Original Article

Language *and* Speech

The Role of Prosody in Disambiguating English Indirect Requests

Language *and* Speech
1–25

DOI: 10.1177/00238309221087715
journals.sagepub.com/home/las
SAGE

Sean Trott
Department of Cognitive Science, UC San Diego, USA

Stefanie Reed
Department of Linguistics, City University of New York Graduate Center, USA

Dan Kaliblotzky
Department of Cognitive Science, UC San Diego, USA

Victor Ferreira
Department of Psychology, UC San Diego, USA

Benjamin Bergen
Department of Cognitive Science, UC San Diego, USA

Abstract

Ambiguity pervades language. The sentence "My office is really hot" could be interpreted as a complaint about the temperature or as an indirect request to turn on the air conditioning. How do comprehenders determine a speaker's intended interpretation? One possibility is that speakers and comprehenders exploit prosody to overcome the pragmatic ambiguity inherent in indirect requests. In a pre-registered behavioral experiment, we find that human listeners can successfully determine whether a given utterance was intended as a request at a rate above chance (55%), above and beyond the prior probability of a given sentence being interpreted as a request. Moreover, we find that a classifier equipped with seven acoustic features can detect the original intent of an utterance with 65% accuracy. Finally, consistent with past work, the duration, pitch, and pitch slope of an utterance emerge both as significant correlates of a speaker's original intent and as predictors of comprehenders' pragmatic interpretation. These results suggest that human and machine comprehenders alike can use prosody to enrich the meaning of ambiguous utterances, such as indirect requests.

Keywords

Indirect requests, pragmatic ambiguity, prosody, inference

Corresponding author:
Sean Trott, Department of Cognitive Science, UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0515, USA.
Email: sttrott@ucsd.edu

Introduction

People often make requests indirectly. For example, "Can you open that window?" is literally a question about the hearer's ability to open the window, but is often intended instead as an implied request for the hearer to do so. Some indirect requests use a highly conventionalized form (in this example, "**Can you** X?"). But others are less conventional, such as "My office is really hot." Indirect requests have been a topic of active research for decades in psycholinguistics (Gibbs, 1979), philosophy (Searle, 1975), cognitive psychology (Holtgraves, 1994), and natural language processing (Perrault & Allen, 1980; Williams et al., 2018). There are at least two major reasons for this.

First, they are exceedingly frequent. In studies eliciting requests from English-speaking participants, anywhere from 80% (Gibbs, 1981) to 97% (Flöck, 2016) were indirect in some way—that is, they used a grammatical construction other than direct imperatives (e.g., "Pass the salt"). In studies of naturally occurring requests (in American English[1]), frequency estimates range from approximately 46% (Flöck, 2016) to 79% (Goldschmidt, 1998). Second, successfully comprehending indirect requests often requires the hearer to make inferences about the speaker's intent, using linguistic and other contextual knowledge, potentially involving diverse cognitive systems. But it still remains largely to be determined what information human comprehenders use to recover the intended interpretation of a potential indirect request.

Previous work suggests that successfully understanding some indirect requests requires the integration of extra-linguistic contextual information. For conventional indirect requests like "Can you X?," comprehenders can use the form of the utterance as a partial cue to its meaning. Consequently, conventional indirect requests might be easier to understand than unconventionally formed ones (Gibbs, 1981), and in some cases the request interpretation may even be the default (Gibbs, 1986). But even conventional indirect requests can pose a challenge: the conventionality of a particular form is still dependent on context (Gibbs, 1986), and canonical forms can lead listeners to misidentify intended questions as requests (e.g., "Can you play tennis?"), especially individuals with anterior aphasia and right-hemisphere brain damage (Hirst et al., 1984).

Less conventional indirect requests, such as "My office is really hot," require the hearer to infer both the speech act (i.e., is it a request?) as well as the intended substance of the request, and are thus thought to incur higher processing costs than both their literal, non-request counterparts (Tromp et al., 2016) and also more conventional indirect requests (Gibbs, 1981). Successful disambiguation of these utterances may benefit from co-speech gesture and eye gaze (Kelly, 2001; Kelly et al., 1999), as well as a representation of what is mutually known across interlocutors (Gibbs, 1987; Trott & Bergen, 2019, 2020).

Indirect requests have also proven challenging for machine language understanding. Wizard-of-Oz style experiments, in which an apparently autonomous machine is controlled by a human operator, show that human speakers continue to use indirect requests when speaking to what they believe to be autonomous robots (Briggs et al., 2017), even when those robots demonstrably cannot understand them (Williams et al., 2018). Current state-of-the-art solutions (Briggs et al., 2017) use rules relating utterance forms to situational contexts to probabilistically derive the intended interpretation of ambiguous utterances like "Can you knock down the red tower?" These solutions work well for established utterance-context mappings, but there are still considerable gaps in the ability of machines to comprehend indirect requests, particularly non-conventional indirect requests.

The subtle and multifaceted nature of indirect requests seems to demand a heavy reliance by the comprehender on context, along with cognitively costly processes like reasoning about the mental states of others. This makes it challenging to provide a systematic account of how humans rapidly

infer a speaker's intended meaning in everyday language use, and to formalize these computations for machine language comprehension. But to the extent that more superficial but systematic cues to an utterance's meaning are available, these cues could aid both human and machine comprehenders in more quickly deciphering a speaker's intent.

One promising but currently under-explored source of disambiguating information for indirect requests is prosody: the intonational, rhythmic, and tonal properties of how an utterance is spoken or signed.

1.1 Prosodic cues for disambiguation

One reason to believe that prosody may disambiguate indirect requests is that previous work on other kinds of linguistic ambiguity has demonstrated a relationship between prosodic cues and speaker intent.

The evidence is strongest for syntactic ambiguity. Several early studies (Beach, 1991; Price et al., 1991) found that prosodic features such as pitch and pause duration can aid participants in identifying the intended parse of sentences involving temporary structural ambiguity. This boost in comprehension may even occur before the ambiguity is encountered, as suggested by differences in the visual scan patterns of listeners tasked with determining which object a speaker was referring to (Snedeker & Trueswell, 2003). Nonetheless, there are still substantive debates about the conditions under which speakers reliably produce such cues—that is, whether the mechanism underlying prosodic differentiation of string-identical syntactic structures is automatic (i.e., a stored pairing of syntactic construction and prosodic signature) or more strategic (i.e., produced selectively, as a function of audience design) (Allbritton et al., 1996; Schafer et al., 2000; Snedeker & Trueswell, 2003; Speer et al., 2011). Regardless, the evidence suggests that when such cues to meaning are available, listeners improve at identifying the intended syntactic parse—pointing to a clear role for prosodic features in syntactic disambiguation.

There is also a growing body of evidence that prosody helps a comprehender decipher a speaker's pragmatic intentions. Early work argued that certain intonational features and contours are reliably associated with the intended pragmatic interpretation of an utterance; in English, for example, a declarative sentence produced as an assertion typically has a falling pitch, while the same sentence produced as a question or covert request has a rising pitch (Pierrehumbert & Hirschberg, 1990). This is corroborated by Ward (2019, Chapter 6), who suggests that declarative sentences uttered as requests (e.g., "It's cold in here") are accompanied by a "late pitch peak" (i.e., a rise in pitch toward the end of an utterance). Accordingly, computational work (Shriberg et al., 1998; Sridhar et al., 2009) has found that including prosodic features from conversational speech (including duration, pause, fundamental frequency [F0], energy, and speech rate) improves a classifier's ability to categorize utterances by dialogue act, above and beyond a model equipped with only statistical word-level features. While these results do not indicate that human comprehenders infer a speaker's intentions on the basis of prosodic-level features, they do suggest that such features are, in principle, useful.

