Journal of Phonetics 82 (2020) 100977

Contents lists available at ScienceDirect

Journal of Phonetics

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

Special Issue: Integrating Phonetics and Phonology, eds. Cangemi & Baumann

Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription

Jason Bishop [a,b,*], Grace Kuo [c], Boram Kim [a,d]

[a] The CUNY Graduate Center, New York, NY 10016, USA
[b] The College of Staten Island-CUNY, Staten Island, NY 10314, USA
[c] National Taiwan University, Taipei City 10617, Taiwan
[d] Haskins Laboratories, New Haven, CT 06511, USA

ARTICLE INFO

Article history:
Received 3 April 2018
Received in revised form 15 March 2020
Accepted 17 March 2020

ABSTRACT

The present study investigated the perception of phrase-level prosodic prominence in American English, using the Rapid Prosody Transcription (RPT) task. We had two basic goals. First, we sought to examine how listeners' subjective impressions of prominence relate to phonology, defined in terms of Autosegmental-Metrical distinctions in (a) pitch accent status and (b) pitch accent type. Second, and in line with this special issue, we sought to explore how phonology might mediate the effects of other cues to prominence, both signal-based (acoustic) and signal-extrinsic (stimulus and listener properties) in nature. Findings from a large-scale RPT experiment ($N = 158$) show prominence perception in this task to vary significantly as a function of phonology; a word's perceived prominence is significantly dependent on its accent status (unaccented, prenuclear accented, or nuclear accented) and to a slightly lesser extent, on pitch accent type (L*, !H*, H*, or L+H*). In addition, the effects of other known cues to prominence—both signal-based acoustic factors as well as more "top-down" signal-extrinsic factors—were found to vary systematically depending on accent status and accent type. Taken together, the results of the present study provide further evidence for the complex nature of prominence perception, with implications for our knowledge of prosody perception and for the use of tasks like RPT as a method for crowdsourcing prosodic annotation.

1. Introduction

1.1. Overview

In the present study we investigated the perception of phrase-level prosodic prominence in American English, using the Rapid Prosody Transcription task (Cole, Mo, & Hasegawa-Johnson, 2010; Cole, Mo, & Baek, 2010). Of particular interest to us was how listeners' subjective impressions of prominence at this relatively macroscopic level relate to (a) *phonology*, (b) *phonetic realization*, and (c) *signal-extrinsic factors*. By *phonology*, we mean the linguistic contrasts related to accentuation within the Autosegmental-Metrical framework (Pierrehumbert, 1980; Gussenhoven, 1984; Beckman & Pierrehumbert, 1986; Ladd, 1996, 2008; see Arvaniti, to appear, for a recent overview), and more specifically, the categories available within the Tones and Break Indices (ToBI) conventions for Mainstream American English (MAE_ToBI; Beckman & Hirschberg, 1994; Beckman & Ayers Elam, 1997). By *phonetic realization*, we mean the gradient acoustic correlates of prominence found in physical speech output. Finally, by *signal-extrinsic factors*, we mean the (non-phonological) "top-down" properties of stimuli and of listeners themselves (i.e., individual differences) that serve as predictors of perceived prominence.

The paper proceeds as follows. First, we discuss some preliminaries to the study of prominence perception in English, as well as the motivations for our study in particular (Section 1.2). We then describe some of the features of the methodology utilized in our investigation, and the specific research questions to be explored (Section 1.3). Following these introductory sections, we present a RPT experiment in English (Section 2), which is followed by a discussion of the findings (Section 3) and concluding remarks (Section 4).

* Corresponding author at: City University of New York, 365 Fifth Avenue, New York, NY 10016, USA. Fax: +1 212 817 1526.
E-mail address: jbishop@gc.cuny.edu (J. Bishop).

https://doi.org/10.1016/j.wocn.2020.100977
0095-4470/

1.2. Predicting perceived prominence

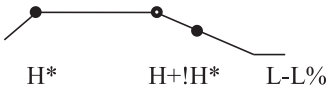
A wealth of studies in recent years have attempted to identify the factors that predict, for English and closely related languages, listeners' impressions that some words stand out as stronger than others (e.g., Baumann, 2006, 2014; Bishop, 2012a, 2013, 2016; Kimball & Cole, 2014, 2016; Streefkerk, Pols, & ten Bosch, 1997; Eriksson, Thunberg, & Traunmüller, 2001; Kochanski, Grabe, Coleman, & Rosner, 2005; Wagner, 2005; Mo, 2008; Jagdfeld & Baumann, 2011; Röhr & Baumann, 2011; Arnold, Möbius, & Wagner, 2011; Baumann & Riester, 2012; Mahrt, Cole, Fleck, Hasegawa-Johnson, 2012; Cole, Mahrt, & Hualde, 2014; Pintér, Mizuguchi, Tateishi, 2014; Baumann & Röhr, 2015; Cole, Hualde, Eager, Mahrt, 2015; Erickson, Kim, Kawahara, Wilson, Menezes, Suemitsu, Moore, 2015; Mixdorff et al., 2015; Baumann, Niebuhr, & Schroeter, 2016; Hualde, Cole, Smith, Eager, Mahrt, & de Souza, 2016 (see also Cole et al., 2019, this Special Issue); Cole, Mahrt, & Roy, 2017; Niebuhr & Winkler, 2017; Roy, Cole, Mahrt, 2017; Turnbull, Royer, Ito, & Speer, 2017; see also relevant earlier research that had somewhat different goals; Terken, 1991, 1994; Rietveld & Gussenhoven, 1985; Turk & Sawusch, 1996; Gussenhoven, Repp, Rietveld, Rump, & Terken, 1997; Cambier-Langeveld & Turk, 1999). In this lively body of work, one surprising lack of consensus has persisted that we think helps motivate an investigation into the role that phonology plays in how people perceive prominence. Across studies of English and closely related Dutch and German, we find strong agreement that fundamental frequency (F0), duration, and intensity (the latter two both contributing to the percept of "loudness") are the primary correlates of both phonological and perceived prominence, but considerable disagreement regarding their importance relative to each other.[1] On the one hand, many studies, especially those in which F0 was manipulated experimentally, have reported changes in perceived prominence to be tightly tied to changes in F0 (e.g., Ladd, Verhoeven, & Jacobs, 1994, and Ladd & Morton, 1997, for English; Rietveld & Gussenhoven, 1985, for Dutch; and Niebuhr & Winkler, 2017, for German). Others—all of which seem to utilize corpus materials—have found F0 to be weakly related to perceived prominence, if related at all (Kochanski et al., 2005; Cole, Mo, & Hasegawa-Johnson, 2010, Cole et al., 2017; but see Mahrt, Cole, Fleck, & Hasegawa-Johnson, 2012). Thus, despite high levels of interest in the matter, confusion remains regarding one of the most basic questions about prominence perception that can be asked.

We point to this particular situation because it highlights a fundamental problem with attempting to model prominence perception directly from quantitative acoustic measures, as many studies have attempted to do, and the fact that corpus-style analyses are the ones that fail to detect listeners' sensitivity to F0 is perhaps not surprising. To see why, consider the case of a regression model designed to predict prominence perception for words on the basis of particular F0 values (rather than F0 turning points or some other more discrete event) applied to the utterances in (1) through (3). The utterance in (1) features a well-known ambiguity regarding the accent status of a word positioned between two H* accents (Beckman, 1996), a familiar challenge to human ToBI transcribers who must make a categorical decision about the word's phonological prominence in the absence of a distinctive F0 target (Beckman & Ayers Elam, 1997). However, it is also problematic for a statistical model designed to predict prominence perception as a function of F0 values at particular time points, since (even allowing for some "sagging" in the transition between the two pitch accents) unaccented *bought* and unaccented *a* will have F0 values very close to those of accented *Marty* and accented *motorcycle*. However, it seems unlikely this roughly-equal F0 will result in unaccented *bought* being as perceptually prominent as the two accented words flanking it. The utterance in (2) presents similar difficulties for prediction based on acoustic measures of F0, but for slightly different reasons. Here, the unaccented article *a* will necessarily have a higher F0 value than nuclear accented *motorcycle*, though this article is specifically not accented, and thus is also unlikely to be perceived as particularly prominent. Finally, while F0 does correlate with the accentual status—and most likely, perceived prominence—of *motorcycle* in (3a) versus (3b), this is not the case for *motorcycle* in (3b) versus (3c). While *motorcycle* is marked by similarly low F0 in both (3b) and (3c), this is for different reasons—interpolation in (3b) but pitch accent in (3c). The point here is that there are many prosodic structures—perhaps most—that serve to weaken an overall correlation between particular F0 values and accentual structure. This, in turn, would have the effect of also weakening the relationship between particular F0 values and perceived prominence in statistical analyses of corpora.

(1)

MARty bought a MOtorcycle.

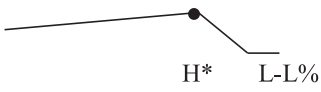
(2)



MARty bought a MOtorcycle.

(3)

a.

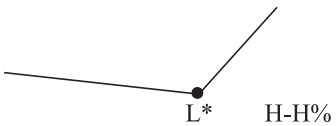


Marty bought a MOtorcycle.

b.

MARty bought a motorcycle.

c.



Marty bought a MOtorcycle?

[1] See also work on other phonetic cues to prominence in Western Germanic languages in Sluijter and Van Heuven (1996), Terken and Hermes (2000) and Epstein (2002). Cues such as voice quality, for example, while surely more marginal in their importance to perceived prominence, are also far less well studied.

In most situations, we think the listener is probably much more like a ToBI transcriber than like the regression model just described, at least in the sense that she interprets most fine-grained phonetic variation only after having performed a parse of the signal into a coarser, more discrete sequence of categorical events. Once a basic phonological structure is thus established—based on, for example, identification of F0 targets and meaningful alignments with the text—the listener can then identify residual acoustic variation. This acoustic variation may be very rich and informative indeed (Cangemi & Grice, 2016), but it is assigned significance mostly in relation to the structure that the listener has already constructed. One consequence of this alternative, and we think much more plausible scenario, is that gradient variation along acoustic dimensions like F0, duration, and intensity will influence prominence perception, *but do so in largely category-specific* ways. That is, acoustic variation will be treated differently by the listener depending on whether the associated word is accented or unaccented, whether it bears a H* or a L*, and so on. This implies that the best models of prominence perception will therefore need to either (a) be applied to data for which the phonological structure is already known, or (b) include algorithms for assigning that structure automatically (e.g., Rosenberg, 2009). Despite the important and multifaceted role that phonology likely plays in prominence perception, it has received relatively little attention. For this reason, investigating its role was a primary goal for us, which we return to in Section 1.3, where we discuss our research questions more specifically.

In addition to phonology, however, subjective judgments of prominence by human listeners reflect other information not found directly in the acoustic signal, and the effect of these additional signal-extrinsic (or "top-down") factors should also be of interest. For example, a clear finding from previous work is that impressions of prominence are inversely related to a word's lexical frequency (Cole, Mo, & Hasegawa-Johnson, 2010; Bishop, 2013; Baumann, 2014; Cole et al., 2017; see also Nenkova et al., 2007) and, to a lesser extent, its number of previous occurrences in the discourse (Cole, Mo, & Hasegawa-Johnson, 2010). It has also been found that words tend to be perceived as more prominent if they occur finally in a prosodic phrase (e.g., Rosenberg, Hirschberg, & Manis, 2010; Jagdfeld & Baumann, 2011; Cole et al., 2017). While it is somewhat unclear what the underlying mechanisms for these effects are, they have more to do with listeners' expectations about the signal than with the signal itself.

Properties of individual listeners—i.e. individual differences—are another source of signal-extrinsic effects. While very little is known about the factors that underlie individual differences in prosody perception, such differences are certainly known to exist (e.g., Cole, Mo, & Hasegawa-Johnson, 2010; Bishop, 2016; Roy, Cole, & Mahrt, 2017; Cole et al., 2017; Baumann & Winter, 2018). Recently, Jun and Bishop (2015) have suggested that at least some such cross-listener variation in prominence perception may be related to individual differences in "cognitive processing styles" (e.g., Ausburn & Ausburn, 1978; for a recent introduction in the context of phonetic research, see Yu, 2013). Following preliminary findings reported by Bishop (2012b; superseded by Bishop, 2017), Jun and Bishop argue that listeners' attention to prosodic prominence may be in part affected by individual differences in so-called "autistic traits", an aspect of cognitive processing style that has been implicated in other well-established speech perception phenomena (e.g., Yu, 2010, 2016; Stewart & Ota, 2008; Ujiie, Asai, & Wakabayashi, 2015). In particular, Jun and Bishop found that listeners who gave more autistic-like responses on the "communication" subscale of the Autism Spectrum Quotient (AQ; Baron-Cohen, Wheelwright, Hill, Raste, Plumb, 2001) were less likely to comprehend syntactically ambiguous sentences based on accentual patterns (i.e., a smaller influence of what Schafer, Carter, Clifton, & Frazier, 1996, referred to as "Focus Attraction"). Consistent with this, English-speaking listeners who give more autistic-like responses appear to be less sensitive to the presence of prenuclear prominences in online lexical processing (Bishop, 2017) and in off-line sentence completion tasks (Hurley & Bishop, 2016). Notably, the communication subscale of the AQ has been used in other work to estimate individual differences in "pragmatic skill" (Kulakova & Nieuwland, 2016; Nieuwland, Ditman, & Kuperberg, 2010; Xiang, Grove, & Giannakidou, 2013; Yang, Minai, & Fiorentino, 2018). Although we acknowledge that the relationship between this measure and an underlying construct like "pragmatic skill" (and related ones, like "Theory of Mind") remains a hypothesis, to the extent that such a relationship exists, the communication subscale of the AQ may serve as a rough proxy for listeners' sensitivity to the relation between prosody and meaning-in-context. This is relevant to prominence perception, since most off-line tasks have shown listeners' responses to reflect, in part, expectations based on contextual meaning (e.g., Vainio & Järvikivi, 2006; Bishop, 2012; Cole, Mahrt, & Hualde, 2014; Turnbull et al., 2017; see also Gussenhoven, 2015, for an even stronger claim). However, it is likely that these meaning-dependent cues have their effects in phonologically-dependent ways, much like the effects of other types of other signal-extrinsic cues have been argued to (Turnbull et al., 2017).[2] Thus signal-extrinsic factors tied to the stimulus and to the listener can be identified as another way in which prominence perception is rich and complex, and importantly for our purposes, best understood in the context of a phonological model of the to-be-perceived speech material.

