution do distributional models provide a good fit against the listener's interpretation of prosodic inputs. This entails that listeners learn and store information about talker-specific cue distributions—to subtract out (normalize) a talker's mean utterance-final F0 and syllable duration, listeners need to first *learn* this information through exposure. Indeed, the models that best describe listeners' interpretation of prosody go one step further: the best-fitting models had access to not only talkers' *overall* cue distribution (across the two categories), but also the talker-specific, category-specific distribution of cues. If future work replicates this pattern, it would suggest that the representations underlying human speech perception consist of many different talker- and/or group-specific models—as proposed in exemplar (Foulkes and Hay, 2015; Hawkins, 2003; Johnson, 2006; Pierrehumbert, 2001), episodic (Goldinger, 1996b, 1998; Sumner and Samuel, 2009), and Bayesian theories of speech perception (Kleinschmidt and Jaeger, 2015; Kronrod et al., 2016; Norris et al., 2015; Theodore and Monto, 2019).

Authors' contributions

Andrés Buxó-Lugo (ABL) and Chigusa Kurumada (CK) designed the behavioral experiments. ABL and CK created the stimuli. ABL implemented the study, collected, and pre-processed the data. Xin Xie (XX) and CK designed and interpreted the ideal observer analyses for Experiments 1 and 2. XX trained and tested the ideal observers. XX conducted all analyses and data visualizations. All authors contributed to theory development and writing.

Acknowledgements

The work presented here was partially supported by institutional research funds (University of Rochester) to CK. We thank T. Florian. Jaeger for theoretical discussions and consultations on quantitative analyses of the data presented here. We also thank Crystal Lee, Timo B. Roettger, Carol Fowler, Maryann Su Lin Tan, Shawn Cummings, Lily Steiger for helpful discussions and feedback during the preparation of this manuscript. Additionally, we thank Dr. James McQueen and an anonymous reviewer for their helpful comments and feedback. An earlier version of the manuscript was presented at the 59th Annual meeting of Psychonomic Society and the 32nd and 33rd meetings of the Annual CUNY Human Sentence Processing Conference.

Appendix I Sentences produced by participants in Experiment 1, once as a statement and once as a question

1. It's closing.
2. It's ending.
3. It's falling.
4. It's freezing.
5. It's loading.
6. It's melting.
7. It's raining.
8. It's snowing.
9. It's starting.
10. It's working.
11. It's ticking.
12. It's printing.
13. It's running.
14. It's coming.
15. It's reading.
16. It's writing.
17. It's moving.
18. It's learning.
19. It's changing.
20. It's stopping.
21. It's walking.
22. It's sinking.
23. It's crashing.
24. It's cooking.

Appendix II A comparison of two variants of talker-independent and talker-specific models (Experiment 1)

We constructed a new set of talker-independent models by training them on data randomly subsampled from the training pool so that the number of training tokens matches those used for talker-specific models. Due to this random sampling, a different training model was constructed for each test talker, resulting in 65 talker-independent models. As shown Fig. A1 (Panel A), performances of such models were more or less equivalent to those of the talker-independent models reported in Section 2.3.2 and significantly poorer than those of the talker-specific models, regardless of the type of cues used.

To further examine whether the better performance of the talker-specific models was indeed due to talker-specificity, instead of more constrained training data (from individual talkers), we tested the trained talker-specific models on test data randomly sampled from test tokens across talkers (the number of test tokens matched those used for talker-specific models). As shown in Fig. A1 (Panel B), such models did not evidence any significant benefits over the talker-independent models. This yields additional support to the conclusion that the higher predicted accuracies associated with talker-specific models stemmed primarily from the congruency between the training data and the test data (i.e., generated by the same, talker-specific, underlying prosodic cue distributions).