There is also some evidence that such features can be used by human comprehenders to decipher a range of pragmatic intents, including interrogatives and declaratives in Italian (D'Imperio & House, 1997) and interrogatives in Swedish (House, 2003). More recently, Hellbernd and Sammler (2016) asked whether trained German speakers could produce cues that identified the intended speech act of one-word utterances—for example, producing the word *Bier* ("beer") as a warning, criticism, or suggestion. In a behavioral task, human listeners successfully identified the speaker's intended speech act for 82% of words (and 73% of non-words). The authors also trained a machine learning classifier to categorize speech act using prosodic features (duration, mean intensity,

harmonics-to-noise ratio, mean fundamental frequency, and pitch rise), obtaining 92% accuracy for words (and 93% for non-words).

Additional evidence that people use prosody to disambiguate comes from research on irony detection. Listeners can identify the presence (or absence) of irony in spontaneously produced speech from radio shows when presented in auditory, but not written, format (Bryant & Fox Tree, 2002); in particular, sarcasm in English has been correlated with lower mean F0 (Cheang & Pell, 2008). More recent studies (Deliens et al., 2018) have confirmed that prosodic features aid in the detection of irony; however, listeners appear to exhibit a speed/accuracy trade-off in the integration of prosodic versus contextual congruity cues, respectively.

Finally, beyond the level of individual speech acts, prosodic features have been shown to improve the detection of a speaker's attitudinal stance and emotional state (Jiang & Pell, 2017; Ladd et al., 1985; Pell et al., 2018; Ward & Hirschberg, 1985; Ward et al., 2017, 2018), even in a foreign language (Pell et al., 2009). Features such as speech rate and pitch can also influence judgments about the perceived politeness of a speech act, including requests (Caballero et al., 2018; Culpeper, 2011), though as has been pointed out, the information conveyed by a given prosodic feature is not necessarily independent from the social-interactional context in which that feature is observed (Culpeper et al., 2003; Wichmann, 2000, 2002).

Together, these findings indicate that speakers are capable of producing signals whose prosodic features provide information about the intended syntactic parse or pragmatic interpretation. Critically, these signals are detectable and reliable enough to be useful to both human and machine comprehenders. However, it is currently unknown whether and when speakers and hearers use prosody to overcome the pragmatic ambiguity intrinsic to indirect requests. This gap is all the more notable since, as observed earlier, indirect requests are the most common way that requests are formulated in American English.

Based on current evidence about prosody and disambiguation more generally, we can delineate several key questions about the role prosody might play in the communication of indirect requests. The first concerns the informativity of prosody. Given the wealth of evidence that prosody aids syntactic disambiguation, it seems likely that prosody is at least partially informative of pragmatic Intent—but the precise reliability of prosody as a cue to a speaker's intended interpretation has not yet been quantified in the case of indirect requests. Addressing this question would help inform exactly which resources comprehenders might bring to bear on the task of interpreting ambiguous input. The second question concerns the circumstances under which prosodic cues are used. On one hand, it could be that the same acoustic cues reliably signal the intended speech act, regardless of factors such as the grammatical form of the utterance; that is, there are stereotypical prosodic features consistently associated with particular pragmatic meanings (e.g., request vs. assertion). On the other hand, different cues may be predictive for different grammatical forms—suggesting that speakers (and hearers) may associate distinct grammatical constructions with different suites of prosodic features to communicate Intent. It could even be that prosody is used to disambiguate the intent of certain grammatical forms but not others—that is, perhaps speakers use prosody strategically.

We addressed these outstanding issues in the current work through several core questions. First, can speakers produce reliable cues to indicate to human listeners whether or not they are making a request? More specifically: how accurately can a human listener recover the intended interpretation of an utterance from the prosody alone? Second, which acoustic cues predict a speaker's intended interpretation? And third, are these the same cues that predict a hearer's actual interpretation? In answering these questions, we focus on two different grammatical constructions that can be used to make indirect requests in English: modal interrogatives (e.g., "Can you open that window?") and declarative statements (e.g., "My laptop is broken"). We use a combination of methods, including a production task with untrained speakers, computational analyses of the

prosodic features extracted from these recorded utterances, and a perception experiment in which participants must determine the intent of an utterance.

Note that all critical data, as well as the code to reproduce the analyses described below, can be found online at: https://github.com/seantrott/pros_scaled.

2 Norming study

We first devised a set of 12 potential indirect requests, each with at least two distinct pragmatic interpretations (e.g., a request vs. a yes/no question, or a request vs. an assertion). Six of these sentences were modal interrogatives (e.g., "Can you open that window?"), and six were declarative statements (e.g., "My office is really hot"). We then asked about the likelihood of each sentence being interpreted as a request, independent of any prosodic features associated with the sentence. Our goal with this norming study was twofold. First, it would allow us to measure the Prior Probability of a given sentence being interpreted as a request; in future experiments, we could then compare the effect of this prior probability on comprehenders' pragmatic interpretations to the effect of a speaker's prosody. Second, we could ask whether the proportion of request interpretations varied systematically as a function of grammatical form (Form).

2.1 Methods

2.1.1 Participants. We recruited 79 participants from Amazon Mechanical Turk (42 females, 37 males). We aimed to recruit 80 participants, but Mechanical Turk under-sampled to 79. The mean age was 37 ($SD = 14$), and ranged from 18 to 69. One participant was not a native speaker of English, so we removed them from the analysis, resulting in a total of 78 participants. The experiment took on average 3.6 minutes to complete (median = 2.5, $SD = 4.97$), and participants were paid US$0.75 for participating, translating to an average of US$12.45 per hour.

2.1.2 Materials. There were 12 critical sentences, each with at least two distinct pragmatic interpretations—one of which was always a request. There were six conventional items, formatted as modal interrogatives (e.g., "Can you lift that box?") and six non-conventional items, formatted as declarative statements (e.g., "My soup is cold"). See Supplementary Table 1 for the complete list of critical items, along with the proportion of request interpretations across all participants.

In addition, there were 12 filler items: 6 were formatted as direct requests (e.g., "Please pass the salt"), and 6 were formatted as propositional statements unlikely to be interpreted as requests (e.g., "Cats are a kind of mammal"). See Supplementary Table 2 for the complete set of filler items.

2.1.3 Procedure. Participants were instructed that they would read a series of sentences. For each sentence, their task was to answer whether they thought the sentence was a request (indicated by pressing either "Yes" or "No"). Each participant performed 24 trials (12 critical items, 12 fillers), presented in random order. We also collected information about each participants' self-reported age, gender, and whether or not they were a native speaker of English. The experiment was implemented using JsPsych (de Leeuw, 2015).

2.2 Results

All analyses were performed in R (R Core Team, 2017), using the *lme4* package (Bates et al., 2014). The analyses below were performed only on critical trials (e.g., conventional vs. non-conventional forms).

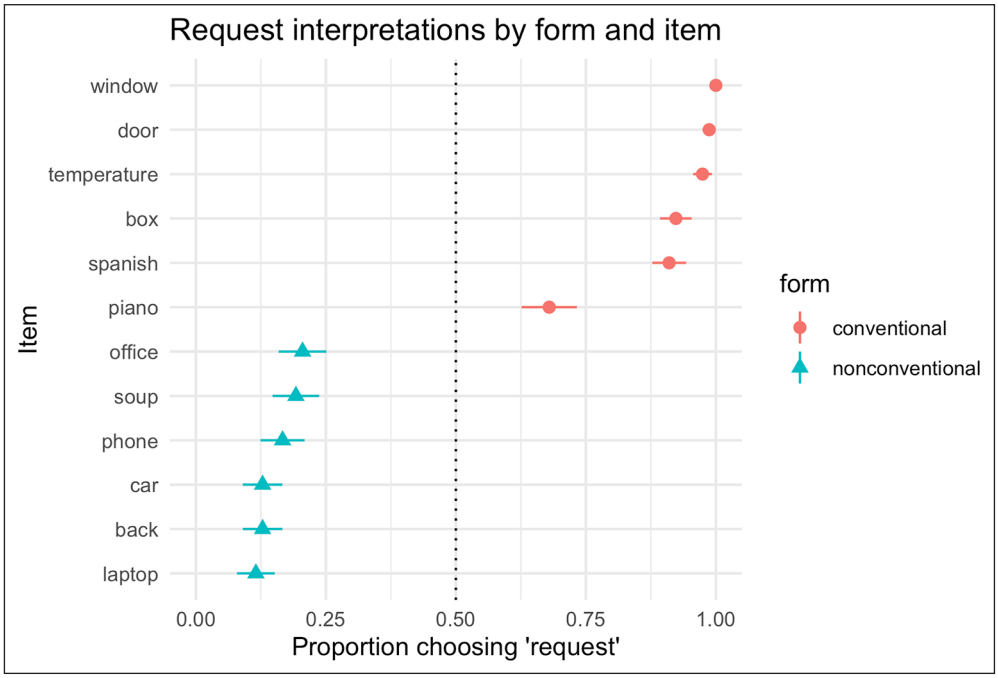

Figure 1. Mean proportion of request responses for each item in the norming study. Conventional items were all more likely than chance to be interpreted as requests, with some variability (e.g., "Can you play the piano?"), whereas non-conventional items had lower likelihoods of being interpreted as requests.