1.3. Present study: exploring prominence perception using Rapid Prosody Transcription

As is clear from the above discussion, many different cues—and types of cues—contribute to prominence perception by human listeners. The overarching goal of the present study was to explore the role of phonology in prominence perception, and to do so (a) in the context of AM theory, and (b) using Rapid Prosody Transcription. Rapid Prosody Transcription (RPT) is a promising new method for exploring the perception of prosody (Cole, Mo, & Hasegawa-Johnson, 2010; Cole et al., 2010b) and for "crowdsourcing" prosodic annotation (Buhmann et al., 2002; Mahrt, 2016; Cole & Shattuck-Hufnagel, 2016; Hasegawa-Johnson, Cole, Jyothi, 2015; Cole et al., 2017). RPT involves the speeded identification of

[2] On this point, see also Calhoun (2006) statistical modeling of prenuclear versus nuclear accents in English, although the focus there is on production rather than perception.

coarsely-defined prosodic events—namely "prominence" and "juncture"—carried out by groups of linguistically-untrained listeners. While RPT has been used to study prominence perception in a number of languages/language varieties (e.g., Smith, 2009; Smith & Edmunds, 2013; Luchkina, Puri, Jyothi, & Cole, 2015; Hualde et al., 2016; see also Cole et al., 2019, this special issue) and under various listening conditions (Cole et al., 2014, 2017), this work has primarily focused on the role that acoustic and non-phonological signal-extrinsic factors play. We therefore asked two basic questions regarding how listeners' prominence perception in RPT relates to phonology, neither of which has been fully explored in English previously:[3]

(1) How do the following phonological distinctions relate to patterns of perceived prominence?
   a. Accent status (a word's status as unaccented, prenuclear pitch accented, or nuclear pitch accented)
   b. Pitch accent type (the particular tone assigned to a pitch accented syllable, e.g., L*, !H*, H*, L+H*, etc.)
(2) Do accent status and accent type mediate the effects of other (signal-based and signal-extrinsic) cues?

The first question is largely confirmatory, in that hypotheses can be straightforwardly derived from theory, and, to some extent, from previous empirical findings from closely-related German (Baumann & Röhr, 2015; Baumann & Winter, 2018; Baumann, 2014). In particular, we predicted that perceived prominence would pattern much like metrical/phonological prominence; listeners should be significantly more likely to judge nuclear accented words as prominent than prenuclear accented words, which in turn should be significantly more likely to be judged as prominent than unaccented words. Similarly, we predicted that perceived prominence should vary as a function of pitch accent type, and our prediction here was based on a characterization of accent type that emphasizes relative pitch level/height. We predicted that words bearing L+H* (which are known to have increased F0; e.g., Bartels & Kingston, 1994) should be most likely to be perceived as prominent by listeners, followed by those with a H* target, followed by those with a !H* target, followed by a L* target. Our basic justification for elevating the importance of level in the present study is based in part on work related to intonational meaning discourse function in English (Hirschberg, Gravano, Nenkova, Sneed, & Ward, 2007; Pierrehumbert & Hirschberg, 1990) and on previous perception work that seems to show level-based perceptual differences between some MAE_ToBI pitch accent categories (Turnbull et al., 2017; see also discussion in Ladd, 1994, Ayers, 1996, and Ladd & Schepman, 2003).[4]

The second question acknowledges the possibility that phonology interacts with other cues (Turnbull et al., 2017) and this question has both confirmatory and exploratory aspects to it. On the confirmatory side, and given our discussion in Section 1.2, we predict that the effect of phonetic cues should not be spread evenly across phonological contrasts. We present as more exploratory the details of how different cues might be weighted in relation to these contrasts, though even here there are some general predictions that can nonetheless be pointed out. For example, in the case of accent status we might assume that F0 will be most useful to cueing perceived prominence for words that are accented—that is, for words that are aligned with an F0 target rather than on the interpolation line (recall our discussion of example (1), above). Conversely, it seems plausible that signal-extrinsic factors like lexical frequency or individual differences in pragmatic skill might have their strongest effects on the perceived prominence of *unaccented* words, given the more ambiguous phonetic and phonological cues to words parsed into metrically weak positions. However, we know of no study that has demonstrated any such phonologically-dependent asymmetries in the importance of various cues to perceived prominence, so we regard the details of our question here to be largely exploratory in nature. We now turn to the RPT experiment that investigated these issues.

2. Experiment

2.1. Methods

2.1.1. Stimuli materials

Speech Corpora. Materials were selected for use in a RPT experiment with native-English speaking listeners from the United States. The stimuli to be presented to listeners consisted of samples of connected speech from four weekly public addresses recorded by United States President Barack Obama, currently in the public domain and stored on a United States Government web archive (Obama, 2013, 2014a, 2014b, 2014c). These recordings had a political purpose, and thus it was assumed that the speech therein would be fairly careful (i.e., likely read and rehearsed). We chose speech of this sort (and by this speaker) for the following reasons. First and foremost, we assumed that (relative to samples of other speech styles) these samples would contain connected speech with fewer disfluencies, and with fewer reduced and ambiguous instances of intonational categories. This was desirable since the goal of the study was to predict prominence perception based on ToBI-defined categories, and other things being equal, it was assumed that spontaneous speech (such as that found in the Buckeye Corpus, Pitt et al., 2007, used in some previous work; Cole, Mo, & Hasegawa-Johnson, 2010) would contain far more disfluencies, more reduction, and in general, more ambiguity. Second, speech produced by Barack Obama in particular was chosen because it was assumed that listeners would be approximately equally familiar with it; since these listeners would be drawn from a population of mostly native New Yorkers, and because variation within this population is considerable (Newman, 2015), we chose a speaker with a dialect we assumed to be equally different from that of most of our listeners, but also equally familiar to them

[3] As referenced below, closely-related questions have recently been investigated for German, but the only study on English we are aware of is Hualde et al. (2016; see also Cole et al., this volume). However, their dataset was primarily designed to explore cross-language differences, and the statistical comparisons in their brief report only directly address our first question (and only with respect to accent status).

[4] We acknowledge, however, that this is not the only way that the inventory of American English pitch accents could be reduced. For their German data, for example, Baumann and his colleagues (Baumann, 2014; Baumann & Röhr, 2015; Baumann & Winter, 2018) group GToBI pitch accents into a combination of levels and movements (e.g., low vs. rising vs. high vs. falling). Although our statistical tests are based on accent level, we also report patterns for individual pitch accents in the results section.

(see Cole et al., 2017 for recent evidence regarding cross-dialectal prosody perception).[5] The four samples selected as stimuli (henceforth Samples A, B, C, and D) came from the "*Your Weekly Address*" series, which contains commentary by President Obama on current events. Each sample was chosen for its quality (sound quality and recording environment varies widely across the many *Your Weekly Addresses* that President Obama recorded during his two terms) and similar length (approximately 3–4 min). An example excerpt from Sample A is shown in (4):

(4) . . .*Today our economy is growing and our businesses are consistently generating new jobs. But decades-long trends still threaten the middle class. While those at the top are doing better than ever, too many Americans are working harder than ever, but feel like they can't get ahead. That's why the budget I sent Congress earlier this year is built on the idea of opportunity for all. It will grow the middle class and shrink the deficits we've already cut in half since I took office*. . .

These samples were downloaded as MP3 files (versions in uncompressed file formats not being available), suitable for analysis of intonation and the basic phonetic correlates of prominence that we were interested in. The MP3 files were converted to WAV files for the practical purpose of making them loadable in Praat (Boersma & Weenink, 2017) for later phonological annotation by trained MAE_ToBI labelers, and for later presentation to linguistically-untrained listeners in the RPT experiment. The only editing carried out on the files was the removal of brief salutations (e.g., "*Hi everybody.*" at the beginning of all four speech samples, and similar messages at the end, such as "*Thanks everybody, and have a good weekend.*"). After this editing, the result was four speech samples, containing a total of 1,821 words, or approximately 10 min of speech material (*Sample A:* 470 words/2.6 min; *Sample B:* 448 words/2.35 min; *Sample C:* 445 words/2.7 min; *Sample D:* 458 words/2.4 min). Finally, orthographic transcripts of the resulting speech materials were produced, with all punctuation and capitalization removed, as in previous RPT work (e.g., Cole, Mo, & Hasegawa-Johnson, 2010; Cole, Mo, & Baek, 2010). These transcripts were set aside, to be used by listeners in the RPT experiments to make their real-time identifications of prosodic events.

Phonological annotation. All four speech samples were phonologically annotated by two labelers with extensive training in the MAE_ToBI conventions. Recall from above that we chose these careful/performance-style speech materials with the goal of having stimuli on the lower end of reduction/phonetic ambiguity. The reason for this was because we wished to investigate the extent to which RPT judgments of prominence are related to phonological categories, and so we utilized realizations of those categories that were clearer/more canonical. Our annotation procedure and our use of the MAE_ToBI annotations were also intended to minimize the number of ambiguous instances of phonological categories that would ultimately be analyzed. First, the two ToBI labelers worked independently, not communicating with each other to resolve disagreements about the labels they considered assigning. Second, disagreements that occurred in each labeler's final annotations were left unresolved; rather than using a third "tie-breaking" annotator or some other method to force a decision, we took the two annotators' inability to agree on a label as evidence that the word's realization was sufficiently ambiguous or otherwise unclear, and simply excluded such words from the analysis. Although agreement rates between the ToBI annotators were not critical to our questions, we report them since they provide additional data on the MAE_ToBI system's inter-rater reliability (see also Pitrelli, Beckman, & Hirschberg, 1994; Syrdal & McGory, 2000; Yoon, Chavarria, Cole, & Hasegawa-Johnson, 2004; Breen, Dilley, Kraemer, & Gibson, 2012). Agreement rates are shown in Table 1, for both the presence versus absence of a pitch accent (disregarding pitch accent type) and pitch accent type (where the possible categories were: unaccented, L*, L*+H, L*+!H, !H*, H+!H*, H*, L+H*, and L+!H*). Rates are displayed both in terms of raw percent agreement and chance-corrected Cohen's kappa values, but here and throughout the paper, we rely primarily on kappa (κ) values for interpretation.[6] It is difficult to make cross-study comparisons with precision, since studies vary considerably with respect to the kinds of speech materials annotated, the level of training of annotators, how categories are defined, and whether chance-corrected measures of agreement are reported. Generally, however, the agreement levels for pitch accent presence and type in the present study were consistent with the range that has been reported previously, and on the higher end of that range (somewhat higher than found by Breen et al., 2012, and more similar to Pitrelli and colleagues' (1994) findings). This was a desirable outcome, since, as described above, only agreed-upon labels could be used for our analyses. Further, although it is important to stress that such cutoffs are rather arbitrary, in practice many researchers follow Landis and Koch (1977), who recommended interpreting kappa values of 0.01–0.20 as "slight agreement", 0.21–0.40 as "fair agreement", 0.41–0.60 as "moderate agreement", 0.61–0.80 as "substantial agreement", and 0.81–1.0 as "near perfect agreement". By these standards, our ToBI annotators agreed at rates squarely in the "substantial" or higher categories.[7] Having obtained MAE_ToBI annotations for the speech materials, these annotations—again, limited to those that both annotators agreed upon—served as our definition of the materials' phonological structure, which would later be used to model linguistically-untrained listeners' perception of prominence in the RPT experiment.

Signal-based and signal-extrinsic factors. The acoustic measures collected from the speech samples were those common to many previous studies using the RPT method (e.g., Mo, 2008; Cole et al., 2017), and were extracted automatically

[5] As pointed out by a reviewer, we did not actually attempt to collect any measure from individual participants regarding their familiarity with Obama's voice, and so it is conceivable that some differences among them might exist. While we doubt that the magnitude of any such differences would likely explain the patterns we were interested in exploring, we acknowledge some possibility and discuss the issue further in Section 3.4.

[6] We suppress the results of significance tests of Cohen's kappa (and, later in the paper, those for Fleiss's kappa). All kappa statistics we present in this paper had a z-score that was significant at the 0.001 level, indicating that agreement among listeners was always above chance level, even when the kappa was relatively low.

[7] We do not have an explanation for why agreement between the annotators was higher for Sample C than for the other samples (especially on pitch accent type). However, we note that this was also true of agreement among RPT listeners, as will be shown later in Section 2.2.2. Indeed, the relative rates of agreement across the four speech samples generally turned out to be quite similar for the ToBI annotators and RPT listeners, which suggests relevant differences internal to the speech samples rather than some aspect of the methodology or task.

Table 1
Interrater agreement for assignment of MAE_ToBI labels to the four speech samples by two trained annotators. Agreement is expressed as Cohen's kappa (with the corresponding percent agreement in parentheses).

	Presence of Pitch Accent	Pitch AccentType
Sample A	0.78 (89%)	0.62 (72%)
Sample B	0.81 (91%)	0.65 (76%)
Sample C	0.90 (95%)	0.83 (88%)
Sample D	0.84 (91%)	0.71 (81%)
All Materials	**0.83 (92%)**	**0.70 (79%)**

in Praat. First, the four speech samples were forced-aligned to word and phone tiers using the Montreal Forced Aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), with hand corrections made where necessary (accuracy was maximized by first segmenting the four speech samples into yet smaller files prior to alignment). Lexically stressed syllables were automatically identified with reference to the Carnegie Mellon Pronouncing Dictionary (ver. cmudict-0.7b) and acoustic measures were extracted for the vowel of the lexically-stressed syllable for each word by a Praat script and z-normalized. Measures included (a) *Max F0* (the maximum F0 during the vowel, measured using autocorrelation and hand corrected where tracking errors clearly occurred); (b) *RMS intensity* (measured uniformly across the frequency spectrum); and (c) the acoustic *duration* of the vowel. Other non-phonological properties of the stimuli included (a) each word's *CELEX frequency* (Baayen, Piepenbrock, & Gulikers, 1996); (b) the *number of previous repetitions* in the speech sample; and (c) *phrasal position* (whether the word was final versus non-final in an intermediate phrase). Finally, listener-based properties included (a) *gender* (the listener's self-declared status as male or female) and (b) *pragmatic skill* (the listener's score on the communication subscale of the Autism Spectrum Quotient (AQ), using the Likert scoring method).