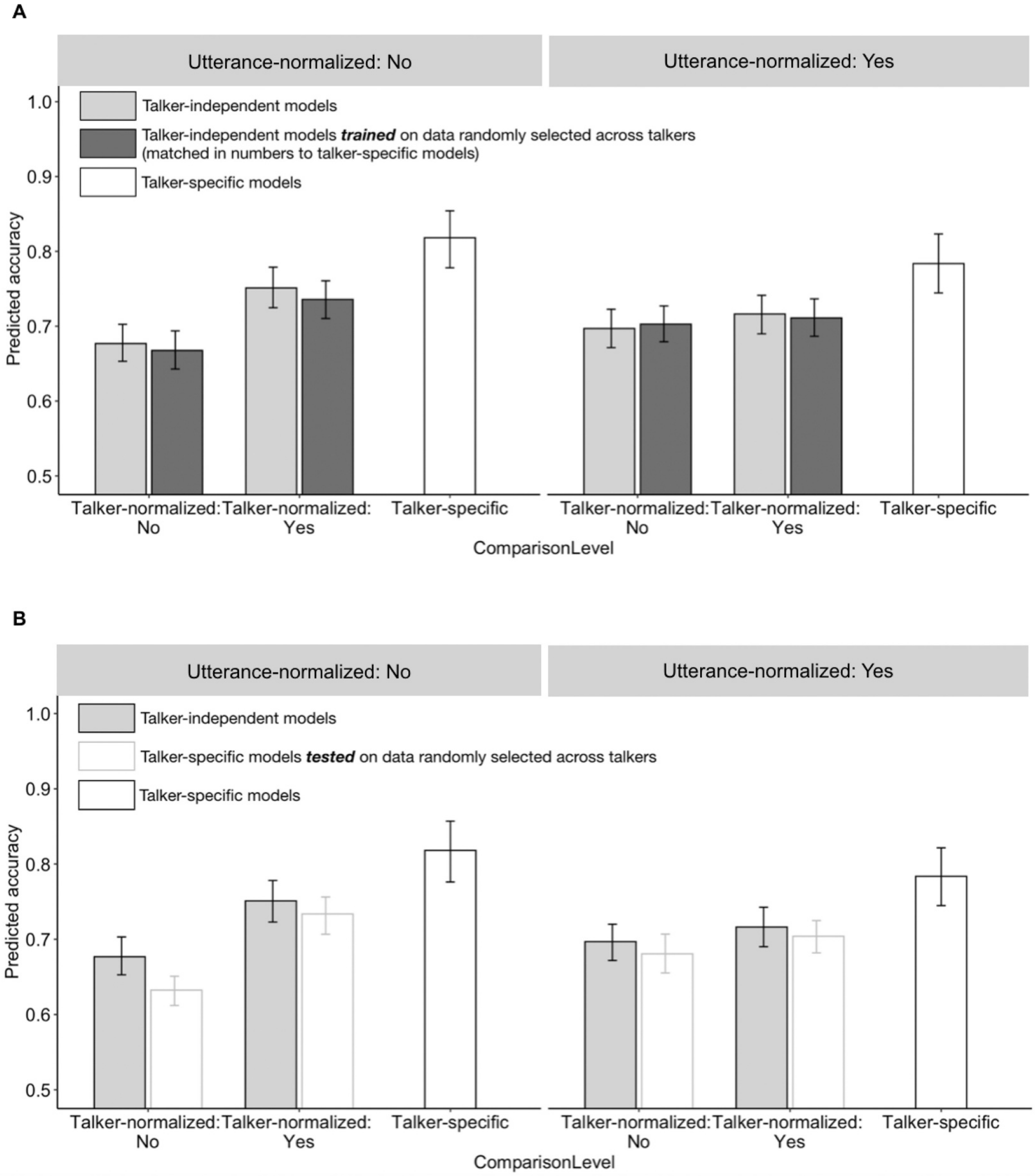

Fig. A1. Predicted accuracy of different ideal observer models. Error bars show bootstrapped 95% CIs of predicted by-talker accuracy (an average of the five folds was computed for each talker, CIs are over these by-talker means). Note that the y-axis starts at chance performance (0.5). In each panel, the talker-independent models (bars in light grey) and the talker-specific models (the rightmost bars) are identical to those shown in Fig. 8 in the main text.

Appendix III Stimuli norming for the comprehension experiment (Experiment 2)

We conducted a norming experiment to assess whether the stimuli created for both talkers spanned the continuum from question to statement, and to ascertain the most ambiguous continuum step. Stimuli created from the female talker were normed by 120 self-reported native speakers of American English, recruited through Amazon Mechanical Turk. 12 participants (10%) were excluded for providing the identical responses to all tokens, leaving 108 participants for analysis. Stimuli created from the male talker were normed by 60 participants recruited in the same manner. 3 participants (5%) were excluded, leaving 57 participants for analysis.

Participants provided two alternative forced choice (2AFC) responses to 12 tokens of *It's X-ing*, each of the six lexical items at two different continuum steps. Six lists were created by varying the nouns at each continuum step to mitigate the potential of item-specific effects. For example, one list used the verb *booting* as Steps 0 and 6, while another list used Steps 2 and 8 for the same item. Order of presentation was randomized across participants. Participants did not receive feedback on their responses.

Results of the norming study are presented in Fig. A2. As expected, listeners' responses were almost categorical at the two continuum ends, gradually shifting from statement (Step 0) to question (Step 11). We identified Step 7 and Step 8 to be the *a priori* most ambiguous items for the female and the male talker (46.7% and 55.3%), respectively. We also determined that responses given to Steps 0 and 1 were more or less identical, and therefore truncated the continuum to 11 steps (Steps 1–11 in the Norming Studies).