Our primary statistical question was whether sentence Form predicted participants' pragmatic interpretations (Response). We constructed a logit mixed effects model with Response (Yes vs. No) as a dependent variable, Form as a fixed effect, by-subject random slopes for the effect of Form, and random intercepts for subjects and items. Model fit was significantly improved by the effect of Form, $\chi^2(1) = 47.5$, $p < .001$. A total 91% of conventional trials received a "Yes" response, as compared to only 16% of non-conventional trials. Figure 1 illustrates the clear effect of Form, as well as evidence of by-item variability. For example, while conventional items were much more likely to be interpreted as requests than non-conventional items, utterances like "Can you play the piano?" exhibited both more variability and a lower request likelihood. See Supplementary Figure 1 for a visualization of request interpretations including filler items.

2.3 Discussion

The goal of the Norming Study was twofold. First, we aimed to measure the prior probability of each sentence being interpreted as a request (independent of its prosody). This would allow us to ask whether the presence or absence of particular prosodic features explain comprehenders' interpretation of an ambiguous sentence above and beyond the sentence's prior probability of being interpreted as a request. Second, we were interested in whether prior probability varied by Form. As expected, conventional items (i.e., modal interrogatives) were much more likely to be interpreted as requests than non-conventional items (i.e., declarative statements), consistent with past work (Gibbs, 1986) suggesting that grammatical form plays an important role in the "default" interpretation of a potential indirect request.

3 Production study

Our first research question was whether speakers can reliably produce prosodic cues that aid human listeners in recovering their intent. That is, given a sentence (e.g., "Can you open that window?") with two salient pragmatic interpretations (e.g., a request vs. a yes/no question), can speakers deploy systematic acoustic cues to convey a particular interpretation of that sentence?

We collected audio recordings from 18 native English speakers.[2] Each speaker produced two versions of the twelve critical sentences—one request and one non-request (e.g., a yes/no question or a statement, depending on the grammatical form); see the "Procedure" section for more details on the recording process. Then, we extracted seven acoustic features from each recorded utterance, and measured the predictive power of these acoustic features.

Based on prior research characterizing the relationship between prosody and speaker intent both in general (Cheang & Pell, 2008; Hellbernd & Sammler, 2016; Pierrehumbert & Hirschberg, 1990; Shriberg et al., 1998; Sridhar et al., 2009) and for indirect requests in particular (Banuazizi & Creswell, 1999; Hedberg et al., 2014; Trott et al., 2019), we focused on utterance-level acoustic features relating to pitch (fundamental frequency, or F0), duration, and intensity. Following Hellbernd and Sammler (2016), we extracted the mean F0, the range of F0, mean intensity, and the number of voiced frames (as a proxy for duration[3]). We also included measures of dispersion for both pitch (standard deviation of F0) and intensity (standard deviation of intensity). Finally, because the non-request interpretation of the conventional utterances (e.g., "Can you open that window?") was a yes–no question about the hearer's ability, and because yes–no questions in English typically have a low-rise pitch contour (Banuazizi & Creswell, 1999; Hedberg et al., 2014; also see the study by Geluykens, 1988), we used the slope of the F0 component (slope of regressing F0 ~ time) as a proxy for the degree to which an utterance exhibited a rising or falling contour (Roche et al., 2019). All measures were taken across the entire sentence; we did not analyze word-level acoustic features (e.g., focal stress) in the current work (see the "Limitations and future work" section for a discussion of this avenue).

We predicted that for conventional items (i.e., modal interrogatives), requests would have a less positive F0 slope than their non-request counterparts (i.e., a yes/no question), given both the results of past work (Trott et al., 2019) and the fact that yes/no questions are typically associated with a low-rise pitch contour (Banuazizi & Creswell, 1999; Hedberg et al., 2014; Pierrehumbert & Hirschberg, 1990). We also predicted that non-conventional utterances should have a longer duration (i.e., more voiced frames) when intended as requests than when intended as literal statements. Because the default interpretation of these declarative sentences was as literal statements, we expected that speakers might "mark" a deviation from this default interpretation by emphasizing specific words or syntactic constituents; if this is the case, it should be detectable in the utterance-level measure of number of voiced frames. Further support for this prediction comes from previous analyses carried out in a smaller pool of speakers (Trott et al., 2019). Finally, we also predicted that both conventional and non-conventional requests should have a higher mean intensity than their non-request counterparts; this prediction was also based on previous work (Trott et al., 2019).

All aspects of the recording design, exclusion criteria, and statistical analyses were preregistered on OSF (https://osf.io/34fc7).

3.1 Methods

3.1.1 Participants. We recruited 24 speakers of American English from the UC San Diego Psychology Department Subject Pool.

The pre-registration included the following exclusion criteria: (1) participants who self-reported as non-native speakers of English and (2) participants for whom the entire set of items was not recorded (e.g., because of recording errors or missing data). Six speakers were excluded because of recording errors or missing data, resulting in a total of 18 speakers. The mean age of these participants was 19.94 ($SD = 1.76$, median = 19.5). Eight speakers identified as females (10 males).

3.1.2 Materials. There were 12 sentences total, each of which could be plausibly interpreted as either a request or as a non-request (i.e., a statement or a question); see Supplementary Table 1 for a complete list, and Figure 1 for an illustration of the norming data for these items.

The sentences were identical to those used in the Norming Study. Six were conventional, with the modal interrogative form "Can you X?," for example, "Can you close that door?." Six were non-conventional, with the form of "My X is Y," for example, "My phone is dying."

3.1.3 Procedure. After signing the audio release and consent forms, speakers were brought to a sound-attenuated room, which contained a computer monitor, chair, and microphone. They were instructed to sit down and remain at a constant distance from the microphone. Before the recording began, speakers were given examples demonstrating how the same sentence could be used as either a request or statement (or question). These example sentences did not appear in the target stimuli, nor were the different versions spoken aloud to participants (so that participants could not simply imitate the prosody of the experimenter).

During the experiment, the target sentence appeared on the monitor screen. Each speaker produced two versions of all 12 sentences. Speakers were instructed to say each utterance twice—once as a request and once as a literal question or statement (counterbalanced for order). They were allowed to produce each version (i.e., request vs. non-request) multiple times; when they produced a version they were satisfied with, they indicated that they were done with the item. This final version was the one used in subsequent work.

Speakers were not given details about the intended recipient of the request or the situation in which the request was produced. Finally, the order of the target sentences was randomized across participants.

The recordings can be found on GitHub: https://github.com/seantrott/pros_scaled.

3.1.4 Data processing. For each of the 432 recordings (18 speakers producing 12 utterances with two versions each), we used Parselmouth (Jadoul et al., 2018), a Python interface to Praat, to extract the seven acoustic features. We then *z*-scored each of these variables with respect to each speaker's mean and standard deviation for that particular feature, to account for considerable variability between speakers overall.

