# Thinking probabilistically in the study of intonational speech prosody

Chigusa Kurumada (University of Rochester)
Timo B. Roettger (University of Oslo)

## Abstract

Speech prosody, the melodic and rhythmic properties of a language, plays a critical role in our everyday communication. Researchers have identified unique patterns of prosody that segment words and phrases, highlight focal elements in a sentence, and convey holistic meanings and speech acts that interact with the information shared in context. The mapping between the sound and meaning represented in prosody is suggested to be probabilistic – the same physical instance of sounds can support multiple meanings across talkers and contexts while the same meaning can be encoded in physically distinct sound patterns (e.g., pitch movements). The current overview presents an analysis framework for probing the nature of this probabilistic relationship. Illustrated by examples from the literature and a dataset of German focus marking, we discuss the production variability within and across talkers and consider challenges that this variability imposes on the comprehension system. A better understanding of these challenges, we argue, will illuminate how the human perceptual, cognitive, and computational mechanisms may navigate the variability to arrive at coherent understanding of speech prosody. The current paper is intended to be an introduction for those who are interested in thinking probabilistically about the sound-meaning mapping in prosody. Open questions for future research are discussed with proposals for examining prosodic production and comprehension within a comprehensive, mathematically-motivated framework of probabilistic inference under uncertainty.

## 1. INTRODUCTION

How humans communicate their thoughts and intentions with each other is at the core of linguistic and cognitive science research. One device that is at the speaker's disposal is speech prosody, a holistic impression of speech rhythm, intonation, and timing. Even with the same set of words, a slight change of intonation contour can sometimes cause a dramatic change in meaning. For instance, a rise in pitch and an elongated syllable can mark emphasis as in "I LOVE you" (not just liking you) vs. "I love YOU" (not someone else) (Figure 1).

(1) A. I LOVE you (not just "liking" you).
B. I love YOU (not someone else).[1]

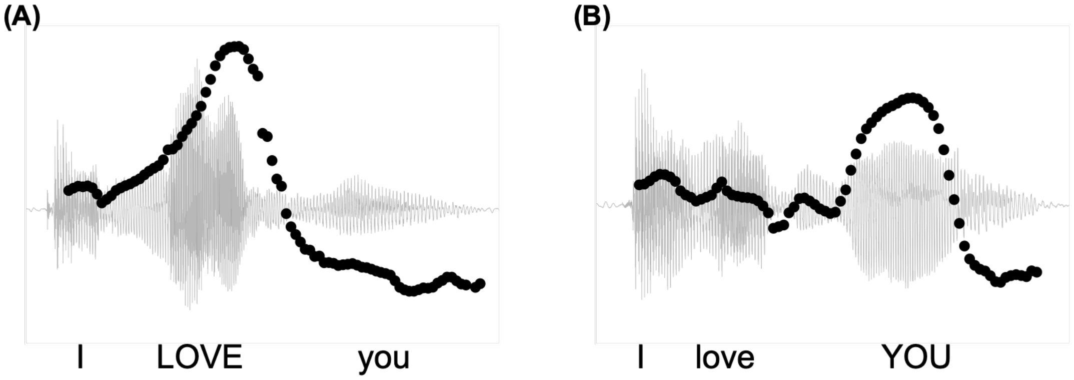

**Figure 1**. Representative waveforms and F0 contours of morpho-syntactically identical statements differing in their information structure. A: "I LOVE you," contrasting the verb with relevant alternatives (e.g. I don't like you, I LOVE you). B: "I love YOU," contrasting the object with relevant alternatives (e.g. I don't love him, I love YOU). The contrastive elements are prosodically marked by a sharp rise-fall in fundamental frequency.

There is now ample evidence suggesting that the information conveyed by prosody is processed online (i.e., in real time), guiding listeners' interpretation of utterances. Speech rhythm and timing influence listeners' segmentation of words (e.g., Baese-Berk et al., 2019; Breen et al., 2014; Cutler, 1990; Heffner et al., 2013; Mattys, 1997; Reinisch et al., 2011b) and phrases (e.g., Carlson et al., 2001; Cutler, 2015; Frazier et al., 2006; Schafer et al., 2000; Snedeker & Trueswell, 2003; Speer et al., 2011; Wagner & Watson, 2010), helping listeners to find linguistic structures in a continuous speech stream. From the earliest stages of language development, prosody provides preverbal infants with auditory envelopes that correspond to what are to be later perceived as words and phrases (e.g., Fernald & Mazzie, 1991; Graf Estes & Hurley, 2013; Soderstrom et al., 2003).

[1] All corresponding data, scripts, and sound files that were used to generate figures in this manuscript, are publicly available here: https://osf.io/b75q9/.

Later in life, proficient listeners can also use the systematic patterns in intonational prosody to anticipate (and potentially predict) incoming linguistic forms and meanings. For instance, English speakers who heard “Hand me the green ball. Now, hand me the YELLOW…*”* often anticipatorily fixate on a yellow object of the same kind (i.e., a yellow ball) more often than other yellow objects in the scene (e.g., Ito & Speer, 2008; see also Kurumada et al., 2014; Tomlinson et al., 2017; Turnbull et al., 2017; Watson et al., 2008; Weber et al., 2006). This is taken as evidence that listeners can anticipate a contextual contrast (e.g., the YELLOW ball, not the green one) signaled by a specific pattern of intonation.

One major puzzle for empirical studies of prosody pertains to the *probabilistic* nature of the sound-meaning mapping (e.g., Calhoun, 2010). As we describe below, what we recognize as a coherent, meaningful pattern of intonation (or “intonational category”) often has a rich internal structure involving acoustically heterogeneous exemplars (e.g., Cole & Shattuck-Hufnagel, 2016). For instance, the contrastive focus on “(Hand me the) YELLOW…” can be produced with distinct acoustic details, depending on who said it, as well as its context and the talker’s emotional state (Ito et al., 2017, see also examples in Section 2.2). This category-internal heterogeneity and gradience are features with which language users richly encode subtle shades of meanings (e.g., Gussenhoven, 1999; Ladd, 2008; Podesva & Calier, 2015; Ward, 2019; Xu, 1997).

The very same features, however, make it difficult for us researchers to understand *how* successful communication is even possible. The heterogeneity means that there is no one-to-one mapping between a cue (e.g., fundamental frequency, henceforth f0) and an intonational category that signals a meaning (e.g., Arvaniti, 2019; Cangemi & Grice, 2016). Further, talkers can vary widely in ways to express themselves via prosody. This is due partly to physiological differences (e.g., vocal tract length correlating with height, Titze (1989)) and their linguistic backgrounds (e.g., regional accents, Arvaniti & Garding, 2007; Cangemi et al., 2015; Clopper & Smiljanic, 2011; Grice et al., 2017; McLarty, 2018). As a result, two instances of utterances that, in one talker, signal distinct intonational categories can map onto a single category in another (for discussion see, Xie et al., 2021). Despite a long history of research, little is known about how listeners can navigate this variability to achieve more or less categorical interpretation of meaning conveyed via intonation (e.g., Arvaniti, 2019; Gussenhoven, 1999; Roessig et al., 2019).

The goal of this paper is twofold. First, we provide an overview of the variable (i.e., probabilistic) mapping between sound and meaning in prosody – how it arises and how it impacts comprehension. Second, we focus on the nature of inter-talker variability to discuss possible ways in which listeners may cope with the ambiguity in the cue-category mapping. To do so, we will examine data of German speakers collected by Grice et al. (2017). In their production experiment, five different talkers produced three different target words under four different intended meanings. These talkers are relatively homogenous in their age and general accent features, and yet their productions vary substantially along several acoustic and phonetic dimensions (see Mücke & Grice 2014, Grice et al. 2017). This dataset thus gives us a window into the nature of variability across different talkers.

What we seek to demonstrate in these discussions is the power of *thinking probabilistically* about prosody. The inherent noise and variability in the prosodic productions make it unlikely that listeners recognize a meaningful pattern of prosody by directly (or deterministically) mapping phonetic cues (e.g., f0) to abstract, intonational categories and then to meanings. Instead, the framework for thinking probabilistically seeks to explain the process by treating the problem as that of *probabilistic inference,* wherein listeners make sense of the ambiguous input by evaluating the observed input and their prior knowledge about how these cues are distributed over possible intonational meanings.

To illustrate, let us briefly consider an example of human perceptual judgments in a visual domain. Just as in the auditory domain, the visual system must routinely navigate the noisy and variable perceptual information flooding the retina and the brain. Under many circumstances, the bottom-up perceptual input itself is ambiguous and does not categorically separate two objects (e.g., a tree or a utility pole). A number of influential theories assume that it is the brain that resolves the ambiguity (e.g., Gregory, 2015): it integrates the sensory signal into the prior probabilistic knowledge to infer the likely source of the visual input (Bar, 2004; Kersten et al., 2004; Knill & Pouget, 2004). If it is a tree (or a utility pole), what types of lines, contours, and shades should the visual system be perceiving? How likely is it to see a tree (as opposed to a utility pole) in a given visual scene (e.g., Are you in a remote campsite in Norway or in a buzzing neighborhood in Tokyo)? “Seeing” an image or an object, thus, is buttressed both by the signal and by the implicit knowledge accumulated over relevant past experiences.

“Hearing” prosody, we argue, can also be productively understood as a probabilistic inference. Because of the ubiquitous variability, the perceived phonetic cues alone seldom distinguish possible categories (and associated meanings) in a deterministic manner. It is the brain that hypothesizes and then infers the category by combining the cues and the knowledge of how different talkers would produce those cues under different circumstances. If the talker is expressing a contrastive meaning, what types of pitch contours, syllable durations, and intensity values should the auditory system be perceiving? How likely is it for the speaker to express the contrastive meaning (as opposed to other information status of the referent) in a given discourse context?

To begin, in Section 2 below, we will provide a brief review of the previous work on prosodic variability and its relevant sources and consequences for speech recognition. Section 3 will discuss possible ways in which within- and across-talker variability might manifest itself in German talkers’ productions of prosodic focus marking. Section 4 will present a probabilistic inference approach that considers the variable cue-category mapping in the prosodic input across talkers. Section 5 will summarize the discussions and present future directions.

## 2. LACK OF INVARIANCE AND FORM-MEANING MAPPING IN THE PRESENCE OF TALKER VARIABILITY

### 2.1. Sound-meaning mapping in encoding of prosody

In this section, we will provide an overview of the sound-meaning mapping, as well as its inherent variability. We note that providing a fully-fledged review of the relevant research in the domain of prosody goes beyond the scope of this paper. Interested readers should consult recent comprehensive summary articles and volumes (e.g., Cole, 2015; Cole & Shattuck-Hufnagel, 2016; Cutler, 2015; Dahan, 2015; Ladd, 2008; McQueen & Dilley, 2020; Speer & Ito, 2009; Wagner & Watson, 2010). Here we instead aim to paint a holistic picture of what constitutes the sound-meaning mapping in prosody and what it means for such a mapping to be "probabilistic."

As we noted above, speech prosody is an amalgam of sound features creating a holistic impression of speech. Among the levels of speech hierarchy, prosody is often characterized as *suprasegmental*, meaning that relevant representations sit on top of smaller atomic units (i.e., *segments* such as vowels and consonants). This distinction between segmental and suprasegmental also pertains to different meanings conveyed. While segments commonly distinguish words (e.g., /p/ig vs. /b/ig), suprasegmental features often convey additional information such as phrasal boundaries (e.g., "Tap the frog with the pen" can be parsed as "Tap the frog | with (= by means of) the pen" or "Tap | the frog with (= in possession of) the pen"), utterance-level prominence (e.g., information status and focus), and speech acts (e.g., question vs. statement). Note, however, that the precise roles suprasegmental features play can vary widely across languages (Jun, 2005). Moreover, processing of segmental and suprasegmental representations is mutually constraining (e.g., Dilley et al., 2010). It suffices to say, for the current purposes, that a given prosodic representation can span units of varying sizes, from a single syllable to an entire utterance, and the meaning that is conveyed can go beyond lexical identities.

Although some accounts assume direct mapping between the sounds and their meanings/functions (e.g., Cooper et al., 1985; Fry, 1955; Xu, 2006), many accounts of intonational speech prosody posit intermediate, phonological representations that usefully group together sound patterns that convey a coherent meaning (e.g., Ladd, 2008; Pierrehumbert, 1980; Pierrehumbert & Hirschberg, 1990). For instance, the prosodic emphasis placed on the words "love" and "you" in Figure 1 is considered to be phonologically equivalent, conveying the meaning of highlighting a contextually relevant contrast. Here and thereafter, we refer to these representations as (intonational) *categories* and their phonetic features as *cues*.[2] A cue can be an absolute value along a given phonetic dimension (e.g., F0, duration, amplitude) or a relative value with respect to a particular linguistic landmark (e.g., alignment between a pitch peak and a lexically stressed syllable).