2.1.2. Participants

Participants for the study were 160 monolingual American English speakers recruited from the Greater New York City area (51 male, 109 female; ages ranged from 18 to 48). "Monolingual" was defined as not having learned a language other than English before the age of ten, and not being (by their own estimation) a fluent speaker in any second language studied after that age. Despite this screening, two participants were later discovered to not meet the requirements, and so their data were excluded from analysis. All participants confirmed that they were free of any history of hearing or communication disorders, and that they lacked any training in prosodic theory or transcription.

2.1.3. Procedure

RPT task: Participants served as listeners in a RPT experiment, designed to elicit coarse prosodic "annotations" of both prominence and juncture (in separate tasks), although we set aside discussion of the latter task. The experiment was carried out in a laboratory setting, in a sound attenuated booth with paper and pencil. In this way our experiment was more similar to the version of the task described by Cole, Mo, & Hasegawa-Johnson (2010) than the more recent experiment reported by Cole et al. (2017), who administered the task electronically and (for some subject groups) remotely via Amazon's Mechanical Turk. Participants in our study each listened to two of the speech samples (half the participants being assigned to Samples A and B, the other half assigned to Samples C and D), identifying prominent words in one of the samples, and instances of juncture in the other (with the ordering of these two tasks, and the speech samples used for them, balanced across participants). The instructions given to participants for the prominence transcription task were intended to direct their attention to the speaker's (i.e., Obama's) voice, rather than to the meaning of utterances, although we assume that interpretation inevitably influences behavior in this task (see Cole et al., 2019, this Special Issue, for evidence supporting this assumption). The word "prominence" was not itself used with participants, and instead the task was described as in (4):

> *(4) "This part of the study is about how people use their voice when pronouncing words in English. When people speak, they use things like "loudness" and "tone of voice" to make some words "stand out" more than others. In this part of the experiment, your job is to listen to President Obama's voice and underline any and all words that he makes stand out in this way. To do this, you will need to listen very carefully to how he pronounces words in "real-time"."*

Participants were to make their identifications of prominent words by underlining them on the printed transcript provided. Although these identifications had to be made in real-time, without the ability to pause or rewind, participants were presented with the speech sample more than once, and were able to make additions and retractions of prominence identifications on each presentation, as done by Cole, Mo, & Hasegawa-Johnson (2010) in their experiment. In our study, listeners had three such chances rather than only two, and we use the term "pass" to refer to each of these three passes through the materials. Additionally, we kept track (via different color markings) which pass responses had been made on. Before beginning this task, participants carried out a brief practice trial intended to familiarize them with the setup. This practice session consisted of one utterance produced by President Obama (one that did not appear in any of the samples used in the experiment).

Individual differences measures: In addition to the RPT task, all participants in the study completed three measures of cognitive processing styles that are arguably related to pragmatic skill: the Autism Spectrum Quotient (Baron-Cohen, Wheelwright, Hill, et al., 2001), the Broad Autism Phenotype Questionnaire (Hurley, Losh, Parlier, Reznick, & Piven, 2007) and the Reading the Mind in the Eyes test (Baron-Cohen, Wheelwright, Skinner, et al., 2001). Due to space considerations, and because we generally found these measures to predict similar things, we focus only on the AQ here. The AQ is a 50-item, self-report questionnaire, measuring "autistic-like" personality traits along five dimensions: *social skills*, *imagination*, *attention to detail*, *attention-switching*, and *communication*. As discussed above, the communication subscale of the AQ (henceforth AQ-Comm) is the measure associated with pragmatic skill in previous work (e.g., Nieuwland et al., 2010; Xiang et al., 2013), and so it is this sub-scale rather than the whole AQ that is utilized in analyses here (although participants completed the whole test). The items pertaining to AQ-Comm are listed in Appendix I. Scoring was done using a 4-

point Likert scale as in Yu (2010) and elsewhere rather than the binary *agree*/*disagree* scoring used in Baron-Cohen, Wheelwright, Hill, et al. (2001) (see Stevenson & Hart, 2017 for some justification for use of the Likert scoring method). The entire experimental session took approximately 45 min to complete.

2.2. Results

2.2.1. Overview

In this section we present mixed-effects logistic regression analyses intended to model RPT listeners' prominence identifications as a function of the phonological, phonetic, and signal-extrinsic factors described above. A special interest here was in illuminating how the effects of the latter two types of cues may be dependent on phonological contrasts related to accent status and accent type. Before going on to the regression analyses, however, we offer a brief analysis of interrater agreement, which provides some useful information about how our RPT listeners compare to those in previous RPT studies, both in terms of agreement with each other as well as with the "consensus" ToBI annotation derived from our two trained transcribers.

2.2.2. Agreement

Considering first agreement among RPT listeners, we calculated Fleiss's kappa (similar to Cohen's kappa, but suitable when the number of raters is more than two) for responses by all participants for each of the four speech samples. Values are shown in Table 2, for responses that RPT listeners gave on the first pass (i.e., after hearing the materials just once) and after all three passes. The first observation we make is that agreement is quite low after just one pass. The second observation is that after three passes, listeners in our study agreed at a rate very similar to those in Cole et al. (2017) study, who agreed at a rate of $\kappa = 0.31$. Thus, RPT listeners—at least if they have multiple opportunities to hear the materials—seem to agree with each other at rates within the "fair agreement" range (by the standards of Landis & Koch, 1977). While this is much lower than the agreement that is achieved by ToBI annotators (in this study and elsewhere), this is not surprising; RPT listeners lack training and instruction and make their decisions quickly in real-time—and without any visual representation of the speech materials.

Turning now to agreement between RPT listeners and the consensus ToBI annotation, we treated prominence identifications by RPT listeners as equivalent to pitch accent identifications (ignoring pitch accent type distinctions) by ToBI annotators. Pairwise Cohen's kappas and raw proportion agreement were then calculated between the consensus ToBI annotation and each of the individual RPT listeners. The results are illustrated in Fig. 1, for RPT responses made after three passes through the materials. In order to illustrate how chance-corrected kappa relates to percent agreement, the figure plots these two measures of agreement against each other. This helps make apparent the extent to which chance inflates the picture of agreement in this binary-choice task; a listener with 75% agreement with ToBI labelers in this dataset has a chance-corrected agreement of only approximately $\kappa = 0.50$ ("moderate agreement"). Another thing it makes apparent is that almost all RPT listeners agree below this level; the individual listener who achieved the highest level of agreement with trained ToBI labelers (a linguistically naïve "super annotator" in the words of Cole et al., 2017) agreed at the $\kappa = 0.56$ level. Though considerably lower than the rate at which ToBI labelers agree with each other, this is approaching "substantial agreement" by common standards (Landis & Koch, 1977). In any case, that some listeners perform this similarly to ToBI annotators is rather impressive given their lack of training and the differences between the tasks. However given that only one or two out of the 158 listeners that we analyzed achieved this, the extent to which such "super annotator" performance is due to chance is rather unclear. We now turn to our statistical modeling of prominence perception in the RPT task, in which we sought to determine the interactive role of phonological, phonetic, and signal-external factors.

Table 2
Interrater agreement, expressed as Fleiss's kappa, among RPT listeners for each of the four speech samples. Agreement was calculated after one and after three passes through the materials.

Materials	One Pass	Three Passes
Sample A	0.185	0.266
Sample B	0.213	0.293
Sample C	0.239	0.320
Sample D	0.180	0.274
Mean	**0.204**	**0.288**

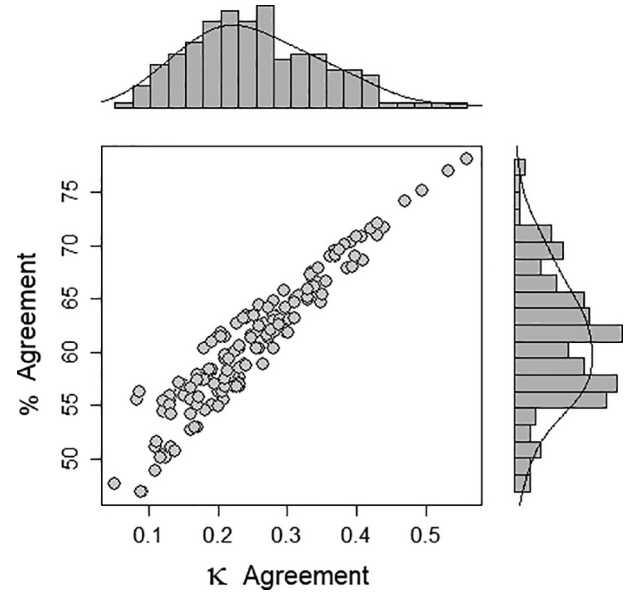


Fig. 1. Levels of agreement between expert ToBI annotators and each individual RPT listener, expressed as both Cohen's kappa (κ) and percent agreement (%). Kappa agreement and percent agreement are plotted against one another in order to highlight the relationship between chance-corrected and non-corrected measures.

2.2.3. Modeling prominence perception

2.2.3.1. Overview. We now turn to our main analyses involving how prominence perception in the RPT task relates to phonological distinctions. Using a series of mixed-effects logistic regression models, our approach in this section involved first determining the extent to which intonational phonological contrasts serve as predictors of perceived prominence, and then

to examine whether the influence of acoustic and signal-extrinsic factors vary *within* these phonologically-defined contrasts. We divide our analysis in this section into two parts, based on the phonological contrast considered.

2.2.3.2. Pitch accent status. Effect of phonology: Considering first the effect of accent status on prominence perception, we derived from the ToBI transcribers' annotations each word's status as unaccented, prenuclear accented or nuclear accented, and calculated the proportion of each judged to be prominent by listeners. This was done for judgments made after just one pass through the materials, and after all three passes that listeners were to make. Table 3 summarizes these grand proportions. First, it is clear that, overall, listeners identified more words as prominent when given additional passes (approximately 1.8 times as many overall). Second, words parsed into stronger metrical positions were more likely to be perceived as prominent. Notably, however, the most metrically prominent words—those bearing nuclear accents—were still far from ceiling-level identification, being judged as prominent less than half the time. Furthermore, unaccented words were judged as prominent at a non-zero rate by RPT listeners. Both of these findings represent clear mismatches between phonology and subjective prominence judgments. It is also worth pointing out that, although listeners identified additional prenuclear and nuclear accents on later passes through the materials, they also identified additional unaccented words as prominent. This indicates that additional passes through the materials to some extent results in listeners simply identifying more words as prominent overall. That is, in relation to ToBI annotators, RPT listeners make more "correct" identifications when given more time, but they also produce more "errors", and thus *overall agreement* between RPT listeners and ToBI annotators does not necessarily increase. This, of course, has methodological implications for RPT's use as a tool for crowdsourcing prosodic annotation (Cole et al., 2017).

To confirm the significance of these numerical patterns, mixed-effects logistic regression was carried out using the *glmer* function in the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) for *R* (R Core Team, 2018). The regression model was used to predict the binary outcome variable "marked prominent by listener" as a function of the fixed-effects factors "accent status", "pass", and "order". Accent status was treated as a discrete ordinal variable so that effects at each level (unaccented < prenuclear accented < nuclear accented) could be simultaneously compared with Tukey corrections applied (using the *ghlt* function in the *multcomp* package for *R*; see Bretz, Hothorn, & Westfall, 2011). Pass was a binary predictor (three passes through the materials vs. just one) and order (a word's linear position in the speech sample, a measure of listeners' progression through the experiment) was a continuous predictor, centered on its mean. Random-effects factors included intercepts for participant and lexical item; the model also included a by-listener random slope for accent status, as its inclusion significantly improved model fit as determined by a log likelihood ratio test (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017).

Table 3
Proportion of unaccented, prenuclear accented and nuclear accented words identified as prominent by RPT listeners on the first and last of three passes through the speech materials.

	One Pass	Three Passes
Unaccented	0.057	0.118
PPA	0.163	0.298
NPA	0.250	0.424

The output of the model is shown in Table 4, and indicates that the likelihood that RPT listeners identified a word as prominent increased along with its metrical prominence; nuclear accented words were most likely to be judged as prominent, followed by prenuclear accented words, followed by words without a pitch accent. The likelihood that a word was judged as prominent also increased significantly from the first to the last pass through the materials. There was a numerical tendency for the identification of words of lower metrical prominence to benefit more from additional passes through the materials than words of higher metrical prominence. Computable from Table 3 is that by the third pass, listeners identified 1.70 times as many nuclear accented words as after the first pass, but 1.83 times and 2.07 times as many prenuclear accented and unaccented words, respectively. However, a log likelihood ratio test indicated that an interaction term between accent status and pass improved model fit only marginally ($\chi^2 = 5.45$, $p < .1$). Thus, although there was a tendency for listeners to identify nuclear accented words sooner/-more readily in the experiment, the effect of additional passes was primarily a simple effect. In the rest of our analyses, we consider prominence judgments that were made after all three passes.

Effects of phonetic and signal-external factors: One thing that the previous section demonstrated was that listeners in the RPT task were not simply identifying nuclear accented words. Indeed, our listeners both (a) failed to identify many nuclear accented words as prominent, and (b) succeeded in identifying many prenuclear accented words—and even some unaccented words—as prominent. The purpose of this section is to determine what other factors, signal-based and signal-extrinsic, predict prominence perception. As discussed in Section 1.2, our assumption was that phonetic cues are unlikely to be weighted equally across phonological distinctions, and so, in addition to signal-extrinsic cues, our focus here was on assessing the relative contribution of F0, duration and intensity for words of different accent status. We were especially curious about whether some of these factors led listeners to identify unaccented words as prominent at the rate that they did (almost 12% of the time in the aggregate).