The full set of extracted features, along with the original audio recordings, can be found here: https://github.com/seantrott/pros_scaled.

3.2 Results

3.2.1 Analysis of individual acoustic features. First, we asked how much independent variance was explained by each feature in turn, comparing a full model (including all seven features) to a model omitting only the feature under consideration. In each case, the full model included Intent (request vs. non-request) as a dependent variable and each of the seven acoustic features as predictors; each model also included by-item random intercepts. We adjusted for multiple comparisons using Holm–Bonferroni corrections (Holm, 1979); we report only the adjusted *p*-values below. In each case, a positive coefficient represents a higher likelihood of a request, while a negative coefficient

represents a higher likelihood of a non-request. Note that in each case, the full model contained all seven features, even though we only had specific, directional hypotheses about some of the acoustic features. We made this analytical decision for two reasons. First, in previous work (Trott et al., 2019), speaker Intent was predicted by different acoustic features for different subset analyses of the data (e.g., only conventional items vs. all items), but each model controlled for all the acoustic features under consideration; thus, one reason to include all seven features was to more closely replicate past work. Second, because this analysis was in part exploratory, we opted for the more inclusive approach when pre-registering the analysis to ensure that we could detect novel correlations that were not identified in past work (Trott et al., 2019).[4]

When predicting Intent across conventional and non-conventional items, only F0 slope emerged as a significant predictor after controlling for multiple comparisons, $\chi^2(1) = 8.36$, $p = .02$. Specifically, items with more positive F0 slopes were less likely to be requests ($\beta = -0.33$, $SE = 0.12$).

For conventional items only, model fit was improved by the inclusion of both number of voiced frames, $\chi^2(1) = 15.93$, $p < .001$, and mean intensity, $\chi^2(1) = 7.9$, $p = .03$. Items with a larger number of voiced frames were less likely to be requests ($\beta = -0.75$, $SE = 0.2$), whereas items with a higher mean intensity were more likely to be requests ($\beta = 0.45$, $SE = 0.17$). The predicted relationship between F0 slope and Intent was not statistically significant after controlling for multiple comparisons ($p = .2$), though it was in the predicted direction ($\beta = -0.33$, $SE = 0.17$).

Finally, for non-conventional items only, model fit was improved by both number of voiced frames, $\chi^2(1) = 29.06$, $p < .001$, and mean F0, $\chi^2(1) = 24.99$, $p < .001$. As predicted, utterances with a larger number of voiced frames were more likely to be requests ($\beta = 0.94$, $SE = 0.19$). Utterances with a larger mean F0 were also more likely to be requests ($\beta = 1.18$, $SE = 0.26$).

3.2.2 Machine learning classifier. The analysis above reveals which features are informative about Intent, but does not directly indicate how much information these features contain—particularly when all seven are combined. One simple way to address this question is to quantify the ability of a machine learning classifier equipped with all seven features to predict the Intent of held-out test items; that is, how successfully does a model generalize to novel samples? To quantify test item accuracy, we used leave-one-out cross-validation (LOOCV). In this procedure, a model is fit to all items in the dataset but one; the model is then used to classify the held-out test item, allowing us to determine whether the model successfully generalized (i.e., whether the predicted label from the model matches the actual label for the held-out test item). This leave-one-out procedure is performed for every item in the dataset, ultimately giving us an accuracy score: the percentage of held-out items that were correctly classified.

In our case, this amounted to 432 splits of the data, corresponding to each of the 432 utterances. For each split, we fit a logistic regression classifier to the 431 training utterances. The classifier was trained to predict an utterance's original Intent from all seven acoustic features and their interaction with Form. This classifier was then used to predict the Intent of the held-out test item. The classifier successfully predicted Intent on 65% of held-out test items, a rate substantially above chance (50%).

As shown in Figure 2, held-out test items that were estimated as being more likely to be Requests (i.e., a larger value for *p*[request]) were, in fact, more likely to have originally been intended as requests. This finding demonstrates that the learned relationship between the set of acoustic features and a speaker's Intent is both systematic and generalizable, at least with respect to held-out test items from the dataset. (For comparison, previous work (Trott et al., 2019) used an identical method on a smaller pool of audio samples and obtained held-out performance of 74%.)

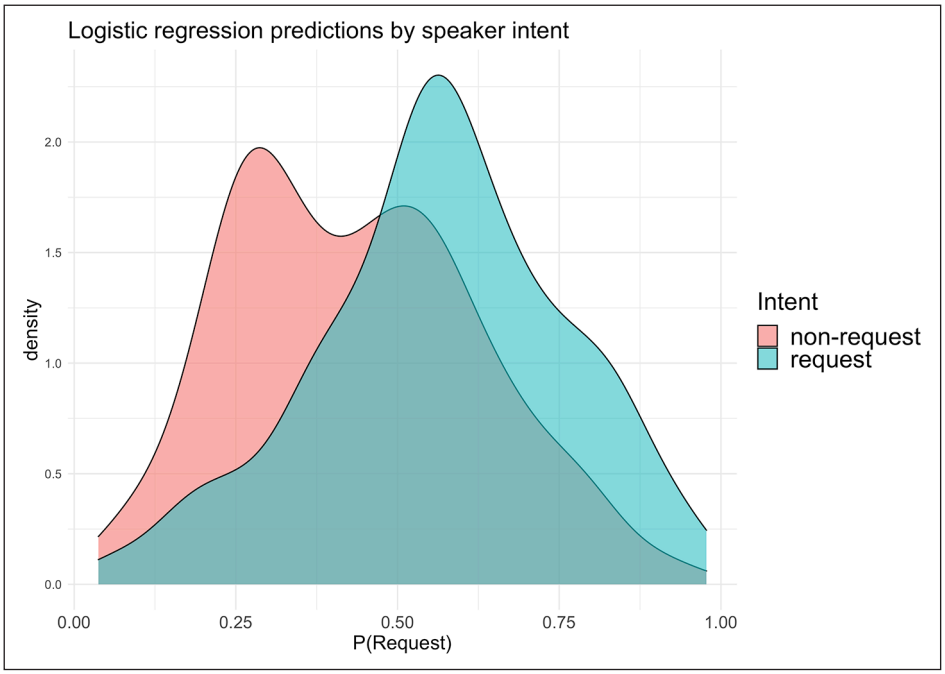

Figure 2. A classifier equipped with all seven acoustic features correctly predicted a held-out test item's true Intent 65% of the time; figure illustrates the distribution of classifier probabilities over classes (e.g., request, non-request), colored by the original Intent of a given item.

3.3 Exploratory analysis of final rise

The analyses described above used acoustic features that were calculated across the entire utterance, limiting their sensitivity to fine-grained prosodic features that might be present only on particular segments of the utterance. In particular, F0 slope was intended as a measure of rising intonation, where a more positive slope should reflect a low-rise intonation—but such a measure might fail to detect more subtle, nonlinear changes in slope that nonetheless signal pragmatic information to listeners (e.g., an initial rise, followed by a fall–rise contour). Thus, we conducted an exploratory (i.e., non-pre-registered) analysis investigating the presence or absence of a final rise in each utterance.

3.3.1 Data coding and processing. First, we inspected all 432 critical utterances by hand in Praat and identified the timestamps corresponding to the beginning and end of either the final Noun Phrase, or NP (for modal interrogatives, e.g., "that box"), or the final word (for declarative utterances, e.g., "broken").

Then, using the hand-coded timestamps, we automatically extracted two acoustic features from the final segment of each utterance using Parselmouth: the F0 slope (i.e., whether the pitch of the final segment increases or decreases over time) and degree of rise (i.e., the difference between the average F0 of the final frame and the average F0 of the initial three frames).