To illustrate, we adopt a particularly influential framework, the "autosegmental-metrical (AM)" framework and its nomenclature to illustrate the architecture often assumed for productions of intonation in languages like German and English (Arvaniti, 2017; Grice et al., 2005; Ladd, 2008; Pierrehumbert, 1980; Silverman et al., 1992, for typologically diverse languages see Jun, 2005). In the AM framework, an

[2] There is a long history of debate in the field of intonational speech prosody about the exact nature and number of intonational categories to be posited for a given language (e.g., Arvaniti, 2019; Breen et al., 2012; Syrdal & McGory, 2000; Wightman, 2002). We cannot exhaustively review the debate here. Note though that our approach to thinking probabilistically developed here is not agnostic to the number of categories for the reasons we elaborate on in Section 3.1.

intonational structure is considered to include a succession of tonal targets (such as H(igh) and L(ow)) which systematically co-occur with structural landmarks, resulting in categories also as known as “tonal events”. There are two cross-linguistically common types of tonal events that are thought to characterize a more holistic structure of intonation. The first type is called *pitch accent* -- a tonal event that is associated with a lexically stressed syllable. In Figure 2, the stressed syllables in “hand” and “yellow” receive pitch accents (e.g., as annotated as H* and L+H*), that mark these lexical items as prominent. The second type is called *edge tone* (or boundary tone as annotated as L-% or H-%) -- a tonal event that occurs at the edge of a constituent.  In Figure 2, “ball” receives an edge tone (e.g., L-%), that marks this lexical item as prominent.

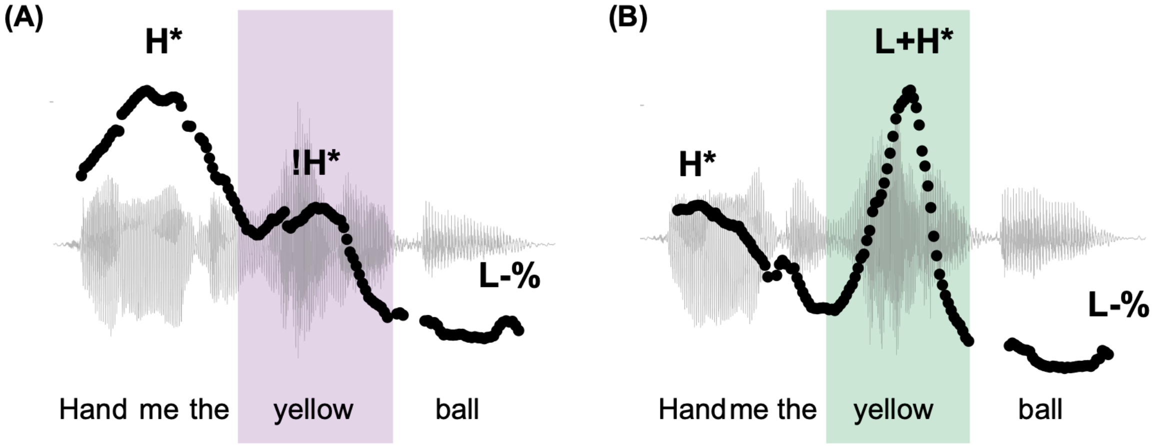

**Figure 2.** Representative waveforms and F0 contours of “Hand me the yellow ball” differing in its information structure. (A): An answer to “What ball do you need?”, rendering “yellow (ball)” as narrow focus. In this case, the constituent is marked by a (down-stepped) H*, characterized by a shallow rise of f0. (B): An answer to “Do you want me to hand you the green ball?”, rendering “yellow” as contrastive focus. In this case, the constituent is marked by an L+H*, characterized by a steep rise in f0 to a high target and a sharp fall.

The contrast between the two types of pitch accents, "H*” (an acute rise in pitch) vs. “L+H*” (a steep pitch rise followed by a fall), constitutes one of the most widely studied examples of tonal targets and their meanings conveyed in context (e.g., Calhoun, 2004; Dilley et al., 2005; Ito et al., 2017; Ito & Speer, 2008; Pierrehumbert & Hirschberg, 1990; Tomlinson et al., 2017; Watson et al., 2008; Weber et al., 2006). Listeners are sensitive to the differences in the shape of pitch movements and can often rapidly and successfully arrive at different meanings (i.e., narrow focus vs. contrastive focus). In the first utterance of Figure 2, the word “yellow” simply modifies the ball whereas in the second utterance it highlights a contextual contrast (e.g., “yellow not green”).

As the labels such as “H(igh)” or “L(ow)” suggest, tonal events (both pitch accent and edge tones) are inherently relative and proportional rather than absolute (Pierrehumbert, 1980). More specifically, these events are defined with respect to local F0 ‘peaks’ or ‘valleys,’ whose specific phonetic values depend on varying factors, such as the degree of emphasis of a stressed syllable or its position in the utterance. Figure 3A demonstrates this heterogeneity (originally included in the now classic work by Liberman and Pierrehumbert (1984), and subsequently incorporated into Ladd and Morton (1997)). The f0 contours depicted are considered instances of the “same” intonation pattern, “with variation in the pitch range signaling different degrees of emphasis independently of what is conveyed by the choice of intonation pattern” (Ladd and Morton, 1997, p.313). Even within the same talker, peaks form a range, and so do valleys.

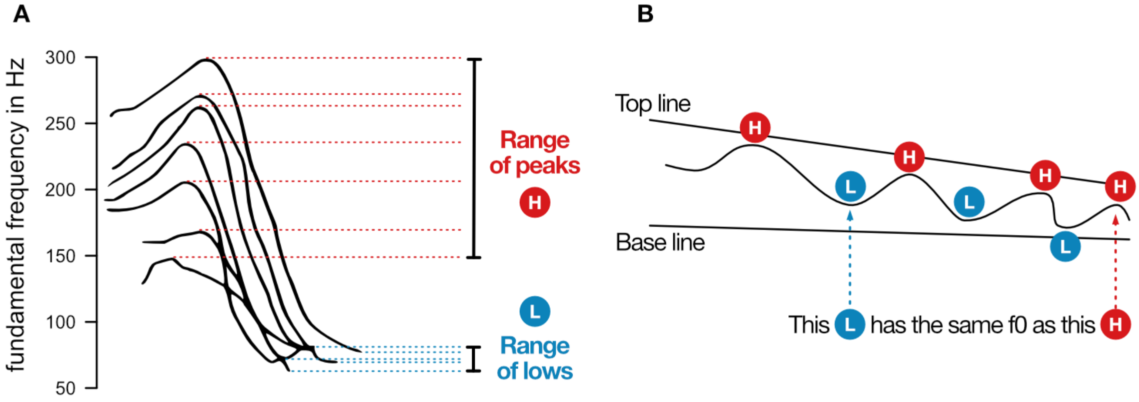

**Figure 3.** Two illustrations of the relative nature of f0 targets. (A): F0 contour for the same sentence spoken with “the same” intonation in different pitch ranges. Adapted from Liberman & Pierrehumbert (1984). (B): illustration of declination. A speaker's pitch range may fall or rise during speech, independently of the falls and rises of $f_0$. Adapted from John Coleman’s website: http://www.phon.ox.ac.uk/jcoleman/intonation.htm

Further, to produce a coherent intonation consisting of multiple tonal events, talkers must sequence these underlying peaks or valleys by connecting them via interpolation of phonetic cue values while adhering to other, more global language-specific constraints (Pierrehumbert, 1980). For instance, it is well known that an English talker's vocal pitch starts higher at the onset of an utterance, and gradually decreases during speech (e.g., *declination* for gradual narrowing of a pitch range, and *downstepping* for more categorical shifts in a baseline) (e.g., Cohen et al., 1982; Grice, 1995; Gussenhoven & Rietveld, 1988; Ladd, 1988). Importantly, these changes occur simultaneously but independently of falls and rises of F0 that indicate peaks and valleys. Consequently, different peaks at different points in an utterance would likely vary in their absolute pitch height, as can be seen in Figure 3B.

In a nutshell, an intonation contour can consist of multiple tonal events, each of which groups together phonetic cues and then connects these cues to meanings in context. While these tonal events are often posited as distinct and discrete, their phonetic realizations are inherently variable and continuous. In extreme cases, given instances of a “high” tone and a “low” tone have the same objective f0 values. This heterogeneity, as we will discuss further below, is compounded by lexical identities and linguistic contexts (Cole, 2015) as well as extra-linguistic differences originating from talkers (e.g., pitch range differences between male vs. female, Bishop & Keating, 2012), speaking styles (e.g., child-directed vs. adult-directed speech, Dilley et al., 2014; Foulkes et al., 2005), and social identities of talkers (e.g., “uptalk,” Warren, 2016). It is these sources of variability that make the mapping between phonetic cues and phonological categories “probabilistic” from the listeners’ point of view, which we turn to next.

## 2.2. Lack of invariance in the form-meaning mapping in prosody

Let us now consider how this variable category-cue mapping may affect *the listener,* whose goal includes recovering intonational meanings intended by the talker. Because observed variations of cues can be due to many linguistic and non-linguistic causes, listeners must be able to distinguish phonetic cues that are indicative of underlying tonal events from those that are indexical or incidental in nature (Arvaniti, 2019). This, however, is not a trivial problem for multiple reasons.

First of all, the transmission of the speech signal is known to be perturbed by random noise and sensory uncertainty generated by neural, articulatory, or auditory mechanisms (e.g., Adank, 2012; Adank et al., 2011; Feldman et al., 2009; Jones et al., 2013; Summerfield, 1981; Summers et al., 1988). When the same talker produced the same utterance twice under (hypothetically) exactly the same condition, the auditory signatories of the productions would hardly be identical. Put differently, what listeners receive as the perceptual cues to an intonational category are rarely stationary. Rather, they regularly form a *distribution,* where data points are clustered around a particular central tendency (e.g., a category mean) with some amount of dispersion around it. When the speech signal is perturbed by multiple sources of noise, the *width* of the dispersion (i.e., variance) expands, which results in increased uncertainty about which category was intended by the talker.

Second, as we mentioned in Section 2.1 above, there are often multiple factors that vary cue-category mapping in a more or less systematic, but not fully predictable, manner. Those include physiological features of talkers (e.g., vocal tract length correlated with age and physiological features of talkers Titze (1989)) as well as their accent features (e.g., Arvaniti & Garding, 2007; Clopper & Smiljanic, 2011; Holliday, 2019; McLarty, 2018; Mücke et al., 2009). For instance, Arvaniti and Garding (2007) found that regional variants of American English can differ in terms of phonetic realizations of pitch accents. Southern Californians, on average, show a pattern where their pitch alignment for the L+H* accent is later compared to Minnesotans, and that some Minnesotan talkers may lack the H* vs. L+H* contrast altogether. Further, prosodic features can be used to encode talkers’ intended speech styles, social identities, and backgrounds (e.g., Eckert, 2016; Grabe & Post, 2002; Halliday, 1967; Holliday, 2019; Podesva, 2011; Ponsot et al., 2018; Warren, 2017).

To further complicate the picture, within a group that shares some prosodic features, talkers often show more random, or idiosyncratic, differences (e.g., Cangemi et al., 2015; Cangemi & Grice, 2016; Chodroff & Cole, 2019; Dahan & Bernard, 1996; Fuchs et al., 2015; Xie et al., 2021). Here we illustrate this with production data from five German talkers’ productions of utterances with either a narrow focus meaning or a contrastive focus meaning (Figure 4A, Grice et al., 2017). The narrow focus is often used to highlight the information status of a referent as new (e.g., “Who does Melanie want to meet?” “(Melanie wants to meet) **Dr. Bahber**”). On the other hand, the contrastive focus can present a contrast between contextual alternatives (e.g., “(Melanie wants to meet) **Dr. Bahber, not Dr. Shmidt**”).

As discussed above, these focus meanings are often expressed via distinct intonational categories, which differ along multiple cue dimensions. For simplicity’s sake, we zoom in on a dimension called *onglide*, which is considered critical in encoding the narrow vs. the contrastive focus in German (Ritter & Grice, 2015). Onglide is a dynamic measure, capturing the pitch movement leading towards the target on the accented syllable (Figure 4B). Figure 4C provides time-normalized, superimposed F0 contours for narrow and contrastive focus for each talker separately. Overall, with both focus categories, pitch rises toward the stressed syllable and falls after reaching a local maximum. Exact ways to produce the focus categories, however, seem to differ across talkers.

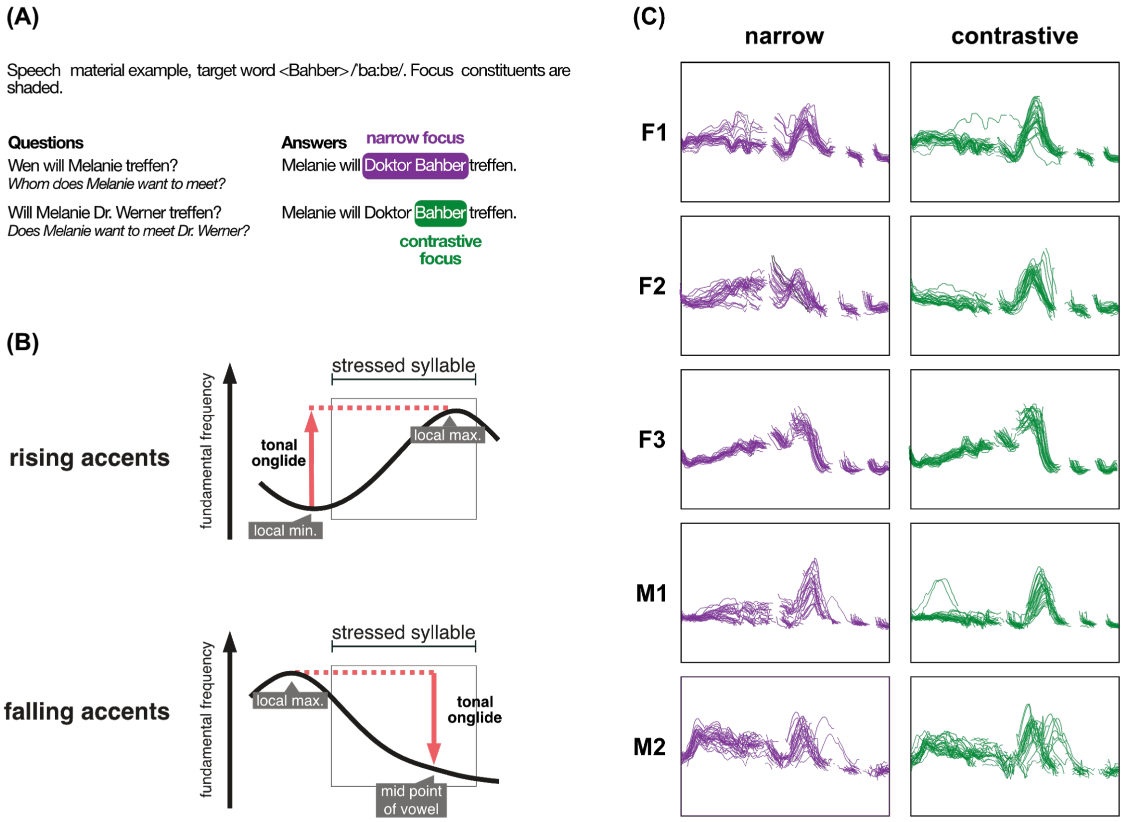

**Figure 4**: (A) Example sentence of the speech materials used in Grice et al. (2017); (B) Schematic depiction of the tonal onglide measurement for both rising and falling accents (adopted from Roessig et al. 2019); (C) Time-normalized, superimposed f0 contours for narrow and contrastive focus displayed for each talker separately.