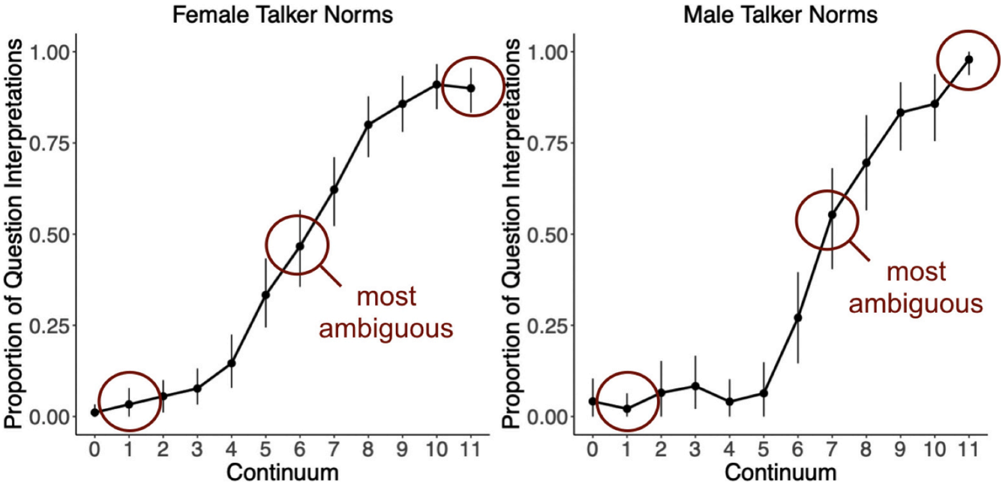

Fig. A2. Proportions of question (as opposed to statement) responses by continuum steps in the Norming study. Error bars indicate bootstrapped 95% confidence intervals. Circles indicate continuum steps used in the exposure phase of Study 2. For both the female and the male talker, steps 1 and 11 were selected as typical statement and question realizations, respectively. Step 6 of the female talker and Step 7 of the male talker were selected as maximally ambiguous token.

Appendix IV 2AFC comprehension responses given to exposure tokens (Experiment 2)

To examine the rate of adaptive changes that happened during the exposure phase, we summarized the responses given to the 30 exposure tokens across the three between subject conditions in Experiment 2 (Fig. A3).

The data support three observations. First, as expected, responses given to the unambiguous tokens (plotted in blue and yellow, Step 1 an 11, respectively) were more or less stable throughout the exposure trials. However, the exact profiles were not identical across the three between subject conditions i.e., Participants seem to have been less accurate in their judgments when they encountered these "unambiguous" tokens in the question-biasing and the statement-biasing conditions, as compared to the no bias condition. We suspect that the presence of the ambiguous tokens might have increased the overall task difficulty by increasing the level of uncertainty in their judgments.

Second, the largest amount of "shift" in the responses given to the ambiguous tokens (i.e., an indication of learning) happened at the beginning of the exposure phase. The change was also incremental (i.e., gradual), happening over multiple trials. We plan to delve into the nature of the adaptation using a Bayesian belief updating models (Kleinschmidt and Jaeger, 2016; Kleinschmidt, 2020).

Finally, to the ambiguous tokens, participants initially provided similar responses across the conditions (i.e., the question-biasing and the statement-biasing) and then gradually learned to provide opposing interpretations (Panel B in the figure below). The adaptive changes seem to be more or less symmetrical between the statement-biasing and the question-biasing conditions although the steepness of the slopes was constrained by the initial biases associated with a given talker. More specifically, the ambiguous tokens sounded somewhat more like a question for the female talker while they sounded somewhat more like a statement for the male talker. We plan to delve into likely sources of these biases as well as into the incremental nature of the adaptation in our future experiments with a wider range of items with increased number of stimuli.