3.4 Analysis and results

All analyses were conducted in R (R Core Team, 2017) using the *lme4* package (Bates et al., 2014). For each set of analyses described below, we considered two dependent variables: F0 slope and

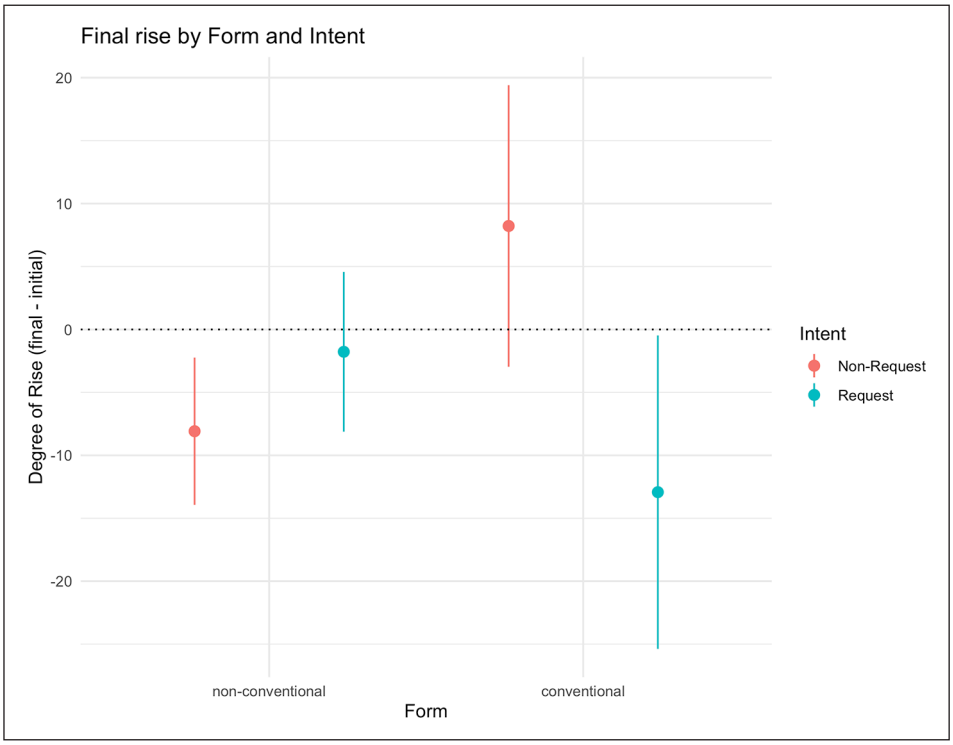

Figure 3. Conventional items (i.e., modal interrogatives) showed a significant difference in degree of rise (i.e., final F0 − initial F0) of the final segment as a function of Intent, while non-conventional items (i.e., declaratives) did not. Conventional items intended as requests had less of a final rise than those intended as yes/no questions (i.e., non-requests).

degree of rise. For each set of analyses, we corrected for multiple comparisons using Holm–Bonferroni correction (i.e., for the two dependent variables considered).

In the first set of analyses, we asked whether a model including an interaction between Form (conventional vs. non-conventional) and Intent (request vs. non-request) explained additional variance in either dependent variable (degree of rise or F0 slope) than a model omitting that interaction. In each case, the full model contained the interaction between Form and Intent (as well as main effects of each), by-speaker slopes for the effect of Form and Intent, and random intercepts for speakers and items. The interaction significantly improved model fit in the case of predicting degree of rise, $\chi^2(1) = 10.42$, $p = .002$, but not when predicting F0 slope ($p = .19$). That is, the effect of Intent on degree of rise depended on the grammatical construction being used; this is illustrated in Figure 3.

We then conducted a series of subset analyses, separating the set of conventional items (216 utterances) from non-conventional items (216 utterances). For each set of items, we asked whether a model including a fixed effect of Intent explained additional variance over a model omitting only that predictor, again considering both dependent variables (degree of rise and F0 slope). All models contained by-subject and by-item random slopes for the effect of Intent, and random intercepts for speaker and item.

For conventional items, the inclusion of Intent significantly improved model fit for both degree of rise, $\chi^2(1) = 4.56$, $p = .03$, and F0 slope, $\chi^2(1) = 5.78$, $p = .03$. Requests were correlated with a less positive degree of rise ($\beta = -21.15$, $SE = 8.82$) and less positive F0 slope ($\beta = -0.196$, $SE = 0.07$).

For non-conventional items, the inclusion of Intent did not improve model fit for either degree of rise or F0 slope (after correcting for multiple comparisons).

3.5 Discussion

We conducted several exploratory analyses of acoustic features associated with the final NP (for modal interrogatives) or final word (for declarative statements) of each utterance. Based on the analysis of coarse-grained features in the primary manuscript, we expected to find a difference in measures of the final rise (either F0 slope or degree of rise) for conventional items, but not necessarily non-conventional items, given that we observed no differences related to pitch contour for the latter.

We found that conventional items (i.e., modal interrogatives) intended as requests had less of a rise in their final NP (as measured by the difference in F0 between the final frames and initial frames) than those intended as non-requests (see Figure 3). This is consistent with our predictions, and with previous research arguing that yes/no questions are more likely to exhibit a rising intonation; see also Figure 4, which shows the entire pitch contours for the request versus non-request versions of the same sentence (produced by the same speaker). Conventional items intended as requests also exhibited a significantly less positive F0 slope, though the effect of Intent on F0 slope was not significantly different across conventional and non-conventional items (i.e., the interaction term was not significant).

Of course, these results should be interpreted with caution, given that they were exploratory. Furthermore, although the features were more granular than those used in the pre-registered analyses, they still have limitations: they will still fail to capture subtle non-linearities in the pitch contour, and the length of the final NP (or word) was not controlled across stimuli (e.g., “that box” vs. “the television”).

4 Behavioral experiment

The analysis of acoustic features above indicates that, consistent with past work (Trott et al., 2019), several utterance-level acoustic cues are in fact predictive of a speaker’s Intent—and combined, the seven acoustic features considered can reliably predict the intent of a held-out item with 65% accuracy. However, this leaves open the question of whether—and to what extent—human comprehenders can also exploit reliable cues in the prosodic signal to determine a speaker’s intended interpretation.

To address this question, we ran a behavioral experiment in which participants listened to the utterances described above. In each trial, comprehenders were presented with a single utterance and asked to indicate via button-press whether the speaker intended that utterance as a request. This design allowed us to determine whether comprehenders could successfully identify whether a given utterance was intended as a request on the basis of its prosody alone. Furthermore, we could ask whether a speaker’s intent predicted participants’ pragmatic interpretations above and beyond the prior probability of a given sentence being interpreted as a request (see the Norming Study).

All aspects of the experimental design and statistical analyses were pre-registered on the Open Science Framework (https://osf.io/mx64e).

4.1 Methods

4.1.1 Participants. We aimed to recruit 80 participants from the UC San Diego Psychology Department Subject Pool. We over-sampled to 82 participants; one participant was excluded