Figure 5A summaries distributions of onglide measures over 208 tokens produced by the five talkers. Utterances produced under the contrastive focus meaning, compared to narrow focus meaning, exhibit an overall larger onglide value. Evidently, however, the two distributions in Figure 5A show a great degree of overlap. A token with the onglide value of five semitones[3], for instance, can be highly ambiguous between the two focus types.[4] Figure 5B demonstrates that this ambiguity is at least partly due to the fact that underlying distributions of onglide values vary across talkers. Some talkers produce a greater magnitude of rises for both of the focus types (e.g., M1) compared to others (e.g., F3). Some talkers are also more consistent in their encoding of the two meanings (relatively narrower variances, e.g., F3) whereas others are more variable across instances of productions (e.g., F2).

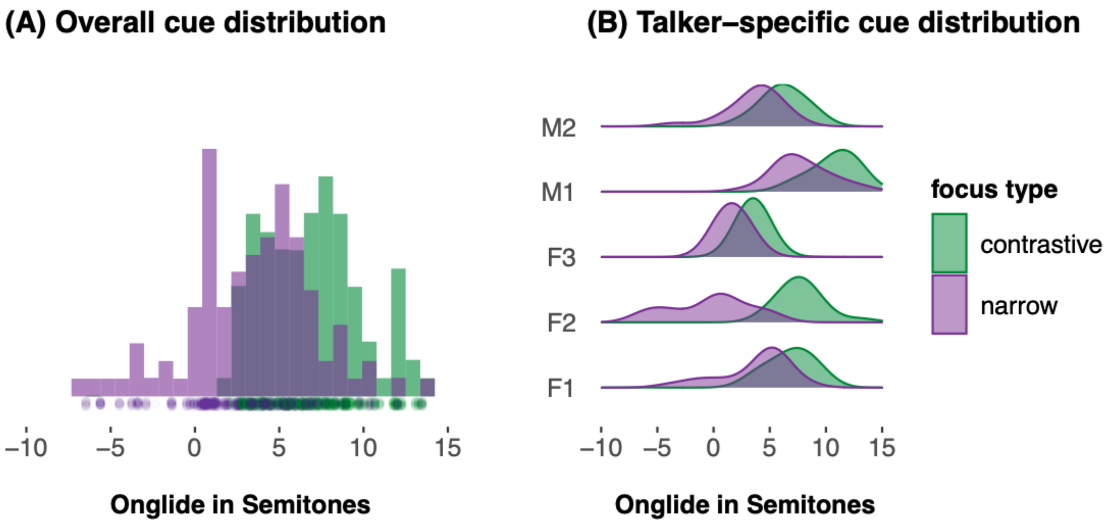

**Figure 5.** Cue distribution of Grice et al. (2017). (A): Histogram of the onglide measurement (in semitones) across all talkers illustrating large overlap between focus categories (color coded). (B): Kernel density plots for all five talkers illustrating large variability in distribution location, spread, and shape.

Recall that an onglide value is inherently *proportional* to a contextually determined anchor point (Figure 4B). The variability is therefore not attributable to baseline differences across talkers (e.g., male talkers typically having overall lower base vocal pitch spanning a narrower pitch range). The data instead suggest that talkers are inherently variable in how they produce the two focus types. This, in turn, creates uncertainty for the listener who must recover the intended intonational categories and their meanings.

[3] A semitone is the smallest interval used in classical Western music. One semitone is equal to a twelfth of an octave or half a tone.

[4] This type of category overlap has been observed across intonational categories. For instance, Arvaniti (2019) discussed two bi-tonal accents, L+H* and L+<H*. Although they are often thought of as distinguishable from one another in their pitch peak alignment (i.e., alignment between a local maximum of F0 and a lexically determined stressed syllable), their phonetic realizations show overlapping distributions (See also Lohfink et al., 2019). Cangemi and Grice (2017), too, provided a similar observation for intonational categories that are typically considered as dissociable in terms of both forms and meanings (e.g., question vs. statement intonations in Neopolitan Italian. For a similar finding for questions vs. statements in English, see Xie et al., 2021). Across languages, empirical studies have so far shown that categories rarely “underlap” (= having no overlap) prior to incorporating talker identities.

### 2.3. Summary

Intonational meaning is conveyed via inherently relative and gradient signals, and their acoustic and phonetic realizations are perturbed by multiple sources of noise and variability. In particular, the significant talker-to-talker variability renders the mapping between cues and intonational meaning *non-stationary*, i.e., from a listener's perspective, the precise correspondence between a phonetic cue value and its associated meaning keeps changing across talkers. This is often called the "lack of invariance" problem in speech perception (e.g., Liberman et al., 1967; Allen, Miller, & DeSteno, 2003; Hillenbrand, Getty, Clark, & Wheeler, 1995; Newman et al., 2001). A number of modern-day speech perception and spoken word recognition models were developed to address this problem within the bounds of human perceptual and computational capacities (e.g., Dahan et al., 2008; Foulkes & Hay, 2015; Kleinschmidt & Jaeger, 2015; Magnuson et al., 2020; Norris et al., 2015; Pierrehumbert, 1994; Samuel & Kraljic, 2009).

Compared to phonemes that can be disambiguated in a lexical contrast, patterns of intonation that distinguish meanings may perhaps be even more vulnerable to this problem because of the abstract and often gradient nature of meanings. Listeners must navigate the variable cue-category mappings without clear feedback on which meaning the talker must have intended even after the sentence is heard. In the sections below, we begin to explore a new approach to this long-standing puzzle—*How do listeners' arriving at a coherent intonational meaning from variable and often ambiguous phonetic cues?* To anticipate, a key to this puzzle can be found in how the human comprehension system makes a smart (and educated) guess under uncertain conditions.

## 3. DISTRIBUTIONAL ASSUMPTIONS AND GENERATIVE MODELS OF PRODUCTIONS

### 3.1 Three accounts on how to navigate talker variability in prosody

Before expounding our own proposal, let us consider existing accounts proposed to address the (subjective) invariance in cue-category mappings in prosodic comprehension. To our knowledge, three classes of accounts have been put forward in the literature.

(1) **Cue integration**: A bottom-up, phonetic cue is bound to be ambiguous and is typically underspecified in its category affiliation, especially when produced by multiple talkers. Listeners (and transcribers alike) resolve the remaining ambiguity by integrating: (a) multiple acoustic/phonetic cues (e.g., Brugos et al., 2018; Cole & Shattuck-Hufnagel, 2016)[5]; and/or (b) top-down information such as expectations about a metrical structure (e.g., Calhoun, 2010), relative

[5] The same logic applies to perception of gradient features such as loudness. Past research has found that human perception of loudness results from an integration of different acoustic dimensions such as frequency ranges (e.g., Fletcher & Munson, 1933; Plack & Carlyon, 1995; Suzuki & Takeshima, 2004), signal durations (e.g., Turk & Sawusch, 1996; Olsen, Stevens, & Tardieu, 2010; Seshadri & Yegnanarayana, 2009), and periodic structures (e.g., Hellman, 1972; Bao & Panahi, 2010).

frequency, and predictability of lexical items (e.g., Cole et al., 2010; Roy et al., 2017) and their meaning (e.g., Arvaniti, 2019).

(2) **Scaling/normalization**: Talker variability can be, at least partially, removed by compensating for differences across talkers (e.g., pitch range normalization, speech rate compensation e.g., Diehl et al., 1980; Féry & Kügler, 2008; Francis et al., 2006). Listeners would achieve invariance by continuously normalizing and scaling phonetic cues according to contextual baselines derived, for example, from syllables upstream in an utterance (Brown et al., 2011, 2015; Dilley & Pitt, 2010; Pitt et al., 2016; Reinisch et al., 2011a).

(3) **Learning/Storage**: Listeners may cope with the variability through storing talker- (and context-) specific patterns of prosodic productions. They can represent and store heard exemplars (Hawkins, 2003; A. Schweitzer, 2019). They may also learn underlying cue distributions (Kurumada, Brown & Tanenhaus, 2017) and/or the mapping between a category and a meaning (Roettger & Franke, 2019) from episodes of language comprehension. They then retrieve these learned representations when they encounter the same talker/context.

These accounts are not necessarily mutually exclusive and can also work in parallel (Lehet & Holt, 2020). For instance, listeners could apply the scaling/normalization over utterances produced by a particular talker and “learn” a normalization constant (Baese-Berk et al., 2014). These classes of approaches, however, show intriguing conceptual oppositions in terms of how listeners *should* deal with the variability in phonetic input. The cue integration account assumes that listeners do not resolve variability seen for a particular cue dimension without leveraging other sources of information. The scaling/normalization account assumes that listeners *remove* or *discard* the variability by perceiving cues relative to an appropriate contextual baseline (McMurray & Jongman, 2011). The learning/storage account, instead, expects listeners to *track*, and perhaps *draw on*, detailed information about the variability across different talkers (e.g., Schweitzer, 2019; Smith & Hawkins, 2012). One major challenge that we face is thus to synthesize these conceptually contrasting accounts and mechanisms to explain prosodic comprehension.

Our proposal is the following. The three accounts above are often built on different assumptions about the nature and structure of talker variability (e.g., where and how talkers differ from one another). In other words, these are not to be considered as three different solutions to the same problem. Each of them, we argue, presupposes and addresses a distinct problem in relation to talker-variability. The first step towards synthesizing these accounts, therefore, is to explore *possible and actual ways in which talkers vary in their prosodic productions*. To this end, we continue to examine the example of German focus marking that we introduced above. We zoom in on two out of the five talkers and discuss how the “probabilistic” mapping between phonetic cues and meaning might vary across talkers.[6]

### 3.2 Models of productions

Figure 6 provides four different distributional representations of onglide measures produced by F3 (female) and M1 (male). All of them are plotted based on the identical set of production data, with different parameters or “distributional assumptions” about the underlying talker variability. We call each of these representations a *model* of these two talkers’ productions. More specifically, we consider each of them as a separate *generative model* – a listener’s mental model of the process whereby an underlying focus type (i.e., coded as purple or green) generates the observed phonetic cue in the context of these two talkers. This follows an influential research tradition in studies of perception and cognition, in which humans categorize perceptual data based on their beliefs about underlying distributions of each category (e.g., Ng & Jordan, 2001). Below we consider these models one by one and explore how different models support the three distinct accounts we listed above.

Before proceeding, an important terminological note is in order. Throughout the discussion below, we refer to each of the purple and green representations as the narrow or contrastive “category.” This is defined as an empirical distribution of one or more phonetic cues, produced under an intended meaning (i.e., focus type). As such, it is not to be equated with an intonational (phonological) category, such as H* or L+H* discussed above. A distributional category we consider here can subsume multiple intonational categories as well as phonetic variations of each. For instance, when a given talker produced more than one type of pitch accent under the contrastive focus meaning, the contrastive focus category (i.e., plotted as a green distribution) includes cue values from these heterogeneous instances of productions. We take this approach to capture the full-range of phonetic and phonological

[6] We wish to acknowledge two caveats. First, for simplicity’s sake, we will continue to focus on onglide as a phonetic dimension crucial to separating the two focus types of our interest. We will come back to this assumption in Section 5 and discuss how the argument we make based on onglide can generalize to other cues or to a combination of cues. Second, the distributional statistics (e.g., category means and variances) that we examine for the two talkers are estimated from a relatively small data set. Each sentence was produced by each talker seven times across three different target words, so our estimates are based on maximally 21 observations. To accurately estimate phonetic cue distributions associated with a given category produced by a given talker, we would ideally want more data in terms of both type and token frequencies of items. However, *how much data we would need* is an empirical question, and we determined that the current dataset is (a) representative of typical production studies in this area, and (b) large enough for illustrative purposes.

variability observed for the two focus types.[7] We will return in Section 5.2 to the question of whether listeners derive an intonational meaning directly from the phonetic cue distributions or through an intermediate level of intonational categories.