To this end, three logistic regression models were constructed, each intended to predict prominence perception for a different accent status category. Fixed-effects factors included in the models fell into three categories, as outlined above in Section 2.1.1. First, these included *acoustic properties*: "F0 Max" (the maximum F0 occurring during the word's stressed syllable), "duration" (the duration of the word's stressed syllable), "RMS intensity" (the root mean square intensity over the word's stressed vowel), as well as "RMS intensity*duration" (the interaction between these two factors, given their complex and dependent relation to the percept of "loudness"; Beckman, 1986, Turk & Sawusch, 1996). Second, these included *signal-extrinsic properties of the stimuli*: "lexical frequency" (CELEX frequency; Baayen et al., 1996), "repeti-

Table 4
Results for fixed-effects factors in the logistic regression model that tested for effects of *accent status* (unaccented, prenuclear accented, or nuclear accented) and *Pass* (first pass through the materials or after the third and final pass). R code for the model is shown in Appendix II.

	B	*SE*	*z*	*p*
(Intercept)	−2.8723	0.0877	−32.75	**<0.001**
Pass (3 vs. 1)	0.9589	0.0177	54.27	**<0.001**
Order	−0.0132	0.0009	−15.27	**<0.001**
Accent Status (PPA vs. Unaccented)	0.7187	0.0584	12.31	**<0.001**
Accent Status (PPA vs. NPA)	−0.6674	0.0539	−12.37	**<0.001**
Accent Status (NPA vs. Unaccented)	1.3861	0.0635	21.83	**<0.001**

tion" (number of times a word had previously occurred in the speech materials), "order" (the location of the word in the passage), and "phrasal position" (a word's positioning as final versus non-final in an intermediate phrase).[8] Finally, we also included in the model two *listener-based properties*: "gender" (the gender, male/female, declared by the listener) and "pragmatic skill" (the participant's score on AQ-Comm; higher values correspond to more autistic-like, and thus poorer, pragmatic skill). As noted earlier, values for all continuous predictors in the model were z-transformed (and thus also centered on their mean). Random-effects factors included intercepts for listener and lexical item, and a by-listener slope for lexical frequency.

The output of these models is shown in Table 5. For ease of exposition we discuss this output by making reference to Fig. 2, which displays the change in the odds ratio (which we express as percent change), a convenient measure of effect size Hosmer, Lemeshow, & Sturdivant, (2013) for all significant factors in the models. Considering first acoustic factors, it is apparent in Fig. 2 that words realized with greater acoustic prominence were generally more likely to be perceived as prominent. However, and confirming our basic prediction, the details depended on phonological status (i.e., distinctions such as accent status and accent type). For example, one standard deviation increases in F0, duration, and intensity all had their largest effects on the perceived prominence of prenuclear accented words, and (except in the case of duration) their weakest effects on the perceived prominence of unaccented words (with no significant effect of intensity at all). One interpretation of the first observation is that nuclear accented words derive their perceived prominence primarily from their structural prominence and semantic significance (Calhoun, 2006); additional acoustic prominence therefore has only a moderate effect on how nuclear accented words are perceived. In the case of unaccented words, which are generally of lower phonetic prominence to begin with, there may simply be too little acoustic variation to produce a large difference, and as discussed earlier, their F0 values reflect interpolation rather than a structurally significant target. Intensity's effect on the perceived prominence of nuclear accented words was only positive at higher durations, as the simple effect was, curiously, actually negative. However as described above, our assumption was that the interaction between intensity and duration, as a somewhat better approximation of loudness, is likely the more reliable measure (especially when the model contains both an interaction and simple effect). At any rate, it seems clear that all three parameters were quite relevant to perceived prominence—including F0, in contrast to some previous claims (e.g., Kochanski et al., 2005; Cole, Mo, & Hasegawa-Johnson, 2010). Indeed, F0 had the largest effect of the three acoustic measures we tested, with a one standard deviation increase in F0 producing a 64% increase in the odds ratio of a "prominent" response for prenuclear accented words. Notably, it is unlikely this effect for F0 would have been apparent if accent status had been invisible to the model.

Next, we consider signal-extrinsic effects on perceived prominence. Apparent in Fig. 2 is that these factors were generally better predictors of perceived prominence for unaccented words than for accented words. As described above, some previous studies of prominence perception have shown that phrase-final words are more readily identified as prominent than non-phrase-final words; while this effect was also apparent in our data, it was clearly a stronger predictor of perceived prominence for unaccented words, with an increase in the odds ratio that was more than three times the size of that for nuclear accented words. Similarly, previous studies have shown lexical frequency to be inversely associated with perceived prominence; here this relationship was found to be significant only in the case of unaccented words, for which the odds ratio for perceived prominence decreased sharply (approximately 40%) given a one standard deviation increase in lexical frequency. Repetition in the speech materials, while it affected nuclear accented words more than unaccented words, had an extremely small effect on both, with additional occurrences of a nuclear accented word corresponding to only an approximately 5% decrease in the odds ratio for perceived prominence. Similarly, a word's order in the speech materials had a statistically significant association with prominence judgments by listeners, indicating that listeners were somewhat more conservative with their prominence judgments as the experiment progressed. But here, too, the effect size was so small (less than a 2% change in the odds ratio, and only for unaccented and nuclear accented words), that we do not discuss it further.

Finally, perceived prominence was significantly predicted by properties of listeners themselves, which not only indicates the presence of individual differences, but that they are systematically related to the particular variables we tested. First, and most important to us, high scores on AQ-Comm (which indicate poorer pragmatic skill) were inversely related to the likelihood of perceived prominence. However, this was only for words of lower metrical prominence; a one standard deviation increase in AQ-Comm was associated with an odds ratio

[8] Brief commentary is required regarding some of these predictors. First, we acknowledge that F0 maximum is only an estimate of the effects that F0 has (excursion being another aspect of F0), but we have limited our measure for methodological and analytical simplicity (and for comparability with other RPT studies; e.g., Cole et al., 2010). Second, we utilized CELEX frequencies for the analyses we report below, but also explored SUBTLEX frequencies (Brysbaert & New, 2009) and did not find differences in the pattern of statistical results. Third, we point out that "phrase position" does not apply in the case of prenuclear accented words, as such words cannot, by definition, be phrase-final.

Table 5
Results for fixed-effects factors in the logistic regression models testing acoustic and signal-extrinsic factors for each accent status category. R code for the models is shown in Appendix II.

Unaccented Words:	*B*	*SE*	*z*	*p*
(Intercept)	−2.3881	0.1343	−17.78	**<0.001**
Phrase Position (Final vs. Non-final)	1.0176	0.1490	6.83	**<0.001**
CELEX Frequency	−0.5255	0.1583	−3.32	**<0.001**
Repetition	0.0198	0.0107	1.85	<0.1
Order	−0.0070	0.0029	−2.44	**<0.05**
Gender (M)	−0.2362	0.1686	−1.40	>0.1
AQ-Comm	−0.2171	0.0779	−2.79	**<0.01**
F0 Max	0.2254	0.0455	4.96	**<0.001**
Duration	0.2122	0.0530	4.00	**<0.001**
RMS Intensity * Duration	0.0766	0.0483	1.58	>0.1
RMS Intensity	0.0357	0.0452	0.79	>0.1
Prenuclear Accented Words:				
	B	*SE*	*z*	*p*
(Intercept)	−1.3742	0.1647	−8.34	**<0.001**
CELEX Frequency	−0.2714	0.2382	−1.14	>0.1
Repetition	0.0490	0.0352	1.39	>0.1
Order	−0.0025	0.0045	−0.56	>0.1
Gender (M)	−0.4075	0.1456	−2.80	**<0.001**
AQ-Comm	−0.1249	0.0668	−1.87	<0.1
F0 Max	0.4954	0.0670	7.39	**<0.001**
Duration	0.3240	0.0922	3.51	**<0.001**
RMS Intensity * Duration	0.2146	0.0784	2.74	**<0.001**
RMS Intensity	0.2452	0.1010	2.43	**<0.05**
Nuclear Accented Words:				
	B	*SE*	*z*	*p*
(Intercept)	−0.6318	0.2045	−3.09	**<0.001**
Phrase Position (Final vs. Non-final)	0.4409	0.1950	2.26	**<0.05**
CELEX Frequency	0.2683	0.1827	1.47	>0.1
Repetition	−0.0594	0.0264	−2.25	**<0.05**
Order	−0.0156	0.0044	−3.51	**<0.001**
Gender (M)	−0.3364	0.1367	−2.46	**<0.05**
AQ-Comm	−0.0443	0.0630	−0.70	>0.1
F0 Max	0.3119	0.0483	6.46	**<0.001**
Duration	0.1790	0.0532	3.37	**<0.001**
RMS Intensity * Duration	0.1925	0.0509	3.79	**<0.001**
RMS Intensity	−0.2409	0.0981	−2.46	**<0.05**

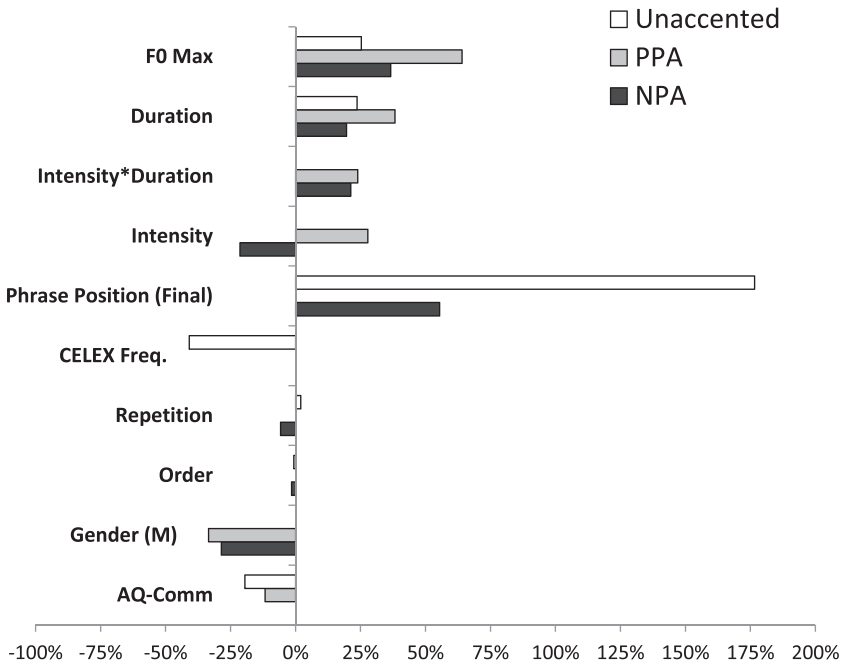


Fig. 2. Effect size (expressed as percent change in the odds ratio for a "prominent" response) for fixed-effects factors in the logistic regression models of unaccented, prenuclear accented (PPA) and nuclear accented (NPA) words. Only factors whose effect was significant are shown; note that phrase position does not apply to PPA words, as all PPA words are by definition either phrase-initial or phrase-medial.

Table 6
Proportion of words identified as prominent as a function of pitch accent type.

	L*	L*+H	L*+!H	!H*	H+!H*	L+!H*	H*	L+H*
# of Observations	1303	312	0	4002	1275	539	9659	4455
Proportion Prominent	0.329	0.170	–	0.304	0.345	0.243	0.367	0.467

Table 7
Proportion of words identified as prominent as a function of pitch accent level and status.

	L*	!H*	H*	L+H*
PPA	0.247	0.224	0.285	0.428
NPA	0.358	0.337	0.484	0.504

decrease of approximately 4% for prenuclear accented words, and 20% for unaccented words, but had no significant effect on nuclear accented words. The effect of pragmatic skill was thus like the other signal-extrinsic factors just explored, in that its effects were largely limited to words parsed into phonologically weak positions. To put this in context, the effect of a one standard deviation change in pragmatic skill was, for unaccented words, comparable to a one standard deviation change in acoustic duration. In addition to pragmatic skill, gender was a significant predictor, though only for accented words; relative to female listeners, male listeners were associated with fewer prominence identifications—a decrease of approximately 34% and 29% in the odds ratio for prenuclear accented and nuclear accented words, respectively. Gender, then, had an effect on accented words that was roughly comparable in size to a one standard deviation change in acoustic duration or intensity. We do not know of this gender difference having been reported previously. We now turn to our consideration of effects related to phonological distinctions in pitch accent type rather than pitch accent status.

2.2.3.3. Pitch accent type/level. Effect of category: As described above, we examined the role of pitch accent type primarily in terms of level, that is, groupings based on the height of the accent's starred tone. Thus L* and L*+H, were grouped as one category, and !H*, H+!H* and L+!H* as another category. The one exception to this classification involved H* and L+H*, which, again, due to L+H*'s association with raised F0, were kept distinct from each other. While our analyses center on these pitch accent level distinctions, we report in Table 6 the proportion of prominence judgments for each individual pitch accent, as well as the number of observations for each.[9]

Considering first the differences in perceived prominence associated with the categories themselves, Table 7 displays the proportion of words judged as prominent by RPT listeners for each accent level, broken down by accent status. Overall, words with L* or !H* pitch accents were least likely to be perceived as prominent, words with a L+H* were most likely to be perceived as prominent, and words with H* showed an intermediate likelihood. Notably, perceived prominence of words with H* seemed to vary the most as a function of accent status; prenuclear H* appears to pattern more like prenuclear L* and !H*, but nuclear H* patterns more like nuclear L+H* in terms of perceived prominence. We therefore explored the role of accent level separately for prenuclear and nuclear accented words.[10] Mixed-effects regression models were thus constructed similarly as for accent status contrasts in the previous section, but in this case the crucial predictor were contrasts in accent level, which was also modeled as a discrete ordinal variable (L* vs. !H* vs. H* vs. L+H*) with Tukey corrections again being applied to the multiple comparisons made.

The results of the two models are shown in Table 8. In general, words marked by pitch accents of a lower accent level were significantly less likely to be perceived as prominent than words bearing a pitch accent of a higher accent level, for both prenuclear and nuclear accented words. One exception to this was the perceived prominence for words bearing a L*, which were no less likely to be perceived as prominent than words bearing !H* (and in fact, words with !H* were numerically less likely to be judged as prominent than words with a L* when considering only nuclear pitch accents; see also Cole et al., 2019, this Special Issue). Additionally, words with H* were more strongly associated with perceived prominence than words with a L* and !H* when nuclear accented, but H* did not differ significantly from !H* in prenuclear accent position. Thus, overall, the findings seem to suggest that accent level corresponds to perceived prominence in a mostly non-gradient way. In nuclear position, the distinction is primarily between high and non-high pitch accents (where downstep is regarded as non-high); in prenuclear position, the distinction seems to be between L+H* and all other levels. It is somewhat unclear why accent status should affect H* more than the other pitch accent levels. One possibility is that this in part reflects the tendency for the second H* in English H*_H*_L-L% sequences to have a phonetically higher F0 than the first. This would of course suggest an important role for within-category phonetic variation (in this case related to F0) in prominence perception. We now turn to listeners' sensitivity to such variation.