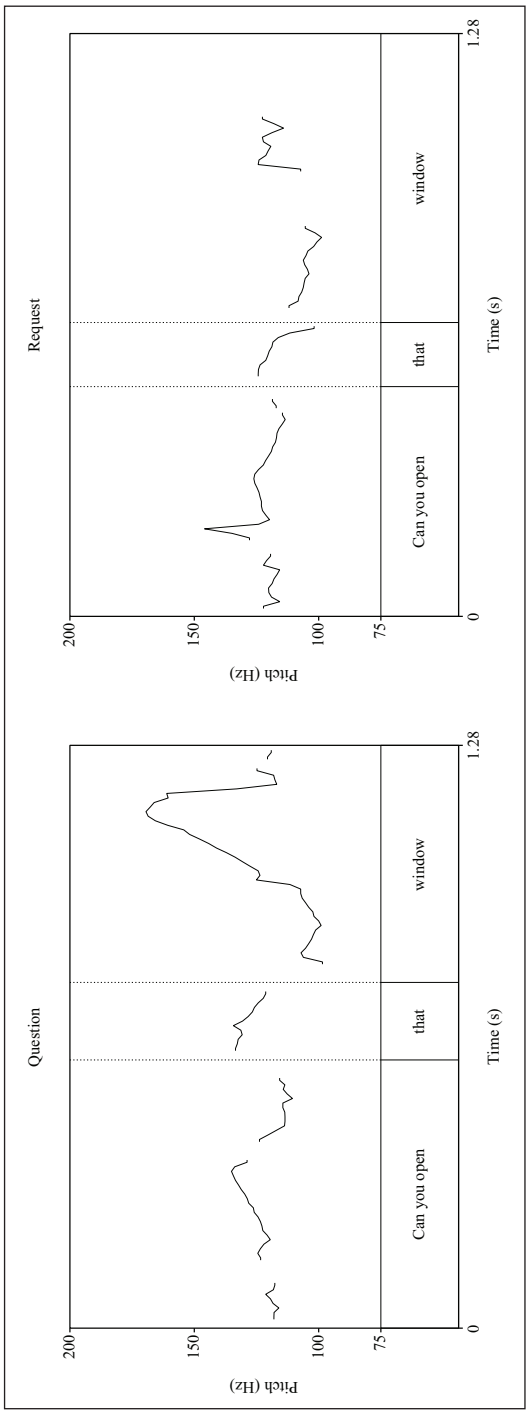

Figure 4. Pitch contours for the utterance "Can you open that window?," produced by the same speaker as a yes–no question (non-request) and request. The question version (left pane) shows more evidence of a final rise than the request version (right pane), and is also longer (consistent with the non-request modal interrogatives having a larger number of voiced frames).

because of self-report as being a non-native speaker of English. Of the final set of 81 participants, 62 self-identified as females (17 males, 1 non-binary, and 1 preferred not to answer). The average self-reported age was 22.6 (median = 20). One participant self-reported as having an age of 220; our pre-registration did not contain any exclusion criteria on this basis, so the participant was included in the analyses below (although we assume the reported age was a typo). The main experiment took on average 7.34 min to complete (median = 5.5 min). Each participant received one course credit for participating.

4.1.2 Materials. We used the 432 utterances recorded and analyzed above (see Production Study for more details). All recordings can be found on GitHub: https://github.com/seantrott/pros_scaled.

4.1.3 Procedure. After completing an audio check, participants were instructed that they would listen to a series of utterances. For each utterance, their task was to determine whether or not the speaker was making a request. Participants indicated their response via button press ("Yes" or "No"). All participants performed the study online.

Participants were assigned to one of two lists, with each list corresponding to conventional (37 participants) or non-conventional (44 participants) items. Each participant heard all versions of all items for a given Form by nine speakers, where those nine were randomly sampled from the total set of 18 speakers. Participants heard each utterance exactly once. This resulted in a total of 108 trials per participant (nine speakers producing six sentences with two versions each). The trials were blocked by speaker, with the order of each item randomized within-block, and the order of each speaker-block randomized. After completing all 108 trials, participants also reported their age, gender, and whether or not they were a native speaker of English.

The experiment was implemented using JsPsych (de Leeuw, 2015).

4.2 Results

All analyses were performed in R (R Core Team, 2017), using the *lme4* package (Bates et al., 2014). Random effects structure was determined by beginning with the maximal model, then reducing as needed for model convergence (Barr et al., 2013). Results were obtained using nested model comparisons. As noted earlier, the pre-registration for these analyses can be found on OSF (https://osf.io/mx64e).

We asked whether participants could reliably detect whether a given utterance was intended as a request or non-request at a rate above chance—that is, whether a speaker's original intent (request vs. non-request) predicted participant response (yes vs. no), above and beyond the form of the utterance (conventional vs. non-conventional) and its prior probability of being interpreted as a request (see Norming Study). Thus, we constructed a *glmer* model with a logit link with response (yes or no) as a dependent variable, fixed effects of intent, form, and prior probability, by-subject random slopes for the effect of both intent and prior probability, by-item random slopes for the effect of intent, and random intercepts for subjects, items, and speaker. Crucially, this full model explained significantly more variance than a model omitting only the fixed effect of Intent, $\chi^2(1) = 14.07$, $p < .001$. Request interpretations were significantly more likely for utterances originally intended as a request ($\beta = 0.6$, $SE = 0.13$). This indicates that comprehenders can recover a speaker's intended interpretation at a rate above chance (overall accuracy was approximately 55%).

Participants' pragmatic interpretations were also strongly correlated with the prior probability of a given sentence being interpreted as a request ($\beta = 6.9$, $SE = 0.97$). Crucially, however, intent

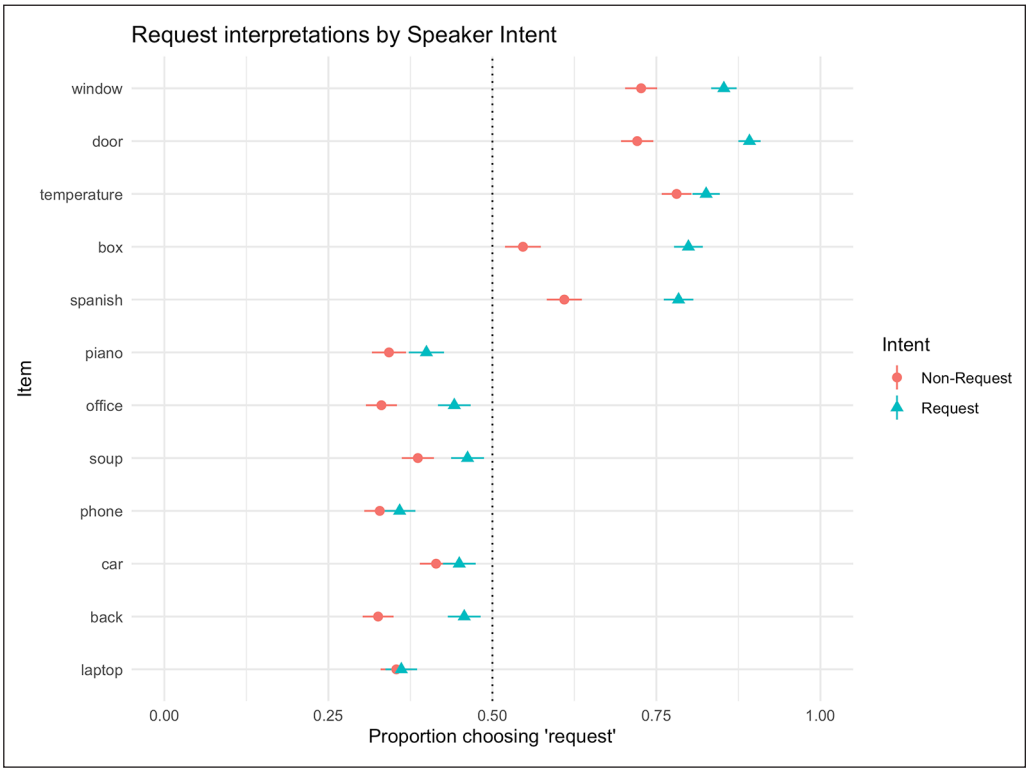

Figure 5. Proportion of request interpretations by item, by speaker intent (request vs. non-request). Items are arranged according to their prior request probability. With one exception ("My laptop is broken"), the rate of request interpretations is higher for each item when the utterance was originally intended as a request, indicating that participants are using reliable prosodic information to infer intent.

influenced their interpretations above and beyond this prior probability, as did a sentence's grammatical form (see Figure 5).

5 Do acoustic cues predict pragmatic interpretations?

The results of the behavioral experiment demonstrate that human comprehenders can exploit reliable cues in the prosodic signal to determine a speaker's intended interpretation. Furthermore, the analysis of acoustic features indicates that, consistent with past work (Trott et al., 2019), several utterance-level acoustic cues are in fact predictive of a speaker's intent—and combined, the seven acoustic features considered can reliably predict the intent of a held-out item with 65% accuracy. This leaves open a third question: which acoustic cues are predictive of a comprehender's pragmatic interpretation? In a pre-registered analysis (https://osf.io/mx64e), we used the acoustic features extracted in the section above to ask whether human comprehenders systematically modulated their pragmatic interpretations as a function of these acoustic cues.