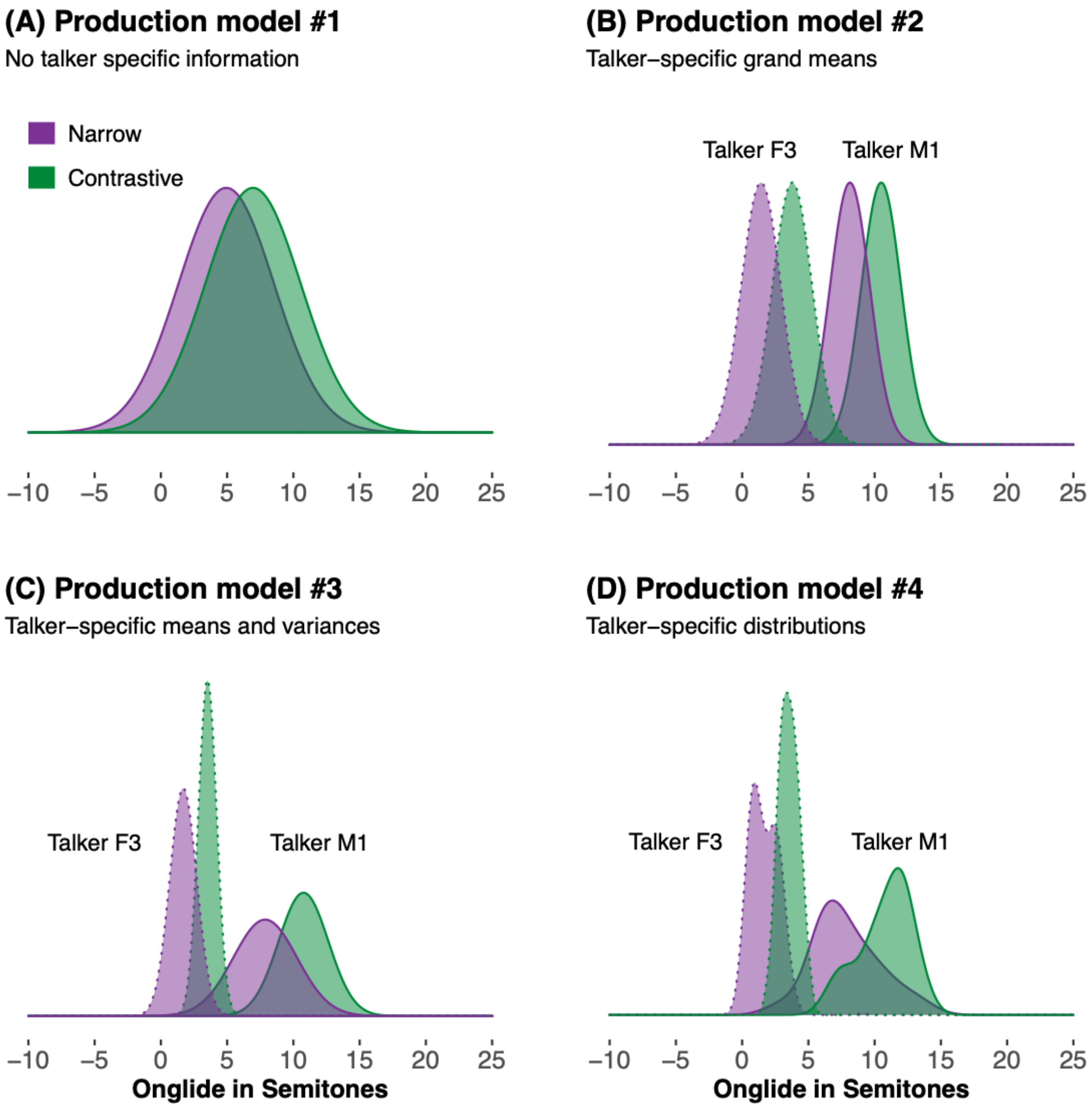

**Figure 6**. Kernel density plots illustrating four possible production models. (A) Production model #1 (top left) assumes Gaussian distributions generated based on the observed means and standard deviations for narrow and contrastive focus aggregated over F3 and M1 in Grice et al. (2017) original dataset. (B) Production model #2 (top right) assumes Gaussian

[7] For a related discussion about empirical, perceptual units of perception vs. linguistic units of analysis, consult Samuel (2020). Our discussion here is inspired by his approach to considering the perceptual system as *omnivorous* – “if the input consistently includes a particular pattern, that pattern can be learned as a “chunk,” and such chunks will be used to recognize speech.” And these chunks do not always straightforwardly correspond to linguistic units such as phonemes or allophones. Likewise, we assume that any recurrent pattern in the prosodic input can form a category as long as it is informative about the latent structure that links an intended meaning and its phonetic realizations. These patterns can, but do not have to, correspond to the phonological units of intonational categories that were proposed originally to describe a structure of language.

distributions generated based on the mean observed difference between categories in F3 and M1, as well as the talkers' grand means. (C) Production model #3 (bottom left) assumes Gaussian distributions generated based on F3 and M1's mean and standard deviation for both categories. (D) Production model #4 (bottom right) is the empirical (smoothed) distributions of F3 and M1 as observed by Grice et al. (2017).

### 3.2.1 Model 1: No talker specificity

The first model, arguably among the simplest one could think of, assumes no talker-specificity in ways that cues are mapped onto categories (Figure 6A). It would simply keep track of onglide values of tokens for the narrow (purple) and contrastive focus (green) categories and pool them together irrespective of who they came from. Each category is characterized as a normal, or Gaussian, distribution centered around a specific mean value. The width (i.e., variance) of a category is assumed here to be constant across the categories. In other words, there is only one parameter to characterize the two categories, i.e., the distance between the overall cue mean and a category mean. Intuitively, listeners who have this model believe that (a) all the tokens were produced by a single talker or (b) all talkers' productions are roughly the same in their underlying cue distributions.

Past accounts that deemed prosodic variability to be vast and insolvable at the phonetic level often pointed to this type of significant category overlap. For instance, Arvaniti (2019) investigates two types of phonetically similar pitch accents which are often assumed to be distinguished by the peak alignment. However, the pitch accents overlap in their phonetic realizations and are not readily separable (See also Lohfink et al., 2019). Arvaniti (2019), then, argued that "(I)nvariance on some phonetic dimension should not be considered essential or used as the main criterion for establishing phonological categories" and reliable prosodic processing should also rely on the listener's knowledge of meaning conveyed in context. Even with contrasts that are considered perceptually highly separable (e.g., a rising vs. a falling tone for expressing a question vs. a statement), empirical data often include many "in between" tokens deemed ambiguous unless the identity of a talker is known (Xie et al., 2021). Findings such as these often undergird the argument that listeners consider information sources other than phonetic cues to resolve prosodic ambiguity (i.e., the "cue integration" view as summarized in Section 3.1).

### 3.2.2 Model 2: Talker-specific cue means

The second model assumes that the production data consists of two distinct sets of distributions, each associated with an individual talker. This model differs from Model 1 by one additional parameter, i.e., a talker-specific cue mean (= the mean of the two category means). The distance between the category means and the spread of the distributions are assumed to be constant across the two talkers. Unlike those that have Model 1, listeners with Model 2 believe that talkers can vary in terms of the overall levels of pitch excursion that they produce to encode the two focus types. Put differently, these listeners evaluate a given onglide value relative to a baseline specific to each talker. In the current example, M1 generally produces a greater magnitude of excursion than F3, and mid-range values depicted in Figure 6B are therefore likely associated with the narrow focus if produced by M1 or the contrastive focus if produced by F3.

Past studies investigating scaling and compensation as a solution to talker-variability (often implicitly) presupposed an underlying distribution as depicted in Model 2. In this view, successful categorization of the input would require adjustments of overall baseline differences across talkers. This can be formalized as identifying the grand mean of a particular cue (e.g., onglide) for each talker, and aligning all talkers by subtracting these mean differences or scaling them to a particular range.

In fact, recent neuro-imaging studies have shown that signatures of intonation processing in the auditory cortex directly reflected the encoding of speaker-normalized relative pitch (as opposed to absolute pitch) (Tang et al., 2017). That is, intonation perception is normalized for a talker's mean pitch from an early stage of auditory processing. This adds credence to the rich body of work showing talker-dependency of pitch perception as well as rate-dependent perception of durational variables (e.g., Diehl et al., 1980; Dilley & Pitt, 2010; Miller et al., 1984). For instance, listeners can perceive syllables of the same physical duration as "short" or "long" depending on their contextual speech rate and do so in a manner contingent on a given talker's global speech rate (Baese-Berk et al., 2014; Reinisch & Maximilian, 2015; but see (Maslowski et al., 2019). Listeners are thus assumed to achieve perceptual invariance by Computing Cues Relative to Expectations (known as the "C-CURE" model, McMurray & Jongman, 2011) in addition to their absolute values.

### 3.2.3 Model 3: Talker-specific category means and variances

Model 3 in Figure 6C has two more parameters: it allows each category (and its distribution) to vary in their relative location from the grand cue mean as well as the amount of spread around each mean. Tighter clustering (i.e., smaller variances) indicates that tokens produced for a given category were relatively similar to one another, whereas looser clustering (i.e., larger variances) indicates that there was a great deal of variability *within* a talker's productions. In the case of the two talkers considered here, F3 was highly consistent when producing the onglide values for both of the focus categories as compared to M1.

*The degree to which* the underlying distributions differ between Models 2 and 3 (i.e., with or without talker-specific mean and variance values) can, and likely do, vary across individual talkers. Some talkers' productions resemble the "typical" or "average" distributional structure assumed in Model 2 while others deviate from them more drastically. When talkers vary substantially from one another in terms of the distributional structure (i.e., both category means and variances), the scaling/ compensation process put forward under Model 2 becomes less effective. The cross-talker variability seen for F3 and M1 here exemplify such a case. Even after subtracting the grand cue mean of each talker, the two sets of distributions in Figure 6C would not line up with each other as squarely as they would in Figure 6B. In other words, there will be some amount of residual uncertainty even if listeners can compensate for baseline differences across the two talkers.

The realization that talkers vary in their distributional structure led some researchers to advocate for the utility of storing (and learning) separate talker (or talker-group) specific representations of the same speech category (e.g., Chodroff & Wilson, 2017;

Kleinschmidt & Jaeger, 2016; Xie et al., 2021). Such an approach becomes most instructive especially when listeners must deal with heterogeneous, and sometimes categorically distinct, ways in which phonetic cues are mapped onto underlying categories. For instance, across variants or accents of English, a yes-no question can be produced with a rising intonation (e.g., in American English) or with a falling intonation (e.g., in British English). To accommodate such differences, compensating for baseline differences across talkers would not be sufficient. Listeners would need to make (at least temporary) adjustments in phonetic and prosodic categorization by storing and learning talker specific uses of phonetic cues to accommodate the cross-talker differences (for a related discussion in segmental speech perception, see Norris et al., 2003; 2016; Vroomen et al., 2007; Kraljic and Samuel, 2007).

### 3.2.4 Model 4: Talker-specific distribution functions

Model 4 is similar to Model 3 except that cues are *not* assumed to be normally distributed. The distributions, therefore, are not characterized with a particular set of means and variances, and their overall shapes are strongly impacted by characteristics of the individual tokens of input. Among the four models we consider here, Model 4 closely represents the underlying, "true" distributions of the data with the highest fidelity. As we mentioned for Model 3, for some talkers and categories, the distributions depicted under Model 4 could be qualitatively indistinguishable from those under Model 2. For others, these two ways to capture the underlying distributions would present starkly different pictures. In the current example, F3's productions of the narrow-focus and M1's productions of the contrastive focus show two peaks (i.e., bi-modality), which would not be captured under Model 2.

The more distinct talkers are from each other in their underlying distributional structures, the more effective it becomes for the listener to learn and store details of the input produced by a particular talker. This resonates with the core idea of the so-called "experience-based" approach in speech perception and spoken word recognition (e.g., Foulkes & Hay, 2015; Goldinger, 1996, 1998; Johnson, 2006; Pierrehumbert, 1994). This approach assumes that listeners store individual exemplars of speech input as part of episodic memory traces. For instance, listeners can store various pronunciation patterns of accent features or other characteristics and retrieve the category-talker contingency when they encounter the same talker (e.g., Goldinger, 1996; Nygaard et al., 1994; Pierrehumbert, 2001; Schweitzer, 2012; Wade, 2007). Accumulating experiences with a talker (or a talker group) can thus contribute to faster and more accurate retrieval of speech categories (Nygaard & Pisoni, 1998). Accounts vary in what exactly is stored for later retrieval and whether any generalization can occur among the stored exemplars. Shared among these experience-based accounts is the assumption that talker-variability serves as an information source that *aids* perception and comprehension rather than a source of noise to be discarded during processing.

### 3.2 Summary

Traditionally, three distinct accounts have been given to explain how listeners may navigate the problem of talker variability in prosodic processing. In an attempt to synthesize these accounts, we have considered four possible generative models that can be posited for the two focus types (Grice et al., 2017). We saw that the same set of data, derived from two German talkers, can support each of the three accounts to different degrees depending on distributional assumptions and ways in which talkers are expected to vary from each other.

Surely, these four models are not meant to form an exhaustive list, and there are many other ways in which listeners believe talkers to vary. The four models considered here nonetheless illuminated a key insight: the relative effectiveness of an approach to overcome the talker variability in *comprehension* depends on the nature and structure of talker variability in *production*. If the true, underlying distributions from the two German talkers varied in a way reducible to differences in their grand cue means or any other simple scaling factor (as assumed under Model 2 above), cue-based normalization would effectively lead to perceptual invariance. Here, however, we observed that the two talkers also varied in their distributional structures (as seen in Models 3 and 4). This predicts that effective categorization likely recruits a mechanism that learns and stores characteristics of these categories for each talker beyond their overall cue mean.