Effects of phonetic factors: Our analysis here focused on effects of signal-based acoustic factors within accent category, setting aside signal-extrinsic factors. We did not expect reexamination of these factors in the context of accent level contrasts to yield additional insights into their effects. Additionally, instead of collapsing pitch accent types into level categories as above, our modeling of within-category phonetics effects excluded the bitonal accents, and thus accent level here refers to whether an accented word bore a L*, !H*, H* or L+H* pitch accent type. We did this both because of the more complex phonetic nature of bitonal accents (especially involving F0), but also because of the relatively small number of

[9] There were no agreed upon instances of L*+!H in our speech materials, and indeed, rather few instances of the non-downstepped L*+H).

[10] We did this for ease of interpretation, given the difficulty associated with interpreting interactions between multi-level categorical predictors. To confirm that such an interaction was likely significant, however, we first compared a model of prominence perception that contained an interaction between accent level and accent status with one that contained only the simple effects of these two factors. A log likelihood ratio test confirmed that the model with the interaction contained a significantly better fit to the data ($\chi^2 = 14.62$, $p < .01$).

Table 8
Results for fixed-effects factors in the logistic regression models that tested the effect of accent level on prominence perception. Prenuclear accented and nuclear accented words were modeled separately. R code for the models is shown in Appendix II.

Prenuclear accented words:				
	B	*SE*	*z*	*p*
(Intercept)	−1.8137	0.2019	−8.98	**<0.001**
Order	−0.0084	0.0036	−2.37	**<0.05**
!H* vs. L*	0.3041	0.2330	1.31	>0.1
H* vs. L*	0.6191	0.1817	3.41	**<0.01**
L+H* vs. L*	1.2144	0.2081	5.84	**<0.001**
H* vs. !H*	0.3149	0.1714	1.84	>0.1
L+H* vs. !H*	0.9103	0.1862	4.89	**<0.001**
L+H* vs. H*	0.5954	0.1345	4.43	**<0.001**
Nuclear accented words:				
	B	*SE*	*z*	*p*
(Intercept)	−0.6138	0.1904	−3.22	**<0.001**
Order	−0.0301	0.0034	−8.79	**<0.001**
!H* vs. L*	−0.1079	0.1760	−0.61	>0.1
H* vs. L*	0.4385	0.1698	2.58	**<0.05**
L+H* vs. L*	0.8156	0.1885	4.33	**<0.001**
H* vs. !H*	0.5464	0.1069	5.11	**<0.001**
L+H* vs. !H*	0.9235	0.1449	6.38	**<0.001**
L+H* vs. H*	0.3771	0.1271	2.97	**<0.01**

Table 9
Results for fixed-effects factors in the logistic regression models testing acoustics for each accent level category. R code for the models is shown in Appendix II.

Words with L:*				
	B	*SE*	*z*	*p*
(Intercept)	−1.3791	0.3782	−3.65	**<0.001**
F0 Max	−0.5128	0.4284	−1.20	>0.1
Duration	0.2524	0.1613	1.57	>0.1
RMS Intensity	0.5174	0.2573	2.01	**<0.05**
Words with !H:*				
	B	*SE*	*z*	*p*
(Intercept)	−1.4938	0.2253	−6.63	**<0.001**
F0 Max	0.1920	0.1141	1.68	<0.1
Duration	0.3142	0.1261	2.49	**<0.05**
RMS Intensity*Duration	0.3506	0.1172	2.99	**<0.01**
RMS Intensity	0.1050	0.2375	0.44	>0.1
Words with H:*				
	B	*SE*	*z*	*p*
(Intercept)	−1.1047	0.1195	−9.25	**<0.001**
F0 Max	0.3301	0.0523	6.32	**<0.001**
Duration	0.3905	0.0661	5.91	**<0.001**
RMS Intensity*Duration	0.2341	0.0587	3.99	**<0.001**
RMS Intensity	0.2334	0.0970	2.41	**<0.05**
Words with L+H:*				
	B	*SE*	*z*	*p*
(Intercept)	−0.4566	0.2102	−2.17	**<0.01**
F0 Max	0.7087	0.1138	6.23	**<0.001**
Duration	0.5326	0.1296	4.11	**<0.001**
RMS Intensity*Duration	−0.1847	0.1251	−1.48	>0.1
RMS Intensity	−0.5439	0.1849	−2.94	**<0.01**

observations we had for some of them (e.g., L*+H and L+!H*, as noted above). We note, however, that the dynamic nature of bitonal pitch accents may influence prominence perception in ways that are not yet clear (see, for example, work in German by Baumann & Röhr, 2015, who characterize pitch accents in more dynamic terms). Finally, to prioritize interpretability and statistical power, we also collapsed nuclear and prenuclear accents in this part of the analysis. Mixed-effects models were otherwise constructed as above, in this case one for each accent level category, with fixed-effects structure that included the same acoustic predictors that we tested for accent status contrasts (i.e., F0 max, duration, intensity, and the interaction between intensity and duration). Random intercepts again included listener and lexical item; in this case, none of the models warranted the inclusion of a random slope (i.e., inclusion of random slopes for any of the variables resulted in either no improvement to model fit, or led to the non-convergence of the model; Matuschek et al., 2017).

The output of each model is shown in Table 9. Again here, we make reference to a graphical representation of effect size for each significant factor, separately for each accent level category, shown in Fig. 3. As with accent status, above, it was generally the case that increased acoustic prominence was associated with increased likelihood of perceived prominence for all accent level categories, indicating that within-category variation was relevant in addition to the effects of category.

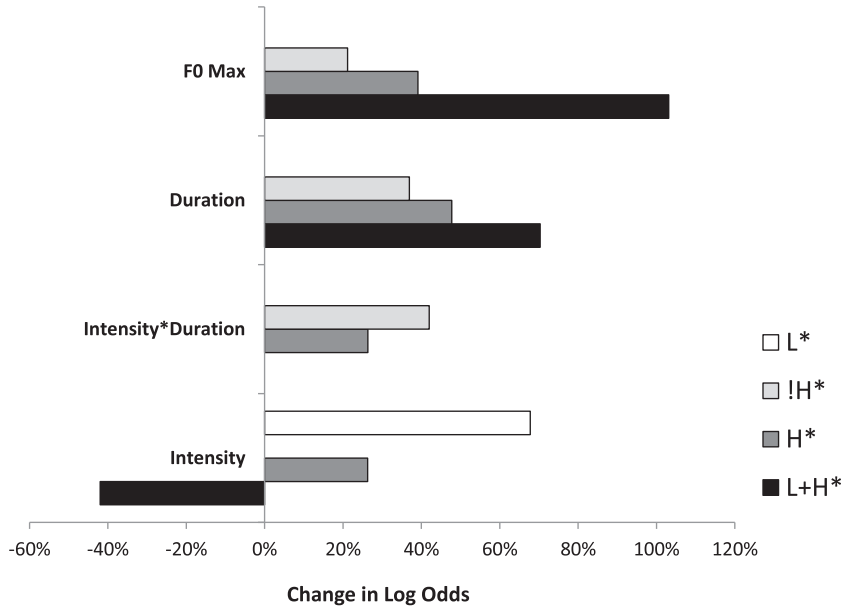


Fig. 3. Effect size (expressed as percent change in the odds ratio for a "prominent" response) for fixed-effects factors in the logistic regression models of words with L*, !H*, H*, or L+H* pitch accents. Only factors whose effect was significant are shown.

But here also, the cues and their levels of importance were not distributed evenly across categories. For example, a one standard deviation increase in F0 Max was associated with an approximately 40% increase in the odds ratio for perceived prominence for words with H*, but only half as much an increase for words with !H*—and no significant effect at all on words bearing a L*. This presumably reflects inherent properties of these categories. That is, the phonological system places limitations on the upward variation in the F0 of a !H*; if too high, it encroaches on the phonetic space of H*.

On the other hand, there are also limitations, perhaps to some extent physiological, on the amount of downward variation that can be achieved for L*; while the perceived prominence of a L* could be expected to have an inverse relationship with Max F0, variation in the low part of a speaker's range is known to be much smaller than in the upper part (e.g., Honorof & Whalen, 2005, and Bishop & Keating, 2012). Notably, however, neither of the sorts of limitations just mentioned account for the oversized effect of F0 Max on the perceived prominence of words with L+H*, which was more than two and a half times that observed for words with H*. The greater sensitivity to Max F0 for words with L+H* is consistent with previous observations about the role of F0 height (or F0 excursion; Ladd & Morton, 1997) for this pitch accent in English (Bartels & Kingston, 1994; Calhoun, 2012; Turnbull, 2017; see also Ladd & Schepman, 2003). Considering the other phonetic parameters tested, duration's effect showed a pattern similar to that just seen for F0 Max, although the differences among !H*, H* and L+H* were somewhat smaller; one standard deviation increases in duration resulted in 37%, 48%, and 70% increases in the odd ratio for perceived prominence for each of these pitch accents, respectively. Duration's effect was also similar to F0's in that it was not significantly related to perceived prominence for words bearing a L*. This was somewhat surprising to us given that, in the absence of much exploitable F0 variation for this pitch accent, speakers might be expected to manipulate (and listeners therefore expected to be sensitive to) increases in duration.[11] Instead, variation in the perceived prominence of L* was best predicted by intensity, with a one standard deviation increase in RMS intensity leading to an approximately 68% increase in the odds ratio for perceived prominence (apparently independently of duration, as an interaction between the two did not contribute to model fit). A one standard deviation increase in simple intensity also led to an increase of approximately 26% in the odds ratio for words marked by H*, and, curiously, increased intensity was inversely associated with perceived prominence for L+H*. Apparently when all other factors are held at their means, intensity factors in this way in this dataset, but we doubt the relationship is actually causal, and instead assume it reflects the relatively low importance of loudness to the perceived prominence of this accent type.[12] Finally, at longer durations, greater intensity was significantly associated with moderate increases in the odds ratio for perceiving both !H* and H* as prominent (increases of approximately 42% and 24%, respectively), as indicated by the interaction term included for the two.

In summary, then, the likelihood that a RPT listener would perceive a word with a given pitch accent as prominent depended to some extent on phonetic variation, but differently

[11] For instance, see Baumann (2014), who suggested that duration may be especially important to the realization of L* in closely-related German.

[12] Wagner and McAuliffe (2019, this Special Issue) point out another possibility: that intensity's counterintuitive association with perceived prominence here might be due, indirectly, to its relationship with phrase position. This is quite plausible; if phrase-final position is a strong top-down cue to prominence (as we found, above), and low intensity is a strong bottom-up cue to phrase boundaries (as found by Wagner & McAuliffe, this volume), then perceived prominence should sometimes systematically co-occur with lower intensity. However, this did not seem to account for the pattern here. We tested a model of L+H* with the same parameters as the one in Table 9, but this time we included an interaction between intensity and a word's status as IP-final or IP-medial; this interaction was not significant ($B=0.0480$, $SE=0.3095$, $z= 0.155$, $p > .1$). We also fit the model of L+H* in Table 9 to a subset of the data that excluded IP-final words; for these words, intensity was found to have the same negative association with perceived prominence, and still significantly so ($B=-0.5114$, $SE=0.2177$, $z=-2.35$, $p<.05$). For the present, we think the main conclusion to draw is that variation in the perceived prominence of L+H* is primarily tied to changes in fundamental frequency; intensity's contribution is comparatively small, and (possibly due to its own complex patterning throughout phrases) rather more difficult to discern.

depending on pitch accent/level category. Prominence perception for words with L* varied primarily as a function of intensity/loudness; prominence perception for H* and !H* varied within modest ranges as a function of F0, duration, and intensity; and perceived prominence for L+H* varied most strongly as a function of F0, and to a lesser extent duration. We discuss the implications of these findings, along with those presented above, in the next section.

3. Discussion

3.1. Overview

The present study explored prominence perception in English, investigating phonological, phonetic, and signal-extrinsic effects on listeners' judgments in Rapid Prosody Transcription (RPT). Our research questions addressed two basic issues that, as described in Section 1.2, represent a gap in the literature to date. First, we asked how phonological categories within the AM framework relate to patterns of perceived prominence. Surprisingly few previous studies have asked this question, and those that have either do not address English (e.g., Baumann & Röhr, 2015; Baumann & Winter, 2018) or provide only a preliminary sketch of phonology's effects (Hualde et al., 2016; Cole et al., 2017; see also Turnbull et al., 2017). Second, we asked how the relative importance of various known signal-based (acoustic phonetic) and signal-extrinsic (lexical statistics, meaning, and individual differences) cues to perceived prominence may vary depending on phonological status (accent status and accent type). We now discuss what we have learned from our findings in relation to both these theoretical questions as well as some more exploratory ones. Before concluding, we also consider some methodological implications for the use of RPT, as well as areas for future research.

3.2. Implications for understanding prominence perception

We began our discussion by arguing that, when giving prominence judgments in tasks like RPT, listeners make their decisions in the context of a phonological parse of the utterance, not directly from acoustics and other cues. Our first research question sought confirmation that phonology itself contributes to prominence perception. Evidence was found in support of this; listeners' prominence judgments were found to be significantly predicted based on metrical strength (i.e., accent status) and tonal shape (accent type/level). This is consistent with the findings of Hualde et al. (2016) and Cole et al. (this Special Issue) that nuclear accented words are significantly more perceptually prominent than unaccented words, which in turn were more prominent than unaccented words. It is also consistent with machine learning results in Bauman and Winter's (2018) recent study, which showed phonological distinctions in accent status and type to be (in that order) by far the most important predictors of prominence judgments by German-speaking listeners. Thus, while studies like ours that investigate prominence perception in the context of a phonological model of sentence prosody are few and recent, the findings so far seem clear: if the goal is to predict which words human listeners will perceive as prominent, some estimation of the listener's phonological parse is necessary.