Based on past work (Trott et al., 2019), theoretical and empirical links between prosody and pragmatic intent, and a previous, unpublished pilot study, we had several predictions about which cues would be predictive of a comprehender's response. First, we predicted that for conventional items (i.e., modal interrogatives), Request interpretations should be less likely for utterances with

a more positive F0 slope. Here, a more positive F0 slope should instead signal that the speaker intends the utterance to be interpreted as a yes/no question. Second, we predicted that for conventional items, items with a higher mean F0 should be less likely to elicit request interpretations. Third, we predicted that for conventional items, longer utterances (i.e., a larger number of voiced frames) should be less likely to elicit Request responses. One explanation for the predictions about mean F0 and number of voiced frames is that the modal verb (e.g., "can") is marked when a speaker intends a yes/no question interpretation—the less likely of the two interpretations for modal interrogatives (see the Norming study). While both measures (mean F0 and number of voiced frames) are taken across the entire utterance, they will necessarily reflect marking on individual lexical or syntactic constituents. Finally, we predicted that utterances with a larger mean intensity should be more likely to elicit request interpretations across both conventional and non-conventional items.

5.1 Results

We conducted a series of nested model comparisons in R. We began with a full model, with interpretation (request vs. non-request) as a dependent variable, fixed effects of all seven acoustic features (as well as their interaction with Form), and random intercepts for subjects, item, and speaker. We then compared that full model to a series of reduced models, each of which omitted the interaction between a given acoustic feature and Form; to test for a main effect of that acoustic feature, we also compared this reduced model to a model omitting the acoustic feature entirely. This involved eight model comparisons altogether. We corrected for multiple comparisons using Holm–Bonferroni corrections; we report the adjusted *p*-values below. Note that unlike in the Production Study, we opted to conduct model comparisons only for features that we had specific, directional hypotheses about.

First, the full model explained significantly more variance than a model omitting only the interaction between F0 slope and Form, $\chi^2(1) = 11.42$, $p = .003$; furthermore, the model omitting only the interaction explained more variance than a model omitting F0 slope altogether, $\chi^2(1) = 71.26$, $p < .001$. Items with a more positive slope were less likely to be interpreted as requests overall ($\beta = -0.16$, $SE = .04$), and even less likely when the item in question was a model interrogative, as indicated by a negative coefficient on the interaction between F0 slope and Form ($\beta = -0.22$, $SE = 0.07$). This finding is illustrated in Figure 6.

The interaction between number of voiced frames and Form also improved model fit, $\chi^2(1) = 114.22$, $p < .001$, but the main effect of number of voiced frames was only marginally significant after adjusting for multiple comparisons ($p = .1$). In the full model, a positive coefficient was assigned to the main effect of number of voiced frames ($\beta = 0.24$, $SE = .04$), and a negative coefficient was assigned to the interaction between number of voiced frames and Form ($\beta = -0.73$, $SE = .07$). In other words, longer non-conventional utterances were more likely to be interpreted as requests, while longer conventional utterances were less likely to be interpreted as requests. This finding is depicted in Figure 7.

Third, the interaction between mean F0 and Form significantly improved model fit, $\chi^2(1) = 12.56$, $p = .002$, though the main effect of mean F0 did not ($p > .2$). Specifically, conventional items with higher mean F0 were less likely to be interpreted as requests ($\beta = -0.26$, $SE = .07$). There was no significant effect of mean intensity, nor an interaction between mean intensity and Form.

6 General discussion

Can speakers and comprehenders use prosody to overcome the ambiguity inherent to indirect requests? We found that human comprehenders were able to identify the intended interpretation of

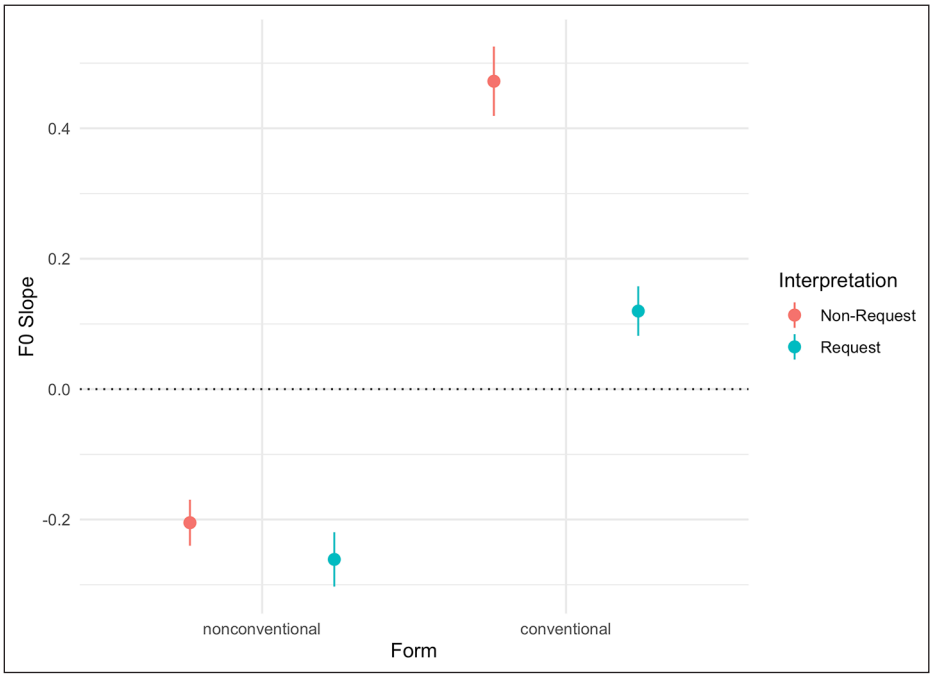

Figure 6. *Z*-scored F0 slope by Form and Interpretation. As predicted, for conventional utterances, non-request interpretations were significantly more likely for utterances with a more positive F0 slope.

an utterance at a rate above chance (55%), even when controlling for the prior probability of a given sentence being interpreted as a request.

We also asked which prosodic features speakers produced as a function of their intent, and which correlated with comprehenders' pragmatic judgments (see Figure 8 for a summary table). We extracted seven acoustic features from each utterance, and asked whether each feature predicted a speaker's original intent (original acoustic features analysis), and whether each feature predicted human interpretations of intent. Number of voiced frames emerged as a significant predictor of both speaker intent and comprehenders' pragmatic interpretations: conventional items (i.e., modal interrogatives) with more voiced frames were associated with non-requests (i.e., yes/no questions), whereas non-conventional items (i.e., declarative statements) with more voiced frames were associated with requests (see Figure 7). Both findings are also consistent with past work conducted with a different sample of speakers (Trott et al., 2019). One explanation is that speakers are deploying this cue to mark a deviation from the expected interpretation of an utterance, given its grammatical form. In both the Norming Study and the behavioral experiment, conventional items were more likely to be interpreted as requests, while non-conventional items were less likely—thus, speakers might use prosody to signal that their intended interpretation is not the canonical meaning by emphasizing particular lexical or syntactic constituents of the utterance, perhaps those associated with the alternative, non-canonical meaning.