An important, open question is: *what do human listeners do?* Which of these models do they have in mind when they categorize prosodic input? Do they rely on normalization or do they learn to apply distinct distributional assumptions to each talker? In the next section, we propose an empirical approach to addressing these questions.

## 4. THINKING PROBABILISTICALLY: LINKING MODELS OF PRODUCTIONS TO MODELS OF COMPREHENSION

A major conceptual advantage of considering multiple generative models is that we can use these models to *simulate* listeners' categorization judgments to be made under each of them. Comparing these simulated judgments against human data will tell us about a type of internal model that listeners are likely entertaining when they categorize the prosodic input. In this section, we provide an overview of how this can be done.

### 4.1 Prosodic categorization as probabilistic inference based on an expected model of cue distributions

We begin by considering a hypothetical scenario in which two German talkers ("Norman" and "Hilde") encode narrow vs. contrastive focus multiple times, respectively. The left panels of Figure 7A and 7B represent distributions of onglide measures for Norman and Hilde. As was seen with F3 and M1 in Section 3, these two talkers' productions are characterized with tighter and looser clustering around the category means: Norman shows a relatively smaller amount of category overlap (Figure 7A) than Hilde (Figure 7B).

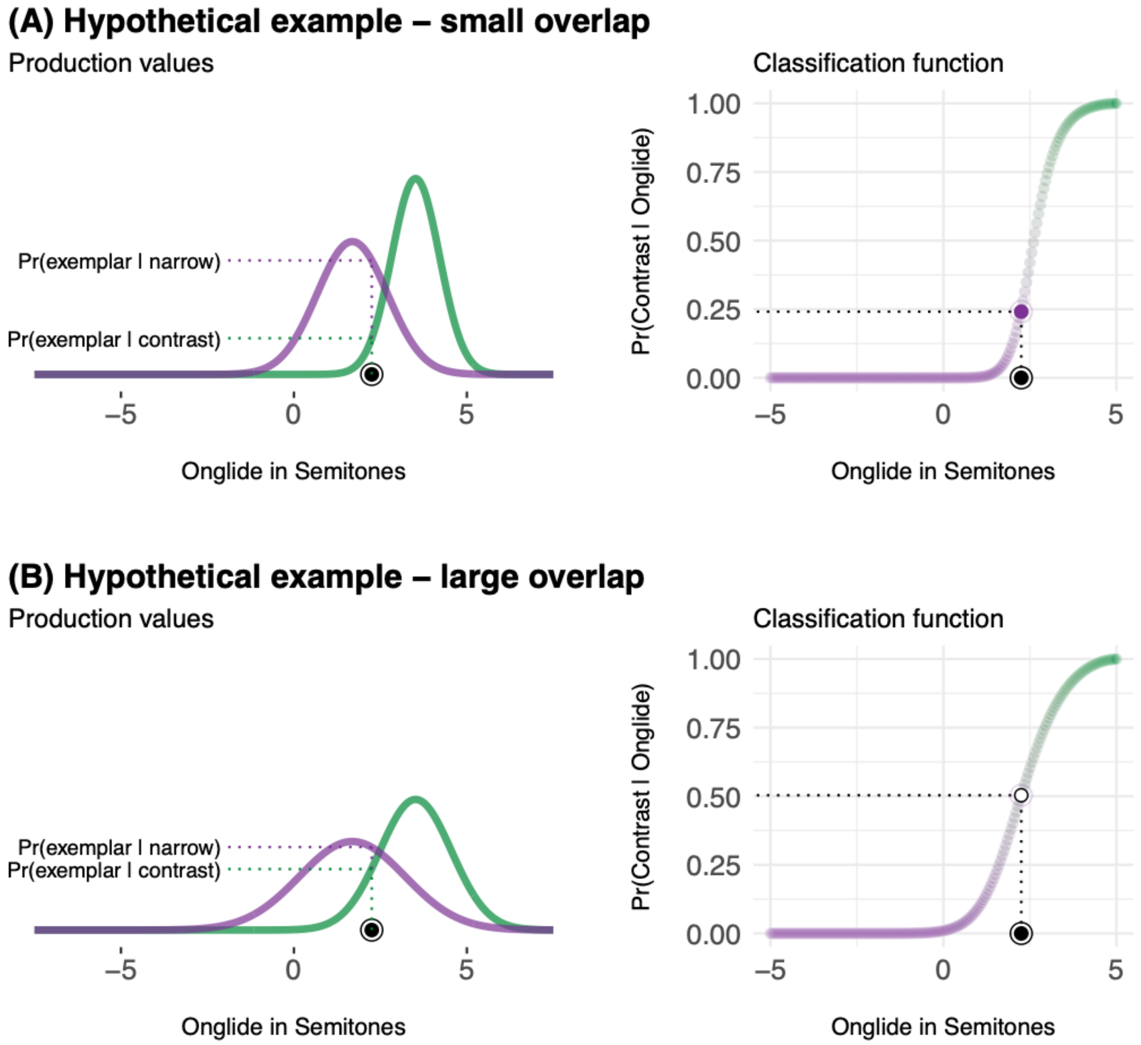

**Figure 7**: Illustration of how classification functions change as a function of phonetic cue distributions. Classification values are calculated by relating the probability of a certain onglide value being contrastive (the height of the green density curve) to the probability of that value being either contrastive or narrow focus (the sum of the heights of both purple and green density curve). (A) is a case of moderate overlap in acoustic cues between categories, resulting in a steep classification curve (top right). The bottom panel (B) illustrates the case of large overlap between categories, resulting in a flatter classification curve and a larger region of ambiguous cue values. Note that dependent on the distributional overlap, the same cue value (here 2.25) is associated with more or less certainty regarding its classification; it likely signals narrow focus in (A) but a coin toss in (B).

Now imagine a listener ("*L*") received a token of prosodic input. For the reasons we discussed in Section 2, the percept of the token is perturbed by various noise sources, and its link to an underlying category can vary across talkers and contexts. The best *L* can do, therefore, is to *infer* a category that is most likely to have generated the observed cue value according to their implicit knowledge of an underlying production (i.e., generative) model. For example, *L* might believe that they heard the onglide measure of 2.25 (indicated with a circle along the X-axis in Figure 7). *L* can then inferentially derive two probabilistic values: the likelihood with which the token is generated from the narrow-focus category (i.e. the height of the purple curve) vs. the likelihood with which the token is generated from the contrastive-focus category (the height of the green distributional curve).[8] As shown in Figure 7A and 7B, these two values can be farther apart from each other (say 76% vs. 24% for Norman, as in Figure 7A) or close together, (say 50 - 50% for Hilde, as in Figure 7B). All else being equal, wider distributions are more likely to overlap with each other. As a result, wider distributions likely result in, for any given onglide value, relatively closer estimates for the two underlying categories.

By relating the two likelihoods along possible onglide values (any point on the x-axis), one can derive a classification function of the relative probability of any given onglide value being classified as a member of the contrastive-focus category (the right panels of Figures 7A and 7B). Notice that the slope of the classification function is steeper for Norman (Figure 7A) than for Hilde (Figure 7B). The shallower slope for Hilde means that *L*'s judgments are expected to be overall more ambiguous for her productions than for Norman's (i.e., many onglide values are associated with probability values closer to 0.5). Importantly, the onglide value of 2.25 is more likely to be judged as a member of the contrastive-focus category for Hilde than for Norman. We express the probability with which a given category is inferred based on a cue value (in this case, onglide) as Pr(Contrast | Onglide). Here, Pr(Contrast | Onglide = 2.25) is higher for Hilde than for Norman.

### 4.2 Deriving ideal categorizations given the expected distributions of data

This way of reasoning laid out in 4.1 straightforwardly links *L*'s implicit model of *production* with their inference in *comprehension.* Past research has demonstrated that this way of reasoning can capture qualitative and quantitative characteristics of human speech categorization judgments (e.g., Clayards et al., 2008; Theodore & Monto, 2019). Most importantly, this approach can explain, in a mathematically principled manner, why the same physical token of input can be assigned to a different underlying category across talkers (and contexts) under distinct generative models. We use this approach to predict categorization judgments expected under Models 1-3 from Section 3.[9] We will then compare the predictions against human judgment data on Grice et al.'s production data set (reported originally by Cangemi et al. (2015)).

[8] In a Bayesian framework of inferences, which we have assumed so far, *L* is integrating these estimates into their estimates of prior probabilities of the categories. In this current example, prior probabilities are conceptualized as Pr(Contrast) and Pr(Narrow), i.e., how likely it is for the talker to produce each of these categories *a priori*. Depending on contextual and talker-specific information, *L* might *a priori* assign different probabilities to these two categories. For simplicity's sake, we assume in the current demonstration that these prior probabilities are uniformly distributed across the categories. We come back to this point in Section 5.3.

[9] Because the inference mechanism explained here assumes that an underlying distribution is Gaussian, we cannot straightforwardly apply it to Model 4. While it is possible to fit a model without the assumption of normal distributions, we do not attempt to do so in the current demonstration.

### 4.2.1 Comprehension of Model 1

Figure 8 represents the categorization function for *L* who assumes Model 1 ("$L_1$"). The two underlying distributions are characterized with wide variances and a large degree of category overlap (Figure 8A). As we noted in Section 4.1, this category overlap predicts a relatively shallow categorization function (Figure 8B), i.e., a larger amount of uncertainty, for a greater number of tokens. That is, $L_1$'s judgment will not be particularly clear-cut in most cases.

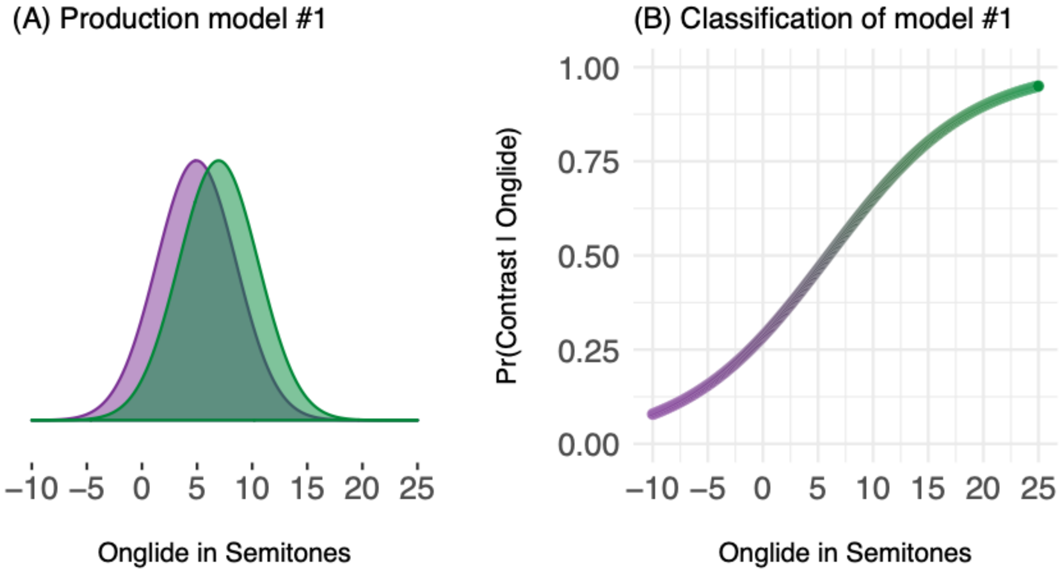

**Figure 8**. Production model #1 and its corresponding classification function. Classification Pr(Contrast | Onglide) is the probability of the intended Contrast focus given a specific Onglide value. High distributional overlap leads to a flat classification function corresponding to high uncertainty about the intended category membership.

### 4.2.2 Comprehension of Model 2

Now consider a listener who assumes Model 2 ("$L_2$"). $L_2$ can represent the input from the two distinct talkers separately in terms of their grand cue means (Figure 9A). While the two distributions are equivocal in terms of their ordinary relationships (e.g., larger onglide values for a contrastive than for a narrow focus), they vary in their baselines such that M1 produces *overall* higher onglide values for both categories compared to F3. With this knowledge, in their categorization judgments, $L_2$ can derive two distinct classification functions with distinct intercepts (with the same slope, Figure 9B).

Notice that each of the underlying distributions is characterized by a smaller variance than what was derived under Model 1 (Figure 8A). This predicts steeper categorization functions for $L_2$'s judgments than those for $L_{1.}$ Put differently, $L_2$'s judgments are expected to be more categorical as compared to those of $L_1$. and tokens of intermediate values (e.g., five semitones) are accurately associated with distinct meanings depending on a talker identity. This increased accuracy expected for $L_2$, in comparison to $L_1$*,* is part and parcel of a benefit of scaling/compensation discussed in Section 3. The more talkers differ in terms of their grand cue means, the more beneficial it is for *L* to consider such baseline differences.