Perhaps the most important and novel findings in our study involve the answers to our second question, which was described as having both confirmatory and exploratory aspects to it. One the one hand, we sought to confirm that cues related to acoustics and to other, signal-extrinsic factors played a phonologically-mediated role in prominence decisions. On the other hand, we also sought to explore which cues were most strongly associated with which contrasts. In fact, the findings confirmed that both kinds of cues—signal-based and signal-extrinsic—were somewhat dependent on phonology. For example, although increased acoustic prominence was overall associated with an increased likelihood of perceived prominence, its effects were stronger for words that were categorically pitch accented. This was particularly true in the case of F0, a result that we predicted based on how tunes are characterized in the AM model. That is, since F0 values reflect interpolation rather than prominence marking for words that are unaccented (Pierrehumbert, 1980), listeners are not expected to treat F0 variation during unaccented words as cueing prominence. In this regard, our findings also help us to understand why some previous studies have failed to find a relation between F0 and perceived prominence (e.g., Kochanski et al., 2005; Cole, Mo, & Hasegawa-Johnson, 2010). Since F0 is only a strong predictor of listeners' decisions about words that are accented—and because accented words tend to be a minority of words overall—an effect for F0 will be hard to detect if distinctions in accent status are invisible to the statistical model. A similar state of affairs holds for distinctions in accent type, since variation in F0 was very important to predicting perceived prominence for some accents (H* and especially L+H*) and not at all for others (L*). Notably, the finding that F0 had such an outsized effect on the perception of words with L+H* could be interpreted as providing support for Ladd and Schepman (2003) argument that the English L+H* is distinguished from H* by prominence rather than a leading low target. What is most clear, however, is that an effect of F0 on perceived prominence in English is evident—and in fairly intuitive ways—when phonological distinctions in accent type and accent status are taken into account.

The results of our experiment reveal patterns that have similarly intuitive interpretations for signal-extrinsic/top-down cues not directly related to phonetics or phonology. As we noted earlier, numerous studies have shown that listeners' prominence judgments are sensitive to factors such as a words' lexical frequency, number of repeated mentions, and phrase-final positioning. We found these effects in our data as well, but they were generally most relevant for the perception of unaccented words, an asymmetry we do not know to have been demonstrated previously. It is perhaps unsurprising, however, that the perception of metrically weaker words is more susceptible to listeners' knowledge and expectations than metrically strong words; listeners may rely more on such factors to make prominence judgments when a word lacks strong phonological and phonetic cues, and is likely lacking in semantic significance as well. A similar pattern was found in relation to individual differences in pragmatic skill, since listeners with poorer pragmatic skill were associated with reduced prominence perception, but only unaccented and prenuclear accented words were significantly affected. While individual differences have been observed to exist among RPT listeners previously

(Baumann & Winter, 2018; Cole et al., 2010; Roy et al., 2017) we do not know of a study that has attempted to attribute any of this variation to a particular source. While the underlying mechanism is still not understood, pragmatic skill had the same basic relationship to prominence sensitivity that has been reported in experiments that are likely more sensitive to differences in how "tuned in" listeners are to prosody-meaning mappings (Bishop, 2012b, 2017, Jun & Bishop, 2015; see also Bishop, 2016). One prediction we make, then, is that RPT judgments will show larger effects of pragmatic skill when the instructions that listeners are given emphasize meaning, as was done in one of the tasks reported in Cole and colleague's study (Cole et al., 2019, this Special Issue).

Interestingly, we found significant gender differences as well, and they involved accented rather than unaccented words. In particular, men were less prolific identifiers of prominence than women, for both prenuclear and nuclear accented words. We have no definitive explanation for this finding, which was in fact one of the exploratory aspects of the study. But we note that women are often claimed to possess, on average, higher levels of pragmatic skill than men (e.g., Baron-Cohen, Wheelwright, Hill, et al., 2001) and we speculate that gender in our dataset may be a proxy for variation in pragmatic skill not captured by the admittedly coarse AQ-based measure. Thus, while we are only beginning to understand the role that individual differences play, the results of our study add to a growing body of work showing that prominence perception reflects a complex integration of phonological knowledge, phonetic realization, and factors quite unrelated to any properties of a word's pronunciation (Vanio & Järvikivi, 2006; Nenkova et al., 2007; Sridhar, Nenkova, Narayanan, & Jurafsky, 2008; Cole, Mo, & Hasegawa-Johnson, 2010; Luchkina, Puri, Jyothi, & Cole, 2015; Calhoun, Kruse-Va'ai, & Wollum, 2019, Cole et al., 2017; Turnbull et al., 2017; among others).

3.3. Implications for the use of Rapid Prosody Transcription as a crowdsourcing method

The primary goal of the present study was to explore questions related to perception, but RPT has also been presented as an alternative to manual annotation by experts relying on a phonological system (Cole & Shattuck-Hufnagel, 2016; Cole et al., 2017), and our results bear on its use as such a "crowdsourcing" tool. The basic idea behind crowdsourcing approaches is that, if the crowd is big enough, and the task is sensitive enough, untrained novices should collectively be able to simulate the performance of a smaller number of expert annotators (Buhmann et al., 2002; Snow, O'Connor, Jurafsky, & Ng, 2008; Hasegawa-Johnson, Cole, Jyothi, & Varshney, 2015; see also Chang, Lee-Goldman, & Tseng, 2016). While the "annotations" that RPT listeners provide are coarse—only simple "+/− prominent" distinctions are made—it is worth reviewing what we found these coarse judgments to reflect in relation to ToBI annotators' transcriptions. First, as predicted, RPT judgments do not reflect all phonological categories equally well; while it was not the case that RPT listeners attended only to nuclear accentuation (as seemed plausible in earlier work; Cole, Mo, & Hasegawa-Johnson, 2010), they certainly identified them at a higher rate than prenuclear accents. After three passes through the materials, RPT listeners collectively identified nuclear accents about 42% of the time, while they identified prenuclear accents only about 30% of the time (Table 3). RPT judgments also asymmetrically reflected different pitch accent types, although this was not independent of pitch accent status. Pitch accent types of lower levels (like L* and !H*) were identified between a quarter of the time and a third of the time, depending on accent status, while words with L+H* were identified between 42% of the time if prenuclear, or as much as 50% of the time if nuclear (Table 7). It should also be noted that RPT listeners identify some words as prominent that ToBI labelers identify as unaccented, as much as about 12% of the time (Table 3). Thus RPT judgments favor some intonational phonological categories over others, and also reflect some mismatches, which researchers using RPT as an alternative to manual annotation by experts will need to take into account.[13] For one thing, such patterns suggest that RPT-crowdsourced annotation may be particularly sensitive to differences in speech style or particular speakers; for example, larger discrepancies between manual ToBI annotations and RPT-crowdsourced annotations are expected if a speech sample (for whatever semantic, stylistic or other reason) contains a larger proportion of L* or !H* pitch accents, since these categories are less likely to be identified as prominent by RPT listeners than H* or L+H*. Finally, unsurprisingly, the details of performance also depend on how many passes through the materials listeners are allowed to make, a methodological decision that has not been explicitly evaluated in previous work. We found RPT listeners to identify almost twice as many words as prominent if they are given three passes through the materials rather than one (see Table 3). However, since many of the additional words identified on later passes are ones that ToBI annotators identify as unaccented, these additional passes are unlikely to actually result in higher rates of agreement between RPT listeners' and ToBI annotators (even though additional passes do seem to result in an increase in agreement *among* RPT listeners; see Table 2). Taken together, however, RPT shows promise as a method for approximating the decisions of trained annotators, and future research is needed to compare the output of RPT with other phonological annotation systems, such as *RaP* (Dilley & Brown, 2005; see also Dilley & Breen, 2018), *KIM* (Kohler, 1991) and *DIMA* (Kügler et al., 2015), or cue-based methods (Brugos, Breen, Veilleux, Barnes, & Shattuck-Hufnagel, 2018).

3.4. Limitations and directions for future research

A number of other questions are left open for further research, some originating from limitations of our study, some following from its findings. One important limitation of our study is the relatively simple characterization of acoustics that we utilized. In particular, we note that the measures used in our analyses likely captured primarily local effects, and so future research should investigate how RPT judgments are dependent on patterns that occur over a wider range. For example,

[13] A reviewer points out that it is unclear what the crucial differences are between the tasks of ToBI annotators and that of RPT listeners, and we agree. It could be ToBI annotators' training in phonology and perception, the ToBI system's richer inventory of response categories, the amount of time ToBI annotators are able to take, and/or ToBI's reliance on a visual representation of the speech signal. While all of these differences in the task likely contribute to differences in their output, one way to get at the question of their relative influence might be to compare RPT with annotation systems with different properties, as we allude to further below.

future work may consider acoustic measures of accented words in relation to more global measures related to pitch range and intensity fluctuations, or acoustic properties of distal preceding material (e.g., Quené & Port, 2005; Dilley & McAuley, 2008; Niebuhr, 2009; Breen, Fedorenko, Wagner, & Gibson, 2010; Morrill, Dilley, McAuley, & Pitt, 2014; Rysling, Bishop, Clifton, & Yacovone, under review). Additionally, our focus was limited to the role that acoustics play in relation to phonology, and we therefore did not attempt to explore here how acoustic factors may interact with each other, or with individual differences variables. As we have pointed out in various places above, previous authors have often noted cross-listener variation in their studies. Baumann and Winter (2018), for example, additionally identified their listeners as clustering into groups that either responded to primarily pitch-related cues (acoustic F0 and phonological accent type) or to signal-extrinsic aspects of the stimuli (lexical and morphosyntactic cues). We leave open for further research whether factors like gender and pragmatic skill are linked to this clustering (but see recent findings with clinical populations that suggest they may be; Grice, Krüger, & Vogeley, 2016; Krüger, 2018). Finally, another individual differences-related issue that we leave for future research involves cross-dialect perception. There is evidence that listeners attend to cues that depend on their L1 in the kinds of tasks in question (Andreeva & Barry, 2012), and since there is evidence for differences in interrater agreement when speaker and listener language mismatch (Cole et al., 2017), it seems likely that listeners may attend to cues in a dialect-specific way as well.[14] Similarly, although we did not predict significant variation among our listeners in terms of experience with Obama's speech specifically, some may have been present.[15] Listeners may also differ in their experience with ethnolinguistic variation related specifically to African American English, although this variety remains understudied in the AM framework (Burdin, Holliday, & Reed, 2018; Jun & Foreman, 1996; Thomas, 2015) and it is somewhat unclear to what extent Barack Obama is a representative speaker of it (Holliday, 2016; Holliday & Villarreal, 2018; Holliday, Bishop, & Kuo, submitted). Investigating the role of sociophonetic differences among speakers and contexts—and individual differences in listeners' sensitivity to them—represents an important area for future work in the study of prominence perception.

4. Conclusion

The study presented here was intended to explore the factors that predict prominence perception by American English-speaking listeners. Using Rapid Prosody Transcription, we asked how phonology, phonetic realization, and signal-extrinsic ("top-down") factors influence prominence perception. Two especially important things were demonstrated regarding phonology's role in listeners' perception of prominence. First, phonological distinctions related to pitch accent status and pitch accent type within the AM-framework were strong predictors of subjective prominence judgments, in line with theoretical predictions based on the AM model and consistent with some recent experimental findings. Second, these phonological distinctions were also found to mediate other cues to perceived prominence, as both signal-based and listener-based effects were to some extent dependent on the phonological distinctions we tested. In conclusion, then, the present study has revealed important details about what Rapid Prosody Transcription data capture, with important implications for both its use as a tool for exploring perception, as well as a method for crowdsourcing prosodic annotation.

Acknowledgements

The authors are extremely grateful to the present volume's Guest Editors, Stefan Baumann and Francesco Cangemi, and to Oliver Niebuhr, Rory Turnbull, and an anonymous reviewer for their thoughtful comments and suggestions on many aspects of this paper. We also thank audiences at the LabPhon15 Workshop on Personality in Speech Perception & Production, the Spring 2016 General Linguistics Seminar at the University of Oxford, and the Second Conference on Prominence in Language at the University of Cologne, where parts of this project were presented. Finally, we acknowledge assistance from Juliana Colon and other members of the CUNY Prosody Laboratory, and partial funding support from Enhanced Grant-46-88 from the Professional Staff Congress of the City University of New York.

Appendix I

Items on the Communication Subscale of the Autism Spectrum Quotient (AQ-Comm) in Baron-Cohen, Wheelwright, Hill, et al. (2001). Higher levels of autistic traits related to communication (and thus lower pragmatic skill) are assigned when a participant responds with 'slightly agree' or 'strongly agree' to statements 1–6, or with 'slightly disagree' or 'strongly disagree' to the statements in 7–10.

1	Other people frequently tell me that what I've said is impolite, even though I think it is polite.
2	When I talk, it isn't always easy for others to get a word in edgeways.
3	I frequently find that I don't know how to keep a conversation going.
4	When I talk on the phone, I'm not sure when it's my turn to speak.
5	I am often the last to understand the point of a joke.
6	People often tell me that I keep going on and on about the same thing.
7	I enjoy social chit-chat.
8	I find it easy to "read between the lines" when someone is talking to me.
9	I know how to tell if someone listening to me is getting bored.
10	I am good at social chit-chat.

Appendix II

R code for mixed-effects models (linear and logistic) presented in the results section.

[14] See Smith and Rathcke (2020, this Special Issue) for recent discussion of dialectal variation in the cueing of prominence patterns.

[15] An interesting possibility pointed out to us by Oliver Niebuhr (pc) is that, as an acoustic expression of "charisma", Barack Obama's speech, particularly this politically-oriented sample, may feature a rather idiosyncratic use of phonetic cues to prominence (for relevant discussion, see Niebuhr, Skarnitzl, & Tylečková, 2018, and Niebuhr, Thumm, & Michalsky, 2018). While we are not in a position to explore this here, such observations make clear that individual differences occur among both speakers and listeners, and indeed the two may interact in complex ways (Cangemi, Krüger, & Grice, 2015).