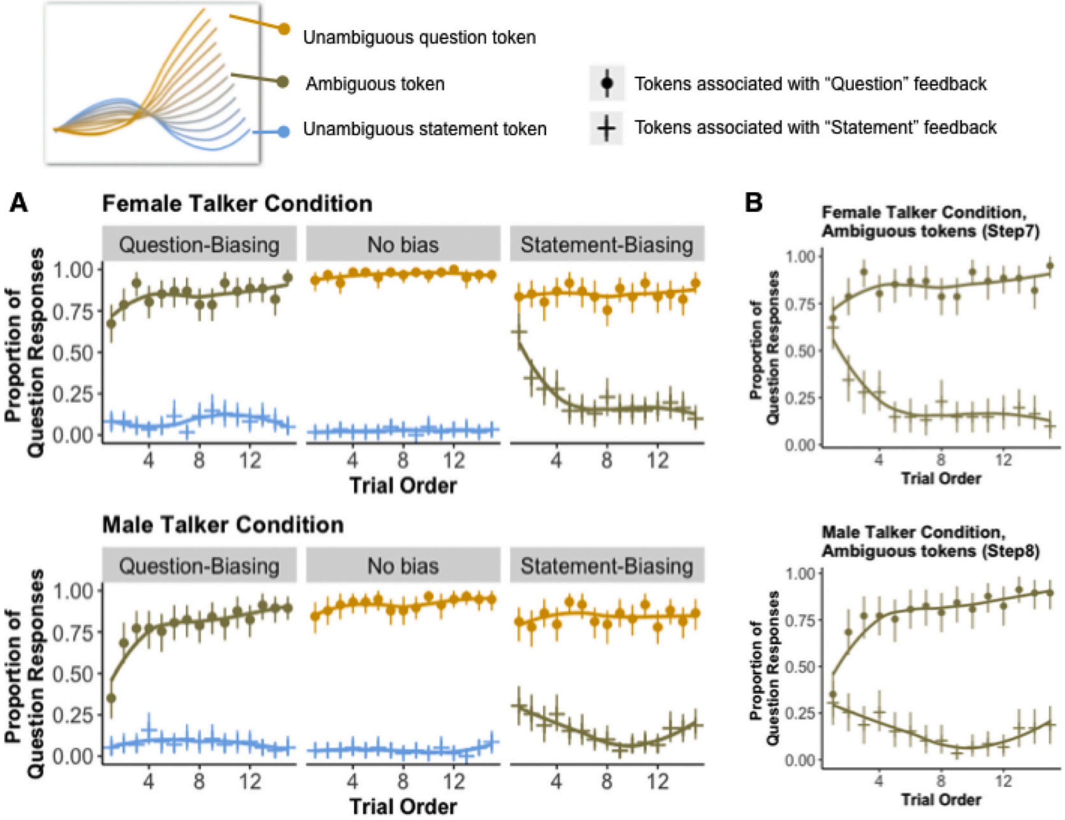

Fig. A3. Proportions of question (as opposed to statement) responses given during the exposure phase by the female vs. the male talker conditions in Experiment 2. Panel A: Overall response patterns across the between subject conditions. The trial order represents the relative ordering of the 15 exposure tokens associated with the question vs. the statement feedback. The colors represent the continuum steps that the stimuli were sampled from: Blue and yellow indicate unambiguous tokens (Step 1 and Step 11, respectively) and green represent the ambiguous items (Step 6 and Step 7 for the female and male talker conditions). Error bars indicate bootstrapped 95% confidence intervals. Circles and crosses indicate the feedback types. Panel B: Responses given to the prosodically ambiguous tokens in the female vs. the male talker conditions; the top and bottom lines represent the Question-Biasing and the Statement-Biasing conditions, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104619.

References

Adank, P., Nuttall, H. E., Banks, B., & Kennedy-Higgins, D. (2015). Neural bases of accented speech perception. *Frontiers in Human Neuroscience, 9*, 1–7. https://doi.org/10.3389/fnhum.2015.00558.

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *The Journal of the Acoustical Society of America, 116*(5), 3099–3107. https://doi.org/10.1121/1.1795335.

Arvaniti, A. (2019). Crosslinguistic variation, phonetic variability, and the formation of categories in intonation. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th international congress of phonetic sciences* (pp. 1–6). Australia: Melbourne.

Arvaniti, A., & Garding, G. (2007). Dialectical variation in the rising accents of American English. In J. Cole, & J. H. Hualde (Eds.), *9. Papers in laboratory phonology* (pp. 547–576). Berlin, New York: Mouton de Gruyter.

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science, 25*(8), 1546–1553. https://doi.org/10.1177/0956797614533705.

Bartels, C. (1999). *The intonation of English statements and questions: A compositional interpretation*. Garland Pub.

Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*. https://doi.org/10.1037/rev0000197.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. (2011). Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLoS ONE, 6*(5), 1–12.

Bishop, J., & Keating, P. A. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *Journal of the Acoustical Society of America, 132*(2), 1100–1112.

Boakye, K., Favre, B., & Hakkani-Tür, D. (2009). Any questions? Automatic question detection in meetings. In *Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 485–489). https://doi.org/10.1109/ASRU.2009.5373293. *ASRU* 2009.

Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer [Computer program]. Software and manual available online at praat.org*.

Bögels, S., & Torreira, F. (2015). Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics, 52*, 46–57. https://doi.org/10.1016/j.wocn.2015.04.004.

Bolinger, D. (1986). *Intonation and its parts*. Stanford: Stanford University Press.

Bolinger, D. (1989). *Intonation and its uses: Melody in grammar and discourse*. London: Arnold.