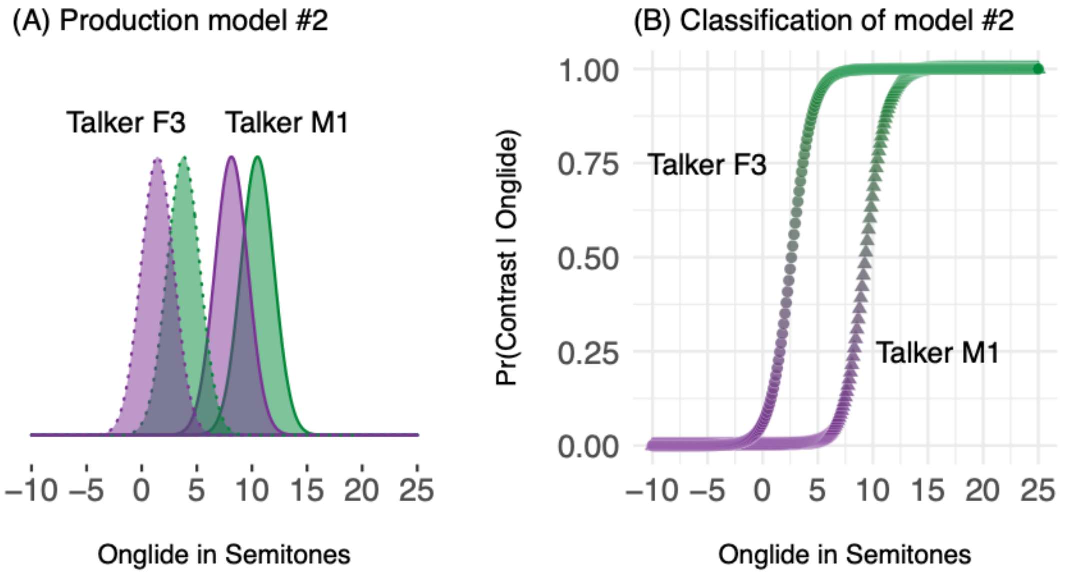

**Figure 9**. Production model #2 and its corresponding classification function. Classification Pr(Contrast | Onglide) is the probability of the intended Contrast focus given a specific Onglide value. Less overlap results in steeper classification functions which are shifted from each other

### 4.2.3 Comprehension of Model 3

Finally, consider a listener who assumes Model 3 ("$L_3$"). As in the case of $L_2$, this listener applies two distinct categorization functions for talker F3 and talker M1. Unlike in $L_2$, $L_3$ can learn and store the structure of the input distributions beyond the grand cue mean for each talker (Figure 10A). $L_3$'s categorization curves can therefore differ from each other in both their intercepts and slopes (Figure 10B). In the current case of F3 and M1's productions, $L_3$ would derive a much steeper categorization function for F3 than for M1. In other words, $L_3$ would experience a larger amount of uncertainty in their categorization judgments in response to tokens produced by M1 than to those produced by F3.

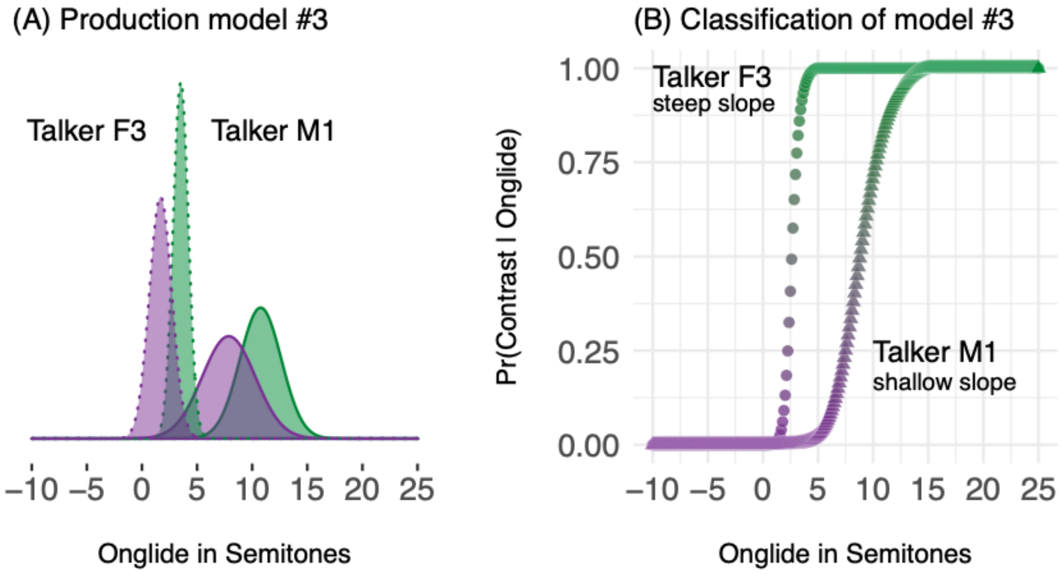

**Figure 10**. Production model #3 and its corresponding classification function. Classification Pr(Contrast | Onglide) is the probability of the intended Contrast focus given a specific Onglide value. Talkers differ in the overlap of their categories resulting in shifted classification functions that differ in their slope. Note that the classification functions are not sigmoid because the distributions have different standard deviations. The irrelevant slope is visually highlighted by transparency.

One caveat is that, compared to estimating the grand cue means (i.e., $L_S$), reliably estimating category means and variances would require exposure to some (and likely a large number of) input tokens. In particular, correctly estimating a category variance would require more tokens than for estimating a category mean. Perhaps $L_3$ may not be able to form a model until they observe some (potentially a large number of) utterances from both talkers to represent how *consistently* each talker would encode each of the focus categories. Further, $L_3$'s categorization likely relies more on memory resources as compared to that of $L_2$ due to the increased number of parameters required to characterize each of the distribution categories for each talker. It is therefore an empirical question whether human listeners can, and in fact do, derive distributional structure in this talker-specific manner.

### 4.2.4 Human listeners' judgments

To gain insight into how human listeners responded to F3 and M1's productions, we examined categorization judgment data derived from native German listeners (Cangemi et al. 2015). In this experiment, 20 native listeners were exposed to production tokens one by one and asked to categorize them into one of four different focus categories: background, broad focus, narrow focus and contrastive focus.[10] For the current purposes, we grouped the responses into two broader classes, i.e.,

[10] "Background" here corresponds to a scenario where an entire sentence provides information that is "given." "Broad focus" constitutes an utterance that presents an entire proportion as new information (in response to a question such as "What happened?").

the contrastive focus vs. everything else and then computed the relative probability of the contrastive focus category including the interaction between onglide values and talker identity (Figure 11).[11]

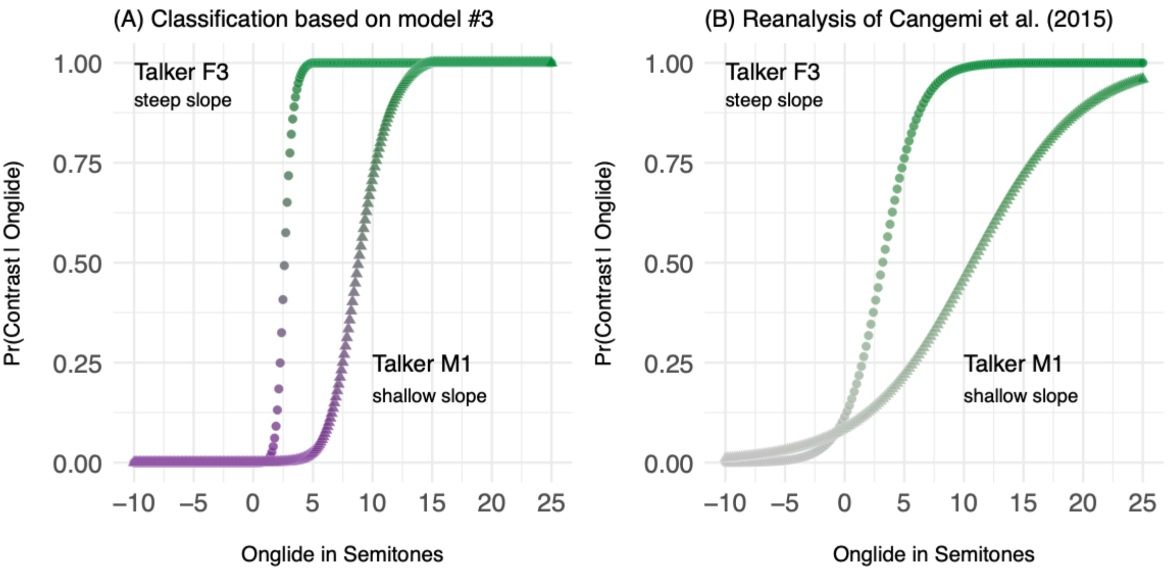

**Figure 11**. (A) Predicted categorization functions based on Model #3, where each talker's category means and variances are estimated based on the production data. (B) Human categorization judgment data (derived from a four-alternative judgment participated by 20 native listeners of German).

The resemblance between the prediction for $L_3$ (Figure 11A) and the human data (Figure 11B) is striking.[12] First of all, human listeners did select distinct categories in response to the same onglide value for F3 and M1. Clearly, the categorization functions (for each talker) are steeper as compared to what is expected for $L_1$. Second, and more specifically, human listeners seemed to condition their categorization judgments according to the grand cue means for each talker. To

[11] The four alternative forced choice format was chosen based on the original study's research question, which is different from our main question here. For exposition purposes, this paper has assumed a simplified scenario in which there are only two alternatives (e.g., Is the token part of the contrastive vs. the narrow focus category?). We note that future examination of this and related predictions can include a simpler, two-alternative task that is conceptually more consonant with the distributional assumptions we have put forward here (e.g., two alternative forced choice judgments).

[12] On first glance, the two models might look rather different. Differences between the model prediction and the human data can be ascribable to the following three sources. First, as we mentioned in Footnote 10, the model prediction considers choices between two categories whereas human listeners were given four alternative choice options. It is possible that the task increased the amount of uncertainty listeners experienced. Second, human listeners would have access to a large number of cues besides onglide. As we discuss in Section 5, other acoustic/phonetic cues could signal a category membership more or less reliably, and human responses could be sensitive to these differential levels of reliability across cues. Third, more generally, human judgments are susceptible to attentional lapses and perceptual noises.

compare, tokens of onglide value of five are more likely to be judged as part of the contrastive focus category when they were produced by F3 than when produced by M1. Finally, and most importantly, human listeners derived a steeper categorization function for F3 than for M1, which can be predictable based on the underlying distributional structure for the two talkers. As predicted for $L_3$ *but not for* $L_2$, human listeners experienced an increased amount of uncertainty in their categorization judgments in response to M1's productions as compared to F3's.

### 4.3 Summary

The fundamental assumption underlying the approach to *thinking probabilistically* is the uncertainty inherent to human perceptual and cognitive judgments. Listeners can never directly observe the cue-category mapping intended by the talker. And critically, the mapping is non-stationary, changing its precise correspondences across talkers and contexts. One way to navigate this uncertainty is to *infer* the intended cue-category mapping by integrating their implicit knowledge about how cues are generally distributed over possible underlying categories in previously encountered tokens of prosodic input.

With this approach, we predicted patterns of categorization judgments expected under each of the generative models. The categorization judgments made by native German listeners were best simulated by a model that expects listeners to represent the structure of cue distributions beyond the overall cue mean (i.e., $L_3$, Figure 11B). This supports the idea that, at least for the types of cues and categories we considered here, listeners learn and store distributional statistics characteristic to a particular talker to navigate the talker-variability in the input.

## 5. WHERE TO NEXT?

The approach to thinking probabilistically leads to a number of new questions that can be empirically tested. Here we discuss three that we believe provide fruitful avenues for future research.

### 5.1 What is the trajectory of learning?

As we saw in Section 4, listeners' categorization judgments examined here were best modeled if listeners are assumed to have access to talker-specific distributional statistics. However, much is unknown about the mechanisms with which listeners come to acquire this knowledge. In particular, it is important to ask how the learning occurs and how quickly listeners come to an accurate estimate of means and variances for a given category. If listeners must learn the distributional knowledge completely anew for each talker, and if the learning process requires a large amount of input, such a process would be of little use in our actual linguistic communication. Listeners would keep mis-categorizing input until talker-specific distributions for all relevant categories are correctly estimated. And the frequent mis-categorization would predict a protracted learning period.

Empirically, however, the listeners seem to be able to accommodate talker-specific patterns of productions relatively rapidly. In the case of the comprehension experiment by Cangemi et al. (2015) that we discussed in Section 4, listeners were exposed to the five talkers for only 12 productions (3 target words x 4 focus conditions). This means that their talker-specific categorization judgments must have been derived from this brief exposure. Logically, for the learning to be rapid and effective, listeners must have a well-calibrated set of expectations as to where their learning begins, and/or what constitute relevant parameters of the learning (Kleinschmidt & Jaeger, 2016). That is, listeners may anchor their learning in a way that a small number of tokens can usefully inform the talker-specific representations of intonational categories.

Of relevance to this idea is a recent proposal that listeners begin with a general expectation for a prototypical talker of a given language and *adapt* their expectation to characteristics of prosodic productions of a particular talker (Bosker, under review; Kurumada et al., 2017; Xie et al., 2021). This would circumvent the need to learn distributional information from scratch in each encounter to a new talker. To test this, Xie et al. (2021) created a 11-step prosodic continuum, ranging from clear statements (e.g., “It’s raining.”) to clear questions (e.g., “It’s raining?”). Two groups of subjects were exposed to distinct input statistics where acoustically ambiguous items (from a mid-point along the continuum) were disambiguated either as a question or a statement. After only 30 tokens of exposure, only 50% of which were ambiguous, those two groups of subjects categorized novel ambiguous items differently to match the input statistics (for similar findings in contrastive intonation and emotional prosody, see Kurumada et al., (2017) and Woodard et al., (2021), respectively). This resonates with the idea of so-called “perceptual learning” in speech (e.g., Norris et al., 2003; Samuel & Kraljic, 2009), whereby listeners constantly optimize their perception of speech categories according to recent exposure.