As predicted, F0 slope was also predictive of comprehenders' pragmatic interpretations: for conventional items, more positive slopes were more associated with non-request (i.e., yes/no question) interpretations (see Figure 5). This is consistent with past empirical findings (Trott et al., 2019), as well as the theoretical prediction that speakers attempting to convey a yes/no question will emphasize prosodic features associated with that speech act, such as a rising pitch contour

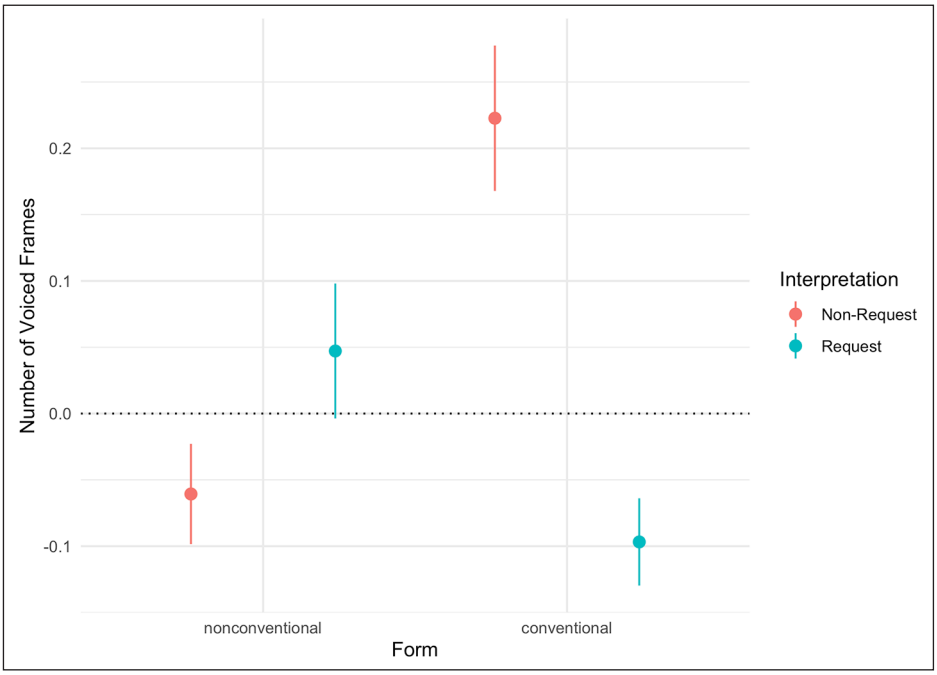

Figure 7. *Z*-scored number of voiced frames by Form and pragmatic Interpretation. As predicted, there was a cross-over interaction: longer conventional items (i.e., declarative statements) were more likely to be interpreted as requests, while longer non-conventional items (i.e., modal interrogatives) were more likely to be interpreted as non-requests (i.e., yes/no questions).

Non-Request (NR) > Request (R) | Request (R) > Non-Request (NR) | No effect

	Conventionality	Predicting Intent	Predicting Interpretation
Mean intensity	Conventional	R > NR	
	Non-conventional		
F0 slope	Conventional		NR > R
	Non-conventional		
F0 duration	Conventional	NR > R	NR > R
	Non-conventional	R > NR	R > NR
Mean F0	Conventional	NR > R	NR > R
	Non-conventional		

Figure 8. Summary table of the four primary acoustic features we had predictions about. Yellow cells indicate that the feature in question was larger for items either intended or interpreted as requests (e.g., F0 duration for non-conventional utterances), and green cells indicate that it was larger for items intended or intended as non-requests (e.g., F0 slope for conventional utterances).

(Banuazizi & Creswell, 1999; Hedberg et al., 2014; Pierrehumbert & Hirschberg, 1990). However, unlike past work, we did not detect a significant relationship between speaker intent and F0 slope for conventional items in the first acoustic features analysis—though there was a significant relationship between intent and F0 slope when the analysis included all items.

As a second-order question, we asked about the predictive power of all features combined: how accurately could a machine learn to classify utterance intent on the basis of prosody alone? Using LOOCV, a machine learning classifier trained on all seven acoustic features (and their interaction with form) successfully identified the intent of potential request utterances 65% of the time. This was higher than human performance on the behavioral task (55%); interestingly, Hellbernd and Sammler (2016) also find that a classifier is more accurate at identifying the speech act of individual words (approximately 92%) than human participants (approximately 82%). One possible explanation for our finding is that the classifier was supervised (i.e., it was given examples of requests and non-requests with particular acoustic features); in contrast, human participants were not given any examples of prosodic features correlated with requests or non-requests, nor were they given feedback as to whether their responses were correct.

The demonstrated utility of prosody for intent recognition (65%) is also consistent with past work (Hellbernd & Sammler, 2016; Shriberg et al., 1998). Thus, these findings also have implications for natural language understanding (NLU) systems, for which intent recognition remains a challenge, particularly for pervasive forms of pragmatic ambiguity such as indirect requests (Briggs et al., 2017; Williams et al., 2018); intent recognition might be improved by equipping existing architectures with a parallel processing stream, which produces a probability distribution over possible intents as a function of specific acoustic features.

6.1 Limitations and future work

Open questions remain. First, the speakers recruited for these studies were all speakers of American English. While there is prior cross-linguistic and cross-cultural work on the frequency and grammatical form of different indirect requests (Blum-Kulka et al., 1989; Holtgraves, 1997; Holtgraves & Joong-Nam, 1990 ; Holtgraves & Yang, 1992; Le Pair, 1996), as well as cross-linguistic consistency and variability in prosody more generally (Fernald et al., 1989; Vaissière, 1983; Yaeger-Dror, 2002), it remains unknown whether the same prosodic features reliably distinguish speaker intent in the case of indirect requests across languages. Thus, future work could also explore the generalizability of the prosodic cues identified in the current work as predictive of speaker intent.

Similarly, we considered only two grammatical forms: modal interrogatives (e.g., “Can you lift that box?”) and declaratives (e.g., “My soup is cold”). Critically, different acoustic features were predictive of intent for each form, and in some cases, the same acoustic feature (e.g., Number of Voiced Frames) was differentially predictive. This suggests that speakers deploy different prosodic cues, or deploy the same prosodic cues in different ways, as a function of which grammatical construction they are using to make a request (or ask a question, etc.). Future work could ask whether these findings generalize to different grammatical forms commonly used to make requests, such as imperatives. Notably, previous work has also demonstrated systematic intonational contours correlated with please-requests (Wichmann, 2004); specifically, the intonation used depends on whether please is placed at the beginning of the request (e.g., “Please open the door”), the middle (e.g., “Could someone please open the door”), or at the end (e.g., “Could you call me please”). Future work could ask how the inclusion of please (and where) influences the utterance-level prosodic features we observed.

Third, the acoustic features we extracted for the pre-registered analyses were taken across the entire utterance. Of course, this is a relatively coarse measure, given that speakers may be

producing cues that are localized to specific syntactic constituents or even words. In an exploratory analysis, we attempted to address this limitation by analyzing several acoustic features associated with the final NP (for conventional items) or word (for non-conventional items). We found that conventional items (i.e., modal interrogatives) intended as requests exhibited less evidence of pitch rise in the final NP than those intended as non-requests (i.e., as yes/no questions); see Figure 3 for an illustration of this difference. However, future work could also analyze the acoustic features associated with other words or syntactic constructions in each utterance. For example, it is possible that speakers are signaling a non-request (i.e., question) interpretation of modal interrogatives like "Can you open that window?" by emphasizing words associated with the semantics of yes/no questions, such as "can" (Hirschberg, 2017); if this is true, word-specific acoustic features such as mean F0 and number of voiced frames should show differentiation as a function of a speaker's intent—and they should also predict a comprehender's pragmatic interpretation.

Relatedly, it is possible that the coarse utterance-level features we analyzed are not themselves what listeners attend to and deploy for pragmatic inference, but that they emerge as a result of (and thus correlate with) these more fine-grained cues; for example, F0 slope and number of voiced frames might be a proxy for focal stress on the modal "can." Importantly, our analyses are correlational and do not demonstrate a causal role for these utterance-level features. A better understanding of the local features that serve to disambiguate intent would also yield a clearer picture of what exactly listeners are perceiving and how this information is integrated with the semantic content of the utterance. Similarly, a more fine-grained analysis of the overall shape of the pitch contour would be useful, potentially using the ToBI framework (Beckman et al., 2004); for example, F0 slope will capture linear increases or decreases in pitch across the length of an utterance, but will not capture nonlinear