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2011). Expectations from preceding prosody influence segmentation in online sentence processing. *Psychonomic Bulletin & Review, 18*(6), 1189–1196. https://doi.org/10.3758/s13423-011-0167-9.

Brugos, A., Barnes, J., Shattuck-Hufnagel, S., & Veilleux, N. (2006). A range of intonation patterns produced in an elicitation task. *The Journal of the Acoustical Society of America, 119*(5), 3301.

Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Maintaining information about speech input during accent adaptation. *PLoS ONE*. Burchill, Zachary: zachary.burchill@rochester.edu: Public Library of Science.

Buxó-Lugo, A., Toscano, J. C., & Watson, D. G. (2016). Effects of participant engagement on prosodic prominence. *Discourse Processes, 6950*(June), 1–19. https://doi.org/10.1080/0163853X.2016.1240742.

Buxó-Lugo, A., Toscano, J. C., & Watson, D. G. (2018). Effects of participant engagement on Prosodic prominence. *Discourse Processes, 55*(3), 305–323. https://doi.org/10.1080/0163853X.2016.1240742.

Cangemi, F., & Grice, M. (2016). The importance of a distributional approach to categoriality in autosegmental-metrical accounts of intonation. *Laboratory Phonology, 7*(1). https://doi.org/10.5334/labphon.28.

Cangemi, F., Krüger, M., & Grice, M. (2015). Listener-specific perception of speaker-specific productions in intonation. In S. Fuchs, D. Pape, C. Petrone, & P. Perrier (Eds.), *Individual differences in speech production and perception* (pp. 123–145). Frankfurt am Main: Peter Lang.

Chodroff, E., & Cole, J. (2019a). Relative influences of information structure and utterance-final position on the prosodic implementation of nuclear pitch accents. *The Journal of the Acoustical Society of America, 145*(3), 1933. https://doi.org/10.1121/1.5102043.

Chodroff, E., & Cole, J. S. (2019b). *Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English, 1966–1970*. https://doi.org/10.21437/interspeech.2019-2684.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics, 61*, 30–47. https://doi.org/10.1016/j.wocn.2017.01.001.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804–809. https://doi.org/10.1016/j.cognition.2008.04.004.

Clopper, C. G., & Smiljanic, R. (2011). Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics, 39*(2), 237–245. https://doi.org/10.1016/j.wocn.2011.02.006.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience, 30* (1–2), 1–31. https://doi.org/10.1080/23273798.2014.963130.

Constant, N. (2012). English rise-fall-rise: A study in the semantics and pragmatics of intonation. *Linguistics and Philosophy, 35*(5), 407–442.

Couper-Kuhlen, E., & Selting, M. (1996). *Prosody in conversation: Interactional studies*. Cambridge University Press.

Crystal, D. (1969). *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press. Retrieved from https://books.google.com/books?id=aS45AAAAIAAJ&pgis=1.

Cutler, A. (2015). *Native listening: Language experience and the recognition of spoken words*. Mit Press. Retrieved from https://mitpress.mit.edu/books/native-listening-0.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology, 42* (4), 317–367. https://doi.org/10.1006/cogp.2001.0750.

Dahan, D. (2015). Prosody and language comprehension. *Wiley Interdisciplinary Reviews: Cognitive Science, 6*(5), 441–452. https://doi.org/10.1002/wcs.1355.

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Attention, Perception, & Psychophysics, 27*(5), 435–443. https://doi.org/10.3758/BF03204461.

Dilley, L. C. (2007). Pitch range variation in English tonal contrasts: Continuous or categorical?. In *Proceedings of the International Congress of Phonetic Sciences. Saarbruecken*. Germany.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science, 21*(11), 1664–1670. https://doi.org/10.1177/0956797610384743.

Doherty, C. P., West, W. C., Dilley, L. C., Shattuck-Hufnagel, S., & Caplan, D. (2004). Question/statement judgments: An fMRI study of intonation processing. *Human Brain Mapping, 23*(2), 85–98. https://doi.org/10.1002/hbm.20042.

Doherty, C. P., West, W. C., Redi, L., Gow, D., Shattuck-Hufnagel, S., & Caplan, D. (2003). *The processing of question-intonation: an fMRI study*.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224–238.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116*(4), 752–782.

Flynn, N., & Foulkes, P. (2011). Comparing vowel formant normalization methods. In *Proceedings of ICPhS XVII, (August)* (pp. 683–686).

Foulkes, P., & Hay, J. B. (2015, January 7). The emergence of sociophonetic structure. In *The Handbook of Language Emergence*. https://doi.org/10.1002/9781118346136.ch13.

Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review, 5*(3), 370–389. https://doi.org/10.3758/BF03208814.

Geffen, S., & Mintz, T. H. (2017). Prosodic differences between declaratives and interrogatives in infant-directed speech. *Journal of Child Language, 44*(4), 968–994. https://doi.org/10.1017/S0305000916000349.