This preliminary evidence opens up a number of new questions. Chief among them are: Do listeners learn both means and variances of the underlying distributions at the same time? How long does the learned knowledge of distributions last? Can listeners track distributions for two talkers at the same time? Can we generalize learned patterns from one person to another (or from one utterance type to another)? Answering these questions requires innovations both in theoretical and methodological approaches to prosody. Related insights have only begun to be explored even in the domain of segmental speech perception, where the importance of constant learning for stable perception has been long acknowledged (e.g., Adank & Janse, 2009; Clopper & Pisoni, 2004; Eisner & McQueen, 2006; Idemaru & Holt, 2011; Kraljic & Samuel, 2011; Nygaard & Pisoni, 1998). Determining the nature and bounds of learning supporting prosodic comprehension, we believe, is a particularly important step for future theory building.

### 5.2 How many/which cues does/should a listener track?

Another class of questions concern what can be stored as part of the listener’s prosodic knowledge. Evidently, there is a large number of cues as well as their combinations that listeners can *in principle* learn and store for a given talker.[13] On the other hand, benefits arising from storing these cues must be balanced against the costs of doing so. It would be infeasible, or at least uneconomical, for the human memory system to store veridical details of all prosodic cues from all tokens produced by all talkers without any decay or abstraction. How should listeners choose what to store and what to do away with?

It is important to keep in mind that not all cues (or combination of cues) are equally useful in inferring an underlying category (Liu & Holt, 2015). Some cues are more variable than others, and some of the variability may be orthogonal to the dimension along which a talker separates categories in productions. To continue with the example of the German narrow vs. contrastive focus, there are many partially co-varying cues besides onglide that listeners can pay attention to. One prominent example is so-called the tonal alignment: the timing of an F0 peak relative to a segmental landmark such as the accented vowel. Note that onglide and alignment are closely related. The former concerns the directionality and amount of F0 movements leading towards a local F0 peak while the latter focuses on the timing of the peak (e.g., early vs. late) relative to a segmental landmark. They can, however, vary independently, and Grice et al. (2017) did in fact report on both individually.

Figure 12 illustrates the distributions of the onglide measure (on the x-axis) and those of peak alignment (on the y-axis) for all rising pitch accents.[14] We can ask how *informative* (in Kleinschmidt 2019's term) it is for listeners to track multiple cues independently and together as multivariate distributions (for detailed discussion, see Xie et al., 2021). For F3 and M1, the distributions of the alignment measures (plotted against y-axis) are largely invariant for the two categories (Figure 12B). In other words, learning and storing distributions of alignment would not add much discriminatory power over underlying categories after onglide is considered. However, talkers such as F2 or F3 distinguished between the two focus categories with both onglide and alignment (indicated with the oblongs in Figure 12B). For these talkers, learning and storing distributions of alignment values is expected to make categorization judgments proportionally more accurate.

One productive hypothesis is that listeners may have implicit knowledge about how informative each of these phonetic cues may be and how the informativity can vary across talkers (Kleinschmidt, 2019). If so, they could track and learn distributional information that is generally more diagnostic about the categories that they infer.

[13] A topic of impotence that we are not discussing in depth here is the existence and degree of individual differences across *listeners*. Roy et al., (2017), for instance, tested 32 monolingual, native listeners of American English on their judgments of prosodic boundaries and prominence. They found that these listeners, as well as trained annotators, tend to rely on a similar set of cues while magnitudes with which each cue influences their judgments varied substantially across listeners. A theory of possible ways to navigate the lack of invariance problem must encompass this variability on the listeners' side as well.

[14] Talkers also produce falling pitch accents, the distribution of which depends on the intended focus category. The relationship between categorical and gradient variation in intonation is, however, outside the scope of the present paper. See Roessig (2021) for a detailed discussion and modelling attempt.

This would reduce the overall amount of information to be stored in memory without sacrificing the robustness of comprehension.

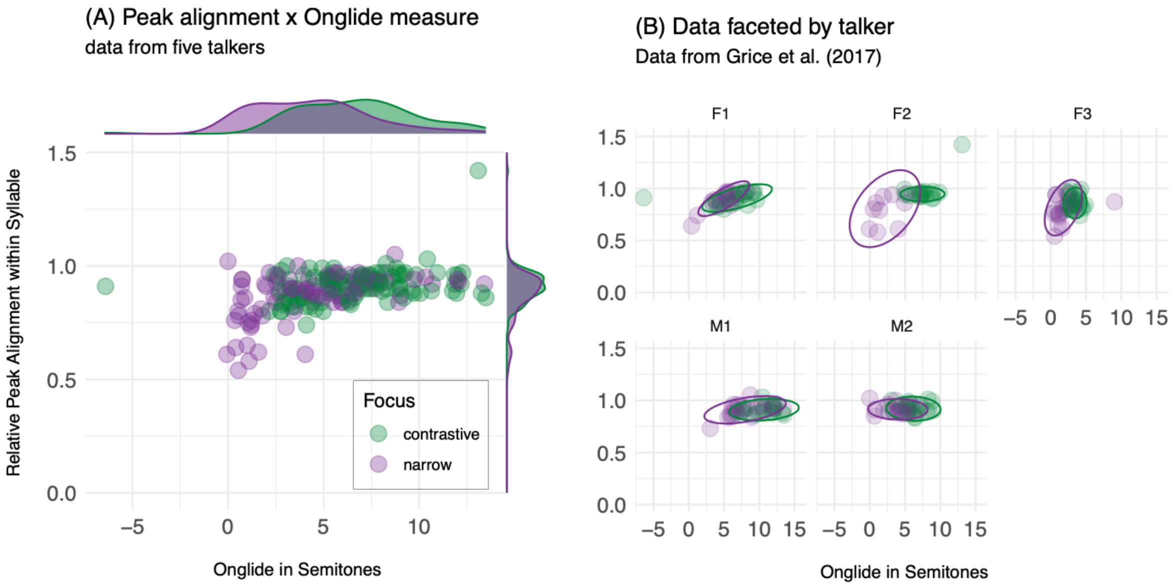

**Figure 12**. (A) Scatterplot and associated kernel density curves for the onglide measure and relative peak alignment within the syllable for all talkers for rising pitch accents; (B) Individual data points and corresponding data ellipses faceted by talkers.

In addition, this logic can be extended to assess the utility of invoking an intonational category (e.g., an H* vs. an L+H* pitch accent) during prosodic processing. We have so far postulated that listeners infer a meaning from phonetic cue distributions (i.e., estimating Pr(Contrast | Onglide), read as the probability of the contrastive focus meaning given a certain onglide value). It is, however, also possible that they first infer an intonational category from the cue distributions (e.g., Pr(L+H* | Onglide)) and then a meaning (e.g., Pr (Contrast | L+H*)). To what extent the intermediate step adds to categorization accuracy is an empirical question, which now we can quantitatively assess in the current framework (e.g., If two intonational categories are discretely separable in terms of their phonetic cue distributions, the intermediate level becomes redundant and hence adds no information).

### 5.3 What else do listeners need to learn for a specific talker?

In this paper, we focused our attention exclusively on the cue-category mapping and its across-talker variability. It, however, goes without saying that there are many other sources of information that impact the listener’s inference over what the talker must have meant.

One of such sources is the *a priori* predictability, or “prior probability” assigned to a category that can be derived as a result of the probabilistic inference. Put simply, listeners are likely expecting the talker to produce different meanings at different rates. These differential probabilities estimated *prior to* receiving actual bottom-up input can serve as an independent source of information to constrain the listener’s inference. For instance, a contrastive meaning is relatively marked (i.e., infrequent and less expected) as opposed to the meaning of “new” information conveyed by the narrow focus. When listeners receive the input that is phonetically ambiguous, these prior probabilities bias them to infer a meaning that is overall more frequent and likely (e.g. Roettger, Mahrt & Cole, 2019). The prior probabilities can be further calibrated across discourse contexts according to the knowledge shared by the talker and the listener, as well as a preceding linguistic context. For instance, “Who did Sally see?” makes a narrow focus more likely for “Sally saw Bob” above and beyond the fact that the narrow focus is overall frequent in conversation. In contrast, “Did Sally see Ted?” makes a contrastive focus more likely for “Sally saw Bob” despite the general rarity of the category.

Furthermore, listeners might also be sensitive to other types of variability in terms of the mapping between categories and meanings. It has been reported that different social and geographical accents have different distributions of accent categories over meanings. For example, African American Vernacular English has been shown to exhibit unique prosodic features (Thomas, 2007, 2015), with an overall higher usage frequency of a contrastive focus accent with a wider range of functional meanings compared to the standard variant (McLarty, 2018). Also, even within a given variant, the same prosodic category might be more or less likely to support a given meaning across talkers: some may use a particular type of pitch movement most exclusively for the contrastive meaning while others might show use patterns that are more liberal or perhaps less canonical according to the standard usage (e.g., L2 learners’ use of prosodic productions, Jackson & O’Brien, 2011).

Recent work has begun to show that listeners are in fact capable of tracking and adjusting the mapping between an intonational category and its meaning in a talker-specific manner (Roettger & Franke, 2019; Roettger & Rimland, 2020). When a given talker’s productions of contrastive focus are not reliable (i.e., not consistently signaling a contextual contrast), listeners rapidly down-weight the intonational information and cancel the inference even when they *hear* the pitch movement that is usually tied to the contrastive meaning. All in all, listeners do seem to condition their comprehension of speech prosody on who produced the input. Future work must address how such talker information may be registered and stored (e.g., at different neural substrates in speech processing) for effective communication.

## 6. CONCLUSION

This paper aimed to lay a foundation for a probabilistic approach to prosodic processing, outlining possible ways in which listeners might navigate the (subjective) invariance between them. The example case of the categorization of German narrow vs. contrastive focus marking demonstrated that listeners would likely benefit from representing distributional statistics for the categories as well as the grand cue means for each talker. A future follow-up to this observation should consist of two steps: 1) collecting more data to accurately estimate the distributional structure (including “co-variation” with other factors such as talkers and contexts) in productions and 2) testing whether, and if so, how, listeners can learn and adapt to the different means and variances observed in the data.

The logic developed here for the cross-talker variability can be extended to other sources of variability (e.g., linguistic and discourse context, social register, speech style). The core contribution of the approach to thinking probabilistically is thus to connect the variability in production to mechanisms supporting comprehension. This approach galvanizes us to elucidate the knowledge of variability that listeners may have and how they use it in probabilistic inference. Doing so will bring us one step closer to answering one of the pertinent questions in cognitive science: *How do humans understand each other through the inherently gradient and variable channel of spoken language?*

**Acknowledgments**

We thank T. Florian Jaeger, Simon Rössig, Lily Steiger and Xin Xie for their input and comments on an earlier version of the manuscript, and Sarah Bybik for her help with recording the examples plotted in Figures 1 and 2. We are further grateful that the German production data were made publicly available by Martine Grice, Simon Rössig, and Henrik Niemann.

**References**

Adank, P. (2012). The neural bases of difficult speech comprehension and speech production: Two Activation Likelihood Estimation (ALE) meta-analyses. *Brain and Language*.

Adank, P., Davis, M. H., & Hagoort, P. (2011). Neural dissociation in processing noise and accent in spoken language comprehension. *Neuropsychologia*, *50*, 70–84. https://doi.org/10.1016/j.neuropsychologia.2011.10.024

Adank, P., & Janse, E. (2009). Perceptual learning of time-compressed and natural fast speech. *Journal of Acoustical Society of America*, *126*(5), 2649–2659. https://doi.org/10.1121/1.3216914

Arvaniti, A. (2019). Crosslinguistic variation, phonetic variability, and the formation of kcategories in intonation. *In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (Eds.) Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, 1–6.

Arvaniti, A, & Garding, G. (2007). Dialectical variation in the rising accents of American English. In Jennifer Cole & J. H. Hualde (Eds.), *Papers in Laboratory Phonology 9* (pp. 547–576). Mouton de Gruyter.

Arvaniti, Amalia. (2017). The Autosegmental-Metrical model of intonational phonology. In S. Shattuck-Hufnagel & J. Barnes (Eds.), *Prosodic theory and practice*. MIT Press.

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L. N., & Banzina, E. (2019). Not just a function of function words: Distal speech rate affects perception of prosodically weak syllables. *Attention, Perception, & Psychophysics*.

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

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Beckman, M., & Ayers Elam, G. (1997). *Guidelines for ToBI labeling. Version 3.0*.

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.

Breen, M., Dilley, L. C., Kraemer, J., & Gibson, E. (2012). Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory*, *8*(2). https://doi.org/10.1515/cllt-2012-0011

Breen, M., Dilley, L. C., McAuley, J. D., & Sanders, L. D. (2014). Auditory evoked potentials reveal early perceptual effects of distal prosody on speech segmentation. *Language, Cognition and Neuroscience*, *29*(9), 1132–1146. https://doi.org/10.1080/23273798.2014.894642

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

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 306–306. https://doi.org/10.1037/a0038689

Brugos, A., Breen, M., Veilleux, N., Barnes, J., & Shattuck-Hufnagel, S. (2018). Cue-based annotation and analysis of prosodic boundary events. *9th International Conference on Speech Prosody*, 245–249.

Calhoun, S. (2004). Phonetic dimensions of intonational categories: The case of L+H* and H*. *Proceedings of Speeeh Prosody 2004*, 103–106.