Model corresponding to results shown in Table 4

glmer(Marked_Prominent ~ Pass + Accent_Status + Order + (1 + Accent_Status | Listener) + (1|Lexical_Item), family = binomial)

Models corresponding to results shown in Table 5

For unaccented words: glmer(Marked_Prominent ~ Celex_Frequency + Repetition + Order + Gender + AqComm + F0max + Duration + RMSintensity*Duration + (1 + Celex_Frequency|Listener) + (1|Lexical_Item), family = binomial)

For prenuclear accented words: glmer(Marked_Prominent ~ Celex_Frequency + Repetition + Order + Gender + AqComm + F0max + Duration + RMSintensity*Duration + (1 + Celex_Frequency|Listener) + (1|Lexical_Item), family = binomial)

For nuclear accented words: glmer(Marked_Prominent ~ Phrase_Position + Celex_Frequency + Repetition + Order + Gender + AqComm + F0max + Duration + RMSintensity*Duration + (1 + Celex_Frequency|Listener) + (1|Lexical_Item), family = binomial)

Models corresponding to results shown in Table 8

For prenuclear accented words: glmer(Marked_Prominent ~ Accent_Level + Order + (1 + Accent_Level|Listener) + (1|Lexical_Item), family = binomial)

For nuclear accented words: glmer(Marked_Prominent ~ Accent_Level + Order + (1 + Accent_Level|Listener) + (1|Lexical_Item), family = binomial)

Models corresponding to results shown in Table 9

For words with L:* glmer(Marked_Prominent ~ F0max + RMSintensity + Duration + (1|Listener) + (1|Lexical_Item) + family = binomial)

For words with !H:* glmer(Marked_Prominent ~ F0max + RMSintensity * Duration + (1|Listener) + (1|Lexical_Item) + family = binomial)

For words with H:* glmer(Marked_Prominent ~ F0max + RMSintensity * Duration + (1|Listener) + (1|Lexical_Item) + family = binomial)

For words L + H:* glmer(Marked_Prominent ~ F0max + RMSintensity * Duration + (1|Listener) + (1|Lexical_Item) + family = binomial)

References

Andreeva, B., & Barry, W. J. (2012). Fine phonetic detail in prosody. Cross-language differences need not inhibit communication. In O. Niebuhr (Ed.), *Prosodies: Context, function, and communication* (pp. 259–288). Berlin/New York: de Gruyter.

Arnold, D., Möbius, B., & Wagner, P. (2011). Comparing word and syllable prominence rated by naive listeners. *Proceedings of Interspeech, 2011*, 1877–1880.

Arvaniti, A. (to appear). The autosegmental-metrical model of intonational phonology. In S. Shattuck-Hufnagel & J. Barnes (Eds.), Prosodic theory and practice. Cambridge, MA: MIT Press.

Ausburn, L., & Ausburn, F. (1978). Cognitive styles: Some information and implications for instructional design. *Educational Communication and Technology, 26*(4), 337–354.

Ayers, G. (1996). *Nuclear accent types and prominence: Some psycholinguistic experiments. Ph.D. dissertation*. The Ohio: State University.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). CELEX 2. [Speech database]. Philadelphia: Linguistic Data Consortium. Retrieved from https://catalog.ldc.upenn.edu/ldc96l14.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001a). The "Reading the Mind in the Eyes" Test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry, 42*(2), 241–251. https://doi.org/10.1111/1469-7610.00715.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001b). The autism-spectrum quotient (AQ): Evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders, 31*(1), 5–17. https://doi.org/10.1023/A:1005653411471.

Bartels, C., & Kingston, J. (1994). Salient pitch cues in the perception of contrastive focus. In P. Bosch & R. van der Sandt (Eds.), *Focus and natural language processing Vol. 1: Intonation and syntax* (pp. 1–10). IBM Deutschland Informations systeme GmbH Scientific Center, Institute for Logic and Linguistics.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

Baumann, S. (2006). *The intonation of givenness: Evidence from German* Ph.D. dissertation. Saarland University.

Baumann, S. (2014). The importance of tonal cues for untrained listeners in judging prominence. In S. Fuchs, M. Grice, A. Hermes, L. Lancia, & D. Mücke (Eds.), *Proceedings of the 10th international seminar on speech production (ISSP)* (pp. 21–24).

Baumann, S., Niebuhr, O., & Schroeter, B. (2016). Acoustic cues to perceived prominence levels: Evidence from German spontaneous speech. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), Proceedings of speech prosody 2016, 711–715. https://doi.org/10.21437/SpeechProsody.2016-146.

Baumann, S., & Riester, A. (2012). Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In G. Elordieta & P. Prieto (Eds.), *Prosody and meaning* (pp. 119–161). Berlin: Mouton de Gruyter.

Baumann, S., & Röhr, C. (2015). The perceptual prominence of pitch accent types in German. In The Scottish Consortium for ICPhS 2015 (Ed.), Proceedings of the 18th international congress of the phonetic sciences (pp. 1–5). Glasgow, UK: The University of Glasgow, ISBN 978-0-85261-941-4. Paper number 0298. Retrieved from https://www.internationalphoneticassociation.org/icphsproceedings/ICPhS2015/Papers/ICPHS0298.pdf.

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics, 70*, 20–38.

Beckman, M. (1986). *Stress and non-stress accent (Netherlands Phonetic Archives 7)*. Dordrecht: Foris.

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, III*, 15–70.

Beckman, M. (1996). The parsing of prosody. *Language and Cognitive Processes, 11*(1–2), 17–68. https://doi.org/10.1080/016909696387213.

Beckmann, M., & Ayers Elam, G. (1997). *Guidelines for ToBI labeling (Version 3)*. The Ohio State University. Unpublished ms.

Beckmann, M., & Hirschberg, J. (1994). *The ToBI annotation conventions*. The Ohio State University. Unpublished ms.

Bishop, J. (2012a). Information structural expectations in the perception of prosodic prominence. In G. Elordieta & P. Prieto (Eds.), *Prosody and meaning* (pp. 239–270). Berlin: Mouton de Gruyter.

Bishop, J. (2012b). Focus, prosody, and individual differences in "autistic" traits: Evidence from cross-modal semantic priming. *UCLA Working Papers in Phonetics, 111*, 1–26.

Bishop, J. (2013). *Prenuclear accentuation in English: Phonetics, phonology, and information structure* Ph.D. dissertation. Los Angeles: University of California.

Bishop, J. (2016). Individual differences in top-down and bottom-up prominence perception. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), *Proceedings of speech prosody* (pp. 668–672). https://doi.org/10.21437/SpeechProsody.2016-137.

Bishop, J. (2017). Focus projection and prenuclear accents: Evidence from lexical processing. *Language, Cognition and Neuroscience, 32*(2), 236–253. https://doi.org/10.1080/23273798.2016.1246745.

Bishop, J., & Keating, P. (2012). Perception of pitch location within a speaker's range: Fundamental frequency, voice quality and speaker sex. *The Journal of the Acoustical Society of America, 132*(2), 1100–1112. https://doi.org/10.1121/1.4714351.

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program] (Version 6.0.35). Retrieved from http://www.praat.org.

Breen, M., Dilley, L., 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), 277–312. https://doi.org/10.1515/cllt-2012-0011.

Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes, 25*(7), 1044–1098. https://doi.org/10.1080/01690965.2010.504378.

Bretz, F., Westfall, P., & Hothorn, T. (2011). *Multiple comparisons using R*. FL: Chapman and Hall/CRC Press.

Brugos, A., Breen, M., Veilleux, N., Barnes, J., & Shattuck-Hufnagel, S. (2018). Cue-based annotation and analysis of prosodic boundary events. In *Proceedings of the 9th international conference on speech prosody* (pp. 245–249). https://doi.org/10.21437/SpeechProsody.2018-50.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977–990. https://doi.org/10.3758/BRM.41.4.977.

Buhmann, J., Caspers, J., van Heuven, V., Hoekstra, H., Martens, J.-P., & Swerts, M. (2002). Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In *Proceedings of the language resources and evaluation conference (LREC)* (pp. 779–785).

Burdin, R. S., Holliday, N., & Reed, P. (2018). Rising above the standard: Variation in L +H* contour use across 5 varieties of American English. In *Proceedings of the 9th international conference on speech prosody* (pp. 582–586). http://dx.doi.org/10.21437/SpeechProsody.2018-72.

Calhoun, S. (2006). *Information structure and the prosodic structure of English: A probabilistic relationship* Ph.D. dissertation. University of Edinburgh.

Calhoun, S. (2012). The theme/rheme distinction: Accent type or relative prominence? *Journal of Phonetics, 40*(2), 329–349. https://doi.org/10.1016/j.wocn.2011.12.001.

Calhoun, S., Wollum, E., & Kruse Va'ai, E. (2019). Prosodic prominence and focus: Expectation affects interpretation in Samoan and English. Language and Speech (Special issue on Prosodic prominence: a cross-linguistic perspective), 1–35 (online first). https://doi.org/10.1177/0023830919890362.

Cambier-Langeveld, T., & Turk, A. (1999). A cross-linguistic study of accentual lengthening: Dutch vs. English. *Journal of Phonetics, 27*(3), 255–280. https://doi.org/10.1006/jpho.1999.0096.

Cangemi, F., & Grice, M. (2016). The importance of a distributional approach to categoriality in Autosegmental-Metrical accounts of intonation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology, 7*(1), 1–20. https://doi.org/10.5334/labphon.28.

Cangemi, F., Krüger, M., & Grice, M. (2015). Listener-specific perception of speaker-specific production in intonation. In S. Fuchs, D. Pape, C. Petrone, & P. Perrier (Eds.), *Individual differences in speech production and perception* (pp. 23–145). Frankfurt: Peter Lang.

Chang, N., Lee-Goldman, R., & Tseng, M. (2016). Linguistic wisdom from the crowd. In *Proceedings of the third association for the advancement of artificial intelligence conference on human computation and crowdsourcing* (pp. WS-15-24.).

Cole, J., Hualde, J., Smith, C., Eager, C., Mahrt, T., & Napoleão de Souzac, R. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics, 75*, 113–147.

Cole, J., Hualde, J., Eager, C., & Mahrt, T. (2015). On the prominence of accent in stress reversal. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences* (pp. 1–5). Glasgow, UK: The University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0771. Retrieved from http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0771.pdf.

Cole, J., Mahrt, T., & Hualde, J. (2014). Listening for sound, listening for meaning: Task effects on prosodic transcription. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Proceedings of speech prosody 7* (pp. 859–863).

Cole, J., Mahrt, T., & Roy, J. (2017). Crowd-sourcing prosodic annotation. *Computer Speech & Language, 45*, 300–325. https://doi.org/10.1016/j.csl.2017.02.008.

Cole, J., Mo, Y., & Baek, S. (2010). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes, 25*(7), 1141–1177. https://doi.org/10.1080/01690960903525507.

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

Dilley, L., & Breen, M. (2018). An enhanced Autosegmental-Metrical theory (AM+) facilitates phonetically transparent prosodic annotation. In *Proceedings of the 6th international symposium on tonal aspects of language* (pp. 67–71). http://dx.doi.org/10.21437/TAL.2018-14.

Dilley, L., & Brown, M. (2005). *The RaP (Rhythm and Pitch) labeling system*. MIT. Unpublished ms.

Dilley, L., & McAuley, D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language, 59*, 294–311.

Epstein, M. (2002). *Voice quality and prosody in English* Ph.D. dissertation. Los Angeles: University of California.

Erickson, D., Kim, J., Kawahara, S., Wilson, I., Menezes, C., Suemitsu, A., & Moore, J. (2015). Bridging articulation and perception: The C/D model and contrastive emphasis. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences* (pp. 1–5). Glasgow, UK: The University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0527. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0527.pdf.

Eriksson, A., Thunberg, G., & Traunmüller, H. (2001). Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. In *Proceedings of the European conference on speech communication and technology* (pp. 309–402).

Grice, M., Krüger, M., & Vogeley, K. (2016). Adults with Asperger syndrome are less sensitive to intonation than control persons when listening to speech. *Culture and Brain, 4*(1), 38–50. https://doi.org/10.1007/s40167-016-0035-6.

Gussenhoven, C. (2015). Does phonological prominence exist? *Lingue e Linguaggio, 14*(1), 7–24. https://doi.org/10.1418/80751.

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

Gussenhoven, C., Repp, B., Rietveld, A., Rump, H., & Terken, J. (1997). The perceptual prominence of fundamental frequency peaks. *The Journal of the Acoustical Society of America, 102*(5), 3009–3022. https://doi.org/10.1121/1.420355.

Hasegawa-Johnson, M., Cole, J., Jyothi, P., & Varshney, L. (2015). Models of dataset size, question design, and cross-language speech perception for speech crowdsourcing applications. *Laboratory Phonology, 6*(3–4), 381–431. https://doi.org/10.1515/lp-2015-0012.

Hirschberg, J., Gravano, A., Nenkova, A., Sneed, E., & Ward, G. (2007). Intonational overload: Uses of the downstepped (H*! H* L-L%) contour in read and spontaneous speech. In J. Cole & J. Hualde (Eds.), *Laboratory phonology 9* (pp. 455–482). Berlin: Mouton de Gruyter. https://doi.org/10.7916/D8KW5QG4.

Holliday, N. (2016). *Intonational variation, linguistic style, and the black/biracial experience* Ph.D. dissertation. New York University.

Holliday, N., Bishop, J., & Kuo, G. (submitted). Prosody and political style: The case of Barack Obama and the L+H* pitch accent. Proceedings of Speech Prosody 2020.

Holliday, N., & Villarreal, D. (2018). How black does Obama sound now? Testing listener judgments of intonation in incrementally manipulated speech. *University of Pennsylvania Working Papers in Linguistics, 24*(2). Article 8. Available at: https://repository.upenn.edu/pwpl/vol24/iss2/8.

Honorof, D., & Whalen, D. H. (2005). Perception of pitch location within a speaker's F0 range. *The Journal of the Acoustical Society of America, 117*(4), 2193–2200. https://doi.org/10.1121/1.1841751.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Third edition). Hoboken, New Jersey: John Wiley & Sons.