Goldinger, S. D. (1996a). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*(5), 1166–1183.

Goldinger, S. D. (1996b). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166.

Goldinger, S. D. (1998). Echoes of echoes?: An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279.

Grabe, E. (2002). Variation adds to prosodic typology. *Speech Prosody, 2002*, 127–132.

Grabe, E., & Post, B. (2002). Intonational variation in the British isles. *Proceedings of Speech Prosody, 2002*, 343–346.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, *64*, 90–107. doi::https://doi.org/10.1016/j.wocn.2017.03.003.

Gussenhoven, C. (1984). *On the grammar and semantics of sentence accents*. Dordrecht: Foris.

Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology. In *Proceedings of Speech Prosody* (pp. 47–57). Aix-en-Provence.

Gussenhoven, Carlos. (1999). Discreteness and gradience in intonational contrasts. Language and Speech, 42(2–3), 283. doi:https://doi.org/10.1177/00238309990420020701.

Haan, J. (2001). Speaking of questions: An exploration of Dutch question intonation. *LOT Dissertation Series, 52*.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31*(3), 373–405. https://doi.org/10.1016/j.wocn.2003.09.006.

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *The Linguistic Review, 23*(3), 351–379.

Hedberg, N., Sosa, J. M., & Görgülü, E. (2017). The meaning of intonation in yes-no questions in American English: A corpus study. *Corpus Linguistics and Linguistic Theory, 13*(2), 321–368. https://doi.org/10.1515/cllt-2014-0020.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristcs of American English vowels. *Journal of the Acoustical Society of America, 97*(5), 3099–3111.

Hirschberg, J., Litman, D., & Swerts, M. (2004). Prosodic and other cues to speech recognition failures. *Speech Communication, 43*(1–2), 155–175. https://doi.org/10.1016/j.specom.2004.01.006.

Hirschberg, J., & Ward, G. (1992). The influence of pitch range, duration, amplitude and spectral features on the interpretation of the rise-fall-rise intonation contour in English. *Journal of Phonetics, 20*, 241–251.

Isaacs, A. M., & Watson, D. G. (2010). Accent detection is a slippery slope: Direction and rate of F0 change drives listeners' comprehension. *Language and Cognitive Processes, 25*(7–9), 1178–1200. https://doi.org/10.1080/01690961003783699.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language, 58*(2), 541–573.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language, 59*(4), 434–446.

Jeong, S. (2018). *Intonation and sentence type conventions : Two Types of rising declaratives, (April)* (pp. 305–356). https://doi.org/10.1093/jos/ffy001.

Johnson, K. (2005a). *Decisions and mechanisms in exemplar-based phonology*. UC Berkeley: Department of Linguistics.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *The Journal of the Acoustical Society of America, 88*(2).

Johnson, K. (2005, January 1). Speaker normalization in speech perception. In *The Handbook of Speech Perception*. https://doi.org/10.1002/9780470757024.ch15.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics, 34*(4), 485–499.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience, 34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698.

Kleinschmidt, D. F. (2020, June 4). *What constrains distributional learning in adults?*. https://doi.org/10.31234/osf.io/6yhbe.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122* (2), 148–203. https://doi.org/10.1037/a0038695.

Kleinschmidt, D. F., & Jaeger, T. F. (2016). What do you expect from an unfamiliar talker? In J. C. Trueswell, A. Papafragou, D. Grodner, & D. Mirman (Eds.), *Proceedings of the 38th annual meeting of the cognitive science society*. Austin, TX.

Kleinschmidt, D. F., Weatherholtz, K., & Jaeger, T. F. (2018). Sociolinguistic perception as inference under uncertainty. *Topics in Cognitive Science, 10*(4), 818–834. https://doi.org/10.1111/tops.12331.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*(2), 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*, 262–268.

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition, 121*(3), 459–465. https://doi.org/10.1016/j.cognition.2011.08.015.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified model of categorical effects in consonant and vowel perception. *Psychological Bulletin and Review*, 1681–1712. https://doi.org/10.3758/s13423-016-1049-y.

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2017). Effects of distributional information on categorization of prosodic contours. *Psychonomic Bulletin and Review, 25*, 1153–1160. https://doi.org/10.3758/s13423-017-1332-6.

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D. F. D. F., & Tanenhaus, M. K. M. K. (2014). Is it or isn't it: Listeners make rapid use of prosody to infer speaker meanings. *Cognition, 133*(2), 335–342. https://doi.org/10.1016/j.cognition.2014.05.017.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 83*(13), 1–26. https://doi.org/10.18637/jss.v082.i13.

Ladd, D Robert. (2008). *Intonational phonology* (2nd ed.). Cambridge University Press.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics, 25*(3), 313–342.