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). Peter Lang.

Carlson, K., Clifton Jr., C., Frazier, L., & Clifton, C. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language*, *45*(1), 58–81. https://doi.org/10.1006/jmla.2000.2762

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

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

Clopper, C., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, *47*(3), 207–238. https://doi.org/10.1177/00238309040470030101

Cohen, A., Collier, R., & ’t Hart, J. (1982). Declination: Construct or Intrinsic Feature of Speech Pitch? *Phonetica*, *39*, 254–273.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*(1–2), 1–31.

Cole, Jennifer, Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, *1*(2), 425–452. https://doi.org/10.1515/labphon.2010.022

Cole, Jennifer, & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology*, *7*(1), 8–8. https://doi.org/10.5334/labphon.29

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, *77*, 2142–2156.

Cutler, A. (1990). Exploiting prosodic probabilites in speech segmentation. *Cognitive Models of Speech Processing*.

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

Dahan, D, & Bernard, J.-M. (1996). Interspeaker variability in emphatic accent production in French. *Language and Speech*, *39*(4), 341–374.

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

Dahan, Delphine, Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition*, *108*(3), 710–718. https://doi.org/10.1016/j.cognition.2008.06.003

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., Ladd, D. R., & Schepman, A. (2005). Alignment of L and H in bitonal pitch accents: Testing two hypotheses. *Journal of Phonetics*, *33*(1), 115–119. https://doi.org/10.1016/j.wocn.2004.02.003

Dilley, L. C., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, *63*(3), 274–294. https://doi.org/10.1016/j.jml.2010.06.003

Dilley, L. C., Millett, A., McAuley, J. D., & Bergeson, T. R. (2014). Phonetic variation in consonants in infant-directed and adult-directed speech: The case of regressive place assimilation in word-final alveolar stops. *Journal of Child Language*, *41*(1), 155–175. https://doi.org/10.1017/s0305000912000670

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

Eckert, P. (2016). Variation, meaning and social change. *Sociolinguistics: Theoretical Debates*, 68–68.

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.

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental Psychology*, *27*, 209–221.

Féry, C., & Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *Journal of Phonetics*, *36*, 680–703.

Foulkes, P., Docherty, G., & Watt, D. (2005). Phonological variation in child-directed speech. *Language*, *81*(1), 177–206.

Foulkes, P., & Hay, J. B. (2015). The emergence of sociophonetic structure. *The Handbook of Language Emergence*, 292–313. https://doi.org/10.1002/9781118346136.ch13

Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *Journal of the Acoustical Society of America*, *119*(3), 1712–1726.

Frazier, L., Carlson, K., & Clifton, C. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, *10*(6), 244–249.

Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, *27*, 765–768.

Fuchs, S., Pape, D., Petrone, C., Perrier, P., & }. (2015). Listener-specific perception of speaker-specific productions in intonation. In *Individual Differences in Speech Production and Perception*. Peter Lang. https://doi.org/10.3726/978-3-653-05777-5/14

Goldinger, S D. (1996). 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, Stephen D. (1996). 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, Stephen D. (1998). Echoes of echoes?: An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

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

Graf Estes, K., & Hurley, K. (2013). Infant-directed prosody helps infants map sounds to meanings. *Infancy*, *18*(5), 797–824. https://doi.org/10.1111/infa.12006

Gregory, R. L. (2015). *Eye and Brain: The Psychology of Seeing—Fifth Edition*. Princeton University Press. https://books.google.com/books?id=MYgVBgAAQBAJ

Grice, M. (1995). Leading tones and downstep in English. *Phonology*, *12*, 183–233.

Grice, M., Baumann, S., & Benzmüller, R. (2005). German Intonation in Autosegmental-Metrical Phonology. In *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press.

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

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

Gussenhoven, C., & Rietveld, A. C. M. (1988). Fundamental frequency declination in Dutch: Testing three hypotheses. *Journal of Phonetics*, *16*, 355–369.

Halliday, M. A. K. (1967). *Intonation and Grammar in British English*. Mouton.

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

Heffner, C., Dilley, L. C., McAuley, J. D., & Pitt, M. A. (2013). When cues combine: How distal and proximal acoustic cues are integrated in word segmentation. *Language and Cognitive Processes*, *28*(9), 1275–1302. https://doi.org/10.1080/01690965.2012.672229

Holliday, N. R. (2019). Variation, race, and multiracial identity in linguistic research. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(1), e1480–e1480. https://doi.org/10.1002/wcs.1480

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology. Human Perception and Performance*, *37*(6), 1939–1956. https://doi.org/10.1037/a0025641

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.

Ito, K., Turnbull, R., & Speer, S. R. (2017). Allophonic tunes of contrast: Lab and spontaneous speech lead to equivalent fixation responses in museum visitors. *Laboratory Phonology*, *8*(1). https://doi.org/10.5334/labphon.86

Jackson, C. N., & O'Brien, M. G. (2011). The interaction between prosody and meaning in second language speech production. *Die Unterrichtspraxis/Teaching German*, *44*(1), 1–11. https://doi.org/10.1111/j.1756-1221.2011.00087.x

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

Jones, P., Moore, D., & Amitay, S. (2013). Reduction of internal noise in auditory perceptual learning. *Journal of the Acoustical Society of America*, *133*(2), 970–981. https://doi.org/10.1121/1.4773864

Jun, S.-A. (2005). *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

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., & 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*.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. https://doi.org/10.1016/j.tins.2004.10.007

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

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, Chigusa, 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

Ladd, D R. (1988). Declination “reset” and the hierarchical organization of utterances. *Journal of the Acoustical Society of America*, *84*(2), 530–544.

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

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

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), 104328–104328. https://doi.org/10.1016/j.cognition.2020.104328

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461.

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

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

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020).

EARSHOT: A minimal neural network model of incremental human speech recognition. *Cognitive Science*, *44*(4), e12823–e12823. https://doi.org/10.1111/cogs.12823

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

Mattys, S. L. (1997). *The use of time during lexical processing and segmentation: A review*. *4*(3), 310–329.

McLarty, J. (2018). African American language and European American English intonation variation over time in the American South. *American Speech*, *93*(1), 32–78. https://doi.org/10.1215/00031283-6904032

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

McQueen, J. M., & Dilley, L. C. (2020). Chapter 34. Prosody and spoken-word recognition. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody*. OUP Oxford.

Miller, J. L., Aibel, I. L., & Green, K. (1984). On the nature of rate-dependent processing during phonetic perception. *Perception and Psychophysics*, *35*, 5–15.

Mücke, D., Grice, M., Becker, J., & Hermes, A. (2009). Sources of variation in tonal alignment: Evidence from acoustic and kinematic data. *Journal of Phonetics*, *37*(3), 321–338. https://doi.org/10.1016/j.wocn.2009.03.005

Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. Generative classifiers: A comparison of logistic regression and naive Bayes. *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, 841–848.

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

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*. MIT.

Pierrehumbert, J. B. (1994). Knowledge of variation. In *Papers from the parasession on variation, 30th meeting of the Chicago Linguistics Society* (Vol. 2, pp. 232–256). Chicago Linguistic Society. http://blog.deming.org/2012/10/knowledge-of-variation/

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

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).

Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, and Psychophysics*, *78*(1), 334–345.

Podesva, R., & Callier, P. (2015). Voice quality and identity. *Annual Review of Applied Linguistics*, *35*, 173–194. https://doi.org/10.1017/S0267190514000270

Podesva, R. J. (2011). Salience and the social meaning of declarative contours: Three case studies of gay professionals. *Journal of English Linguistics*, *39*(3), 233–264.

Ponsot, E., Burred, J. J., Belin, P., & Aucouturier, J. J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(15), 3972–3977. https://doi.org/10.1073/pnas.1716090115

Reinisch, E., Jesse, A., & McQueen, J. M. (2011a). 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., Jesse, A., & McQueen, J. M. (2011b). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 978–996. https://doi.org/10.1037/a0021923

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

Ritter, S., & Grice, M. (2015). The Role of Tonal Onglides in German Nuclear Pitch Accents. *Language and Speech*, *58*(1), 114–128. https://doi.org/10.1177/0023830914565688

Roessig, S., Mücke, D., & Grice, M. (2019). The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PLOS ONE*, *14*(5), e0216859. https://doi.org/10.1371/journal.pone.0216859

Roessig, S. (2021). Categoriality and continuity in prosodic prominence. Language Science Press: Berlin.

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

Roettger, T. B., Mahrt, T., & Cole, J. (2019). Mapping prosody onto meaning–the case of information structure in American English. *Language, Cognition and Neuroscience*, *34*(7), 841-860. https://doi.org/10.1080/23273798.2019.1587482

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

Roy, J., Cole, J., & Mahrt, T. (2017). Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1).

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically-modulated sub-phonetic variation on lexical competition. *Cognition*, *105*, 466–476.

Samuel, A G, & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207

Samuel, Arthur G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language*, *111*(November 2019), 104070–104070. https://doi.org/10.1016/j.jml.2019.104070

Sasha Calhoun. (2010). The centrality of metrical structure in signaling information structure: A probabilistic perspective. *Language*, *86*(1), 1–42. https://doi.org/10.1353/lan.0.0197

Schafer, A., Speer, S. R., Warren, P., & White, D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, *29*(2), 169–182.

Schweitzer, A. (2019). Exemplar-theoretic integration of phonetics and phonology: Detecting prominence categories in phonetic space. *Journal of Phonetics*, *77*, 100915–100915. https://doi.org/10.1016/j.wocn.2019.100915

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

Silverman, K., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C. W. S., Price, P., & Hirschberg, J. (1992). ToBI: A standard scheme for labeling prosody. *Proceedings of the 2nd International Conference on Spoken Language Processing*, 867–879.

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

Snedeker, J., & Trueswell, J. C. (2003). Memory and language using prosody to avoid ambiguity: Effects of speaker awareness and referential context. *Journal of Memory and Language*, *48*(1), 103–130.

Soderstrom, M., Seidl, A., Kemler Nelson, D. G., & Jusczyk, P. W. (2003). The prosodic bootstrapping of phrases: Evidence from prelinguistic infants. *Journal of Memory and Language*, *49*, 249–267.

Speer, S. R., & Ito, K. (2009). Prosody in first language acquisition: Acquiring intonation as a tool to organize information in conversation. *Linguistics and Language Compass*, *3*, 90–110. https://doi.org/10.1111/j.1749-818X.2008.00103.x

Speer, S. R., Warren, P., & Schafer, A. J. (2011). Situationally independent prosodic phrasing. *Laboratory Phonology*, *2*(1), 35–98. https://doi.org/10.1515/labphon.2011.002

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1074–1095. https://doi.org/10.1037/0096-1523.7.5.1074

Summers, W. V., Pisoni, D. B., Bernacki, R. H., Pedlow, R. I., & Stokes, M. A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *The Journal of the Acoustical Society of America*, *84*(3), 917–928.

Syrdal, A. K., & McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. *Proceedings of the ICSLP-20003*, *3*, 235–238.

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., & 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

Thomas, E. R. (2007). Phonological and Phonetic Characterstics of AAVE. *Language and Linguistics Compass*, *1*(5), 451–475. https://doi.org/10.1111/j.1749-818x.2007.00029.x

Thomas, E. R. (2015). Prosodic features of African American English. In *The Oxford Handbook of African American Language* (pp. 420–438).

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, *85*(4), 1699–1707. https://doi.org/10.1121/1.397959

Tomlinson, J. M., Gotzner, N., & Bott, L. (2017). Intonation and pragmatic enrichment: How intonation constrains ad hoc scalar inferences. *Language and Speech*, *60*(2), 200–223. https://doi.org/10.1177/0023830917716101

Turnbull, R., Royer, A. J., Ito, K., Speer, S. R., & Turnbull, R. (2017). Prominence perception is dependent on phonology, semantics, and awareness of discourse. *Language, Cognition and Neuroscience*, *0*(0), 1–17. https://doi.org/10.1080/23273798.2017.1279341

Wade, T. (2007). Implicit rate and speaker normalization in a context-rich phonetic exemplar model. *ICPHS*.

Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes*, *25*(7–9), 905–945. https://doi.org/10.1080/01690961003589492

Ward, N. (2019). *Prosodic patterns in English conversation*. Cambridge University Press.

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

Watson, D. G., Tanenhaus, M. K., & Gunlogson, C. A. (2008). Interpreting pitch accents in on-line comprehension: H* vs. L+H*. *Cognitive Science*, *32*(7), 1232–1244.

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.

Wightman, C. W. (2002). ToBI or not ToBI? *Proceedings of Speech Prosody*, 25–29.

Woodard, K., Plate, R. C., Morningstar, M., Wood, A., & Pollak, S. D. (2021). Categorization of Vocal Emotion Cues Depends on Distributions of Input. *Affective Science*. https://doi.org/10.1007/s42761-021-00038-w

Xie, X., Buxó-Lugo, A., & Kurumada, C. (2021). Encoding and decoding of meaning through structured variability in speech prosody. *Cognition*.

Xu, Y. (2006). Speech prosody as articulated communicative functions. In *Department of Phonetics and Linguistics*. Unversity College London.

Xu, Y. I. (1997). Contextual tonal variations in Mandarin. *Journal of Phonetics*, *25*, 61–83.

## Further Reading

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, *30*(1–2), 1–31.

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

McQueen, J. M., & Dilley, L. C. (2020). Chapter 34. Prosody and spoken-word recognition. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody*. OUP Oxford.