Hualde, J., Cole, J., Smith, C. L., Eager, C. D., Mahrt, T., & de Souza, R. N. (2016). The perception of phrasal prominence in English, Spanish and French conversational speech. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), *Proceedings of speech prosody 2016* (pp. 459–463). https://doi.org/10.21437/SpeechProsody.2016-94.

Hurley, R., & Bishop, J. (2016). Interpretation of "only": Prosodic influences and individual differences. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), *Proceedings of speech prosody 2016* (pp. 193–197). https://doi.org/10.21437/SpeechProsody.2016-40.

Hurley, R. S., Losh, M., Parlier, M., Reznick, J. S., & Piven, J. (2007). The broad autism phenotype questionnaire. *Journal of Autism and Developmental Disorders, 37*(9), 1679–1690. https://doi.org/10.1007/s10803-006-0299-3.

Jagdfeld, N., & Baumann, S. (2011). Order effects on the perception of relative prominence. In *Proceedings of the 17th international congress of phonetic sciences* (pp. 958–961). Hong Kong, China: The City University of Hong Kong. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Jagdfeld/Jagdfeld.pdf.

Jun, S.-A., & Bishop, J. (2015). Priming implicit prosody: Prosodic boundaries and individual differences. *Language and Speech, 58*(4), 459–473. https://doi.org/10.1177/0023830914563368.

Jun, S.-A., & Foreman, C. (1996). Boundary tones and focus realization in African American English intonations. *The Journal of the Acoustical Society of America, 100*(4). 2826 2826.

Kimball, A., & Cole, J. (2016). Pitch contour shape matters in memory. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), *Proceedings of speech prosody 2016* (pp. 1171–1175). https://doi.org/10.21437/SpeechProsody.2016-241.

Kimball, A., & Cole, J. (2014). Avoidance of stress clash in perception of conversational American English. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Proceedings of speech prosody 7* (pp. 497–501).

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America, 118*(2), 1038–1054. https://doi.org/10.1121/1.1923349.

Kohler, K. (1991). A model of German intonation. *Arbeitsberichte Des Instituts Für Phonetik Der Universität Kiel (AIPUK), 25*, 295–360.

Krüger, M. (2018). *Prosodic decoding and encoding of referential givenness in adults with autism spectrum disorders* Ph.D. dissertation. University of Cologne.

Kulakova, E., & Nieuwland, N. (2016). Pragmatic skills predict online counterfactual comprehension: Evidence from the N400. *Cognitive, Affective, & Behavioral Neuroscience, 16*(5), 814–824. https://doi.org/10.3758/s13415-016-0433-4.

Kügler, F., Smolibocki, B., Arnold, D., Baumann, S., Braun, B., Grice, M., ... Wagner, P. (2015). DIMA: Annotation guidelines for German intonation. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences* (pp. 1–5). Glasgow, UK: The University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0317. Retrieved from: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0317.pdf.

Ladd, D. R. (1994). Constraints on the gradient variability of pitch range, or, Pitch Level 4 lives! In P. Keating (Ed.), *Papers in laboratory phonology 3: Phonological structure and phonetic form* (pp. 43–63). Cambridge, UK: Cambridge University Press.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge/New York: Cambridge University Press.

Ladd, D. R. (2008). *Intonational phonology* (Second edition). Cambridge/New York: Cambridge University Press.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: Continuous or categorical? *Journal of Phonetics, 25*(3), 313–342. https://doi.org/10.1006/jpho.1997.0046.

Ladd, D. R., & Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of Phonetics, 31*(1), 81–112.

Ladd, D. R., Verhoeven, J., & Jacobs, K. (1994). Influence of adjacent pitch accents on each other's perceived prominence: Two contradictory effects. *Journal of Phonetics, 22*(1), 87–99.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*(2), 363–374. https://doi.org/10.2307/2529786.

Luchkina, T., Puri, V., Jyothi, P., & Cole, J. (2015). Prosodic and structural correlates of perceived prominence in Russian and Hindi. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th international congress of phonetic sciences* (pp. 1–5). Glasgow, UK: The University of Glasgow. ISBN 978-0-85261-941-4. Paper number 0793. Retrieved from: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0793.pdf.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *Proceedings of Interspeech, 2017*, 498–502.

Mahrt, T. (2016). LMEDS: Language markup and experimental design software (Version 2.4). Retrieved from https://github.com/timmahrt/LMEDS.

Mahrt, T., Cole, J., Fleck, M., & Hasegawa-Johnson, M. (2012). F0 and the perception of prominence. *Proceedings of Interspeech, 2012*, 2421–2424.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001.

Mixdorff, H., Cossio-Mercado, C., Hönemann, A., Gurlekian, J., Evin, D., & Torres, H. (2015). Acoustic correlates of perceived syllable prominence in German. *Proceedings of Interspeech, 2015*, 51–55.

Mo, Y. (2008). Acoustic correlates of prosodic prominence for naïve listeners of American English. In *Proceedings of the 34th annual meeting of the Berkeley Linguistic Society* (pp. 257–267).

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

Nenkova, A., Brenier, J., Kothari, A., Calhoun, S., Whitton, L., Beaver, D., & Jurafsky, D. (2007). To memorize or to predict: Prominence labeling in conversational speech. In *Proceedings of NAACL human language technology conference* (pp. 9–16).

Newman, M. (2015). *New York City English*. Boston: De Gruyter.

Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica, 66*, 95–112.

Niebuhr, O., Skarnitzl, R., & Tylečková, L. (2018). The acoustic fingerprint of a charismatic voice – Initial evidence from correlations between long-term spectral features and listener ratings. In *Proceedings of the 9th international conference on speech prosody* (pp. 359–363). http://dx.doi.org/10.21437/SpeechProsody.2018-73.

Niebuhr, O., Thumm, J., & Michalsky, J. (2018). Shapes and timing in charismatic speech – Evidence from sounds and melodies. Proceedings of the 9th International Conference on Speech Prosody, 582–586. http://dx.doi.org/10.21437/SpeechProsody.2018-118

Niebuhr, O., & Winkler, J. (2017). The relative cueing power of F0 and duration in German prominence perception. *Proceedings Interspeech, 2017*, 611–615. https://doi.org/10.21437/Interspeech.2017-375.

Nieuwland, M., Ditman, T., & Kuperberg, G. (2010). On the incrementality of pragmatic processing: An ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language, 63*(3), 324–346. https://doi.org/10.1016/j.jml.2010.06.005.

Obama, B. (2013). Remarks of President Barack Obama: Weekly Address, 28, November, 2013. Retrieved from https://obamawhitehouse.archives.gov/briefing-room/weekly-address.

Obama, B. (2014). Remarks of President Barack Obama: Weekly Address, 4 May, 2014. Retrieved from https://obamawhitehouse.archives.gov/briefing-room/weekly-address.

Obama, B. (2014). Remarks of President Barack Obama: Weekly Address, 5 April, 2014. Retrieved from https://obamawhitehouse.archives.gov/briefing-room/weekly-address.

Obama, B. (2014). Remarks of President Barack Obama: Weekly Address, 8 March, 2014. Retrieved from https://obamawhitehouse.archives.gov/briefing-room/weekly-address.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation* Ph.D. dissertation. Massachusetts Institute of Technology.

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

Pintér, G., Mizuguchi, S., & Tateishi, K. (2014). Perception of prosodic prominence and boundaries by L1 and L2 speakers of English. *Proceedings of Interspeech, 2014*, 544–547.

Pitrelli, J. F., Beckman, M. E., & Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the 3rd international conference on spoken language processing* (pp. 123–126).

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). *Buckeye corpus of conversational speech (2nd release) [www.buckeyecorpus. osu. edu]*. Columbus, OH: Department of Psychology. Ohio State University (Distributor).

Quené, H., & Port, R. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica, 62*, 1–13.

R Core Team (2018). R: A language and environment for statistical computing, ver. 3.5.1. Foundation for Statistical Computing, Vienna, Austria. Available at: www.R-project.org.

Rietveld, A., & Gussenhoven, C. (1985). On the relation between pitch excursion size and prominence. *Journal of Phonetics, 13*, 299–308.

Röhr, C., & Baumann, S. (2011). Decoding information status by type and position of accent in German. In *Proceedings of the 17th international congress of phonetic sciences* (pp. 1706–1709). Hong Kong, China: The City University of Hong Kong. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2011/OnlineProceedings/RegularSession/Rohr/Rohr.pdf.

Rosenberg, A. (2009). *Automatic detection and classification of prosodic events* Ph.D dissertation. Columbia University.

Rosenberg, A., Hirschberg, J., & Manis, K. (2010). Perception of English prominence by native Mandarin Chinese speakers. *Proceedings of Speech Prosody, 2010*. https://doi.org/10.7916/D8BR91N2, 100982:1-4.

Roy, J., Cole, J., & Mahrt, T. (2017). Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology, 8*(1), 1–36. https://doi.org/10.5334/labphon.108.

Rysling, A., Bishop, J., Clifton, C., & Yacovone, A. (under review). Preceding syllable cues are necessary for the accent advantage effect. Ms, University of California, Santa Cruz.

Sluijter, A. M., & Van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *The Journal of the Acoustical Society of America, 100*(4), 2471–2485.

Smith, C. (2009). Naïve listeners' perceptions of French prosody compared to the predictions of theoretical models. Proceedings of the third symposium prosody/discourse interfaces, 335–349.

Smith, C., & Edmunds, P. (2013). Native English listeners' perceptions of prosody in L1 and L2 reading. *Proceedings of Interspeech, 2013*, 235–238.

Smith, R., & Rathcke, T. (2020). Dialectal phonology constrains the phonetics of prominence. *Journal of Phonetics, 78* 100934.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263).

Sridhar, V. K. R., Nenkova, A., Narayanan, S., & Jurafsky, D. (2008). Detecting prominence in conversational speech: Pitch accent, givenness and focus. In P. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the 4th conference on speech prosody* (pp. 453–456).

Stevenson, J. L., & Hart, K. R. (2017). Psychometric properties of the autism-spectrum quotient for assessing low and high levels of autistic traits in college students. *Journal of Autism and Developmental Disorders, 47*(6), 1838–1853.

Stewart, M., & Ota, M. (2008). Lexical effects on speech perception in individuals with "autistic" traits. *Cognition, 109*(1), 157–162. https://doi.org/10.1016/j.cognition.2008.07.010.

Streefkerk, B., Pols, L., & Ten Bosch, L. F. (1997). Prominence in read aloud sentences, as marked by listeners and classified automatically. In R. J. J. H. van Son (Ed.), *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, 21* (pp. 101–116).

Syrdal, A. K., & McGory, J. (2000). Inter-transcriber reliability of ToBI prosodic labeling. In *Proceedings of the 6th international conference on spoken language processing* (pp. 235–238).

Terken, J. (1991). Fundamental frequency and perceived prominence of accented syllables. *The Journal of the Acoustical Society of America, 89*(4), 1768–1776. https://doi.org/10.1121/1.401019.

Terken, J. (1994). Fundamental frequency and perceived prominence of accented syllables II: Nonfinal accents. *The Journal of the Acoustical Society of America, 95*(6), 3662–3665. https://doi.org/10.1121/1.409936.

Terken, J., & Hermes, D. (2000). The perception of prosodic prominence. In M. Horne (Ed.), *Prosody: Theory and experiment. [Text, Speech and Language Technology Vol. 14]* (pp. 89–127). Dordrecht: Springer.

Thomas, E. (2015). Prosodic features of African American English. In J. Bloomquist, L. Green, & S. Lanehart (Eds.), *The Oxford handbook of African American Language* (pp. 420–438). Oxford/New York: Oxford University Press.

Turk, A., & Sawusch, J. R. (1996). The processing of duration and intensity cues to prominence. *The Journal of the Acoustical Society of America, 99*(6), 3782–3790. https://doi.org/10.1121/1.414995.

Turnbull, R. (2017). The role of predictability in intonational variability. *Language and Speech, 60*(1), 123–153. https://doi.org/10.1177/0023830916647079.

Turnbull, R., Royer, A., Ito, K., & Speer, S. (2017). Prominence perception is dependent on phonology, semantics, and awareness of discourse. *Language, Cognition and Neuroscience, 32*(8), 1017–1033. https://doi.org/10.1080/23273798.2017.1279341.

Ujiie, Y., Asai, T., & Wakabayashi, A. (2015). The relationship between level of autistic traits and local bias in the context of the McGurk effect. *Frontiers in Psychology, 6*, 891. https://doi.org/10.3389/fpsyg.2015.00891.

Vainio, M., & Järvikivi, J. (2006). Tonal features, intensity, and word order in the perception of prominence. *Journal of Phonetics, 34*(30), 319–342. https://doi.org/10.1016/j.wocn.2005.06.004.

Wagner, P. (2005). Great expectations-introspective vs. perceptual prominence ratings and their acoustic correlates. In Proceedings of the Interspeech 2005 - Eurospeech, ninth European conference on speech communication and technology. 2381–2384.

Wagner, M., & McAuliffe, M. (2019). The effect of focus prominence on phrasing. *Journal of Phonetics, 77* 100930.

Xiang, M., Grove, J., & Giannakidou, A. (2013). Dependency-dependent interference: NPI interference, agreement attraction, and global pragmatic inferences. *Frontiers in Psychology, 4*, 708. https://doi.org/10.3389/fpsyg.2013.00708.

Yang, X., Minai, U., & Fiorentino, R. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. *Frontiers in Psychology, 9*, 1720. https://doi.org/10.3389/fpsyg.2018.01720.

Yoon, T.-J., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004). Inter-transcriber reliability of prosodic labeling on telephone conversation using ToBI. *Eighth International Conference on Spoken Language Processing*, 2729–2732.

Yu, A. (2010). Perceptual compensation is correlated with individuals' "autistic" traits: Implications for models of sound change. *PLoS ONE, 5*(8), 1–9. https://doi.org/10.1371/journal.pone.0011950.

Yu, A. (2013). Individual differences in socio-cognitive processing and the actuation of sound change. In A. C. L. Yu (Ed.), *Origins of sound change: Approaches to phonologization* (pp. 201–227). Oxford, UK: Oxford University Press.

Yu, A. (2016). Vowel-dependent variation in Cantonese /s/ from an individual-difference perspective. *The Journal of the Acoustical Society of America, 139*(4), 1672–1690. https://doi.org/10.1121/1.4944992.