Lee, C.-Y. (2009). Identifying isolated, multispeaker Mandarin tones from brief acoustic input: A perceptual and acoustic study. *Journal of the Acoustical Society of America, 125*, 1125–1137.

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition, 202*(May), Article 104328. https://doi.org/10.1016/j.cognition.2020.104328.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff, & R. Oehrle (Eds.), *Language Sound Structure* (pp. 157–233). Cambridge MA: MIT Press.

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition, 174*, 55–70.

Liu, C., & Rodriguez, A. (2012). Categorical perception of intonation contrasts: Effects of listeners' language background. *The Journal of the Acoustical Society of America, 131* (2012). https://doi.org/10.1121/1.4710836. May. EL427.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance, 41*(6), 1783–1798. https://doi.org/10.1037/a0025641.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America, 49*(2B), 606–608. https://doi.org/10.1121/1.1912396.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001.

Mahrt, T., Cole, J., Fleck, M., & Hasegawa-Johnson, M. (2012). F0 and the perception of prominence. In *13th annual conference of the International Speech Communication Association 2012, INTERSPEECH 2012* (pp. 2421–2424). *3*(January).

de Marneffe, M.-C., & Tonhauser, J. (2019). Inferring meaning from indirect answers to polar questions: The contribution of the rise-fall-rise contour. *Questions in Discourse*, 132–163. https://doi.org/10.1163/9789004378322_006.

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). How the tracking of habitual rate influences speech. *perception, 45*(1), 128–138.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization?: Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review, 118*(2), 219–246. https://doi.org/10.1037/a0022325.What.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). *Within-category VOT affects recovery from "lexical" garden-paths: Evidence against phoneme-level inhibition. 60* pp. 65–91).

Monahan, P. J., & Idsardi, W. J. (2010). Auditory sensitivity to formant ratios: Toward an account of vowel normalization. *Language and Cognitive Processes, 25*(6), 808–839. https://doi.org/10.1080/01690965.2010.490047.

Morrill, T., Baese-Berk, M. M., Heffner, C., & Dilley, L. C. (2015). Interactions between distal speech rate, linguistic knowledge, and speech environment. *Psychonomic Bulletin and Review, 22*(5), 1451–1457. https://doi.org/10.3758/s13423-015-0820-9.

Morrill, T. H., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2014). Distal rhythm influences whether or not listeners hear a word in continuous speech: support for a perceptual grouping hypothesis. *Cognition, 131*(1), 69–74. https://doi.org/10.1016/j.cognition.2013.12.006.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9* (5–6), 453–467.

Nakamura, C., Harris, J., & Jun, S.-A. (2019). Listeners' beliefs about the speaker and adaptation to the deviant use of prosody. In *Poster presented at the 32nd annual CUNY Conference on Human Sentence Processing, Boulder, CO*.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America, 109*(3), 1181–1196.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review, 115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204–238.

Norris, D., McQueen, J. M., & Cutler, A. (2015). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience, 31*(1), 4–18. https://doi.org/10.1080/23273798.2015.1081703.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics, 60*(3), 355–376.

Patel, R., & Brayton, J. T. (2009). Identifying prosodic contrasts in utterances produced by 4, 7, and 11 Year old children. *Journal of Speech, Language, and Hearing Research*, (June), 790–801.

Patel, R., & Grigos, M. I. (2006). Acoustic characterization of the question-statement contrast in 4, 7 and 11 year-old children. *Speech Communication, 48*(10), 1308–1318. https://doi.org/10.1016/j.specom.2006.06.007.

Patel, R., & Schroeder, B. (2007). Influence of familiarity on identifying prosodic vocalizations produced by children with severe dysarthria. *Clinical Linguistics & Phonetics, 21*(10), 833–848. https://doi.org/10.1080/02699200701559476.

Petrone, C., & D'Imperio, M. (2011). From tones to tunes: Effects of the f0 prenuclear region in the perception of Neapolitan statements and questions. In S. Frota, G. Elordieta, & P. Prieto (Eds.), *Prosodic categories: production, perception and comprehension* (pp. 207–230). Dordrecht: Springer.

Pierrehumbert, J., & Steele, S. A. (1987). *How many rise-fall-rise contours?*. Murray Hill, NJ.

Pierrehumbert, J. B. (1979). The perception of fundamental frequency declination. *Journal of the Acoustical Society of America, 66*(2), 363–369.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). John Benjamins.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46*(Pt 2–3), 115–154. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14748442.

Pierrehumbert, J. B., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 271–311).

Pisoni, D. B., & Luce, P. A. (1987). Acoustic-phonetic representations in word recognition. *Cognition, 25*(1), 21–52.

Prieto, P., Estebas-vilaplana, E., & Vanrell, M. (2010). *The relevance of prosodic structure in tonal articulation Edge effects at the prosodic word level in Catalan and Spanish. 38* pp. 687–705). https://doi.org/10.1016/j.wocn.2010.10.004.

R Core Team. (2020). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*. https://www.R-project.org/.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue. *Language and Speech, 54*(2), 147–165. https://doi.org/10.1177/0023830910397489.

Reinisch, E. V. A., & Maximilian, L. (2015). Speaker-specific processing and local context information : The case of speaking rate. *Applied Psycholinguistics, 37*, 1–19. https://doi.org/10.1017/S0142716415000612.

Roettger, T. B., & Franke, M. (2019). Evidential strength of intonational cues and rational adaptation to (un-)reliable intonation. *Cognitive Science, 43*(7), Article e12745. https://doi.org/10.1111/cogs.12745.

Roettger, T. B., & Rimland, K. (2020). Listeners' adaptation to unreliable intonation is speaker-sensitive. *Cognition, 204*, 104372. https://doi.org/10.1016/j.cognition.2020.104372.

Rohde, H., & Kurumada, C. (2018). Alternatives and inferences in the communication of meaning. In K. D. Federmeier & D. G. Watson (Eds.), *Current topics in language, psychology of learning and motivation* (Vol. 68, pp. 215–261). Academic Press. doi::https://doi.org/10.1016/bs.plm.2018.08.012.

Ryalls, J., Le Dorze, G., Lever, N., Ouellet, L., & Larfeuil, C. (1994). The effects of age and sex on speech intonation and duration for matched statements and questions in French. *The Journal of the Acoustical Society of America, 95*(4), 2274–2276.

Saindon, M. R., Trehub, S. E., Schellenberg, E. G., & Van Lieshout, P. (2017). When is a question a question for children and adults? *Language Learning and Development, 13* (3), 274–285. https://doi.org/10.1080/15475441.2016.1252681.

Sawusch, J. R., & Newman, R. S. (2000). Perceptual normalization for speaking rate II: Effects of signal discontinuities. *Perception and Psychophysics, 62*(2), 285–300. https://doi.org/10.3758/BF03205549.

Schweitzer, K. (2012). *Frequency effects on pitch accents: Towards an exemplar-theoretic approach to intonation*. Universität Stuttgart.

Smith, R., & Hawkins, S. (2012). Production and perception of speaker-specific phonetic detail at word boundaries. *Journal of Phonetics, 40*(2), 213–233. https://doi.org/10.1016/j.wocn.2011.11.003.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America, 8*, 185–190. https://doi.org/10.1121/1.1915893.

Studdert-Kennedy, M., Hadding, K., & Hadding-Koch, K. (1973). Auditory and linguistic processes in the perception of intonation contours. *Language and Speech, 16*(4), 293–313. https://doi.org/10.1177/002383097301600401.

Sumner, M., & Samuel, A. G. (2009). The effects of experience on the perception and representation of dialect variants. *Journal of Memory and Language, 60*(4), 487–501. https://doi.org/10.1016/j.jml.2009.01.001.

Tang, C., Hamilton, L. S., & Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science, 801*(August), 797–801.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America, 125*(6), 3974–3982. https://doi.org/10.1121/1.3106131.

Theodore, R. M., & Monto, N. R. (2019). Distributional learning for speech reflects cumulative exposure to a talker's phonetic distributions. *Psychonomic Bulletin and Review, 26*(3), 985–992. https://doi.org/10.3758/s13423-018-1551-5.

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia, 45*(3), 572–577.

Ward, N. G. (2019). *Prosodic patterns in English conversation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316848265.

Warren, P. (2016). *Uptalk: The phenomenon of rising intonation*. Cambridge University Press.

Warren, P. (2017). The interpretation of prosodic variability in the context of accompanying sociophonetic cues. *Laboratory Phonology, 8*(1), 1–21. https://doi.org/10.5334/labphon.92.

Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. *Oxford Research Encyclopedia of Linguistics*.

Weber, A., Braun, B., & Crocker, M. W. (2006). Finding referents in time: Eye-tracking evidence for the role of contrastive accents. *Language and Speech, 49*(3), 367–392.

Witteman, M. J., Weber, A., & McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, & Psychophysics, 75*(3), 537–556. https://doi.org/10.3758/s13414-012-0404-y.

Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language, 97*, 30–46. https://doi.org/10.1016/j.jml.2017.07.005.

Xie, X., Theodore, R. M., & Myers, E. B. (2016). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance, 43*(1), 206–217. https://doi.org/10.1037/xhp0000285.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *The Journal of the Acoustical Society of America, 33*(2), 248. https://doi.org/10.1121/1.1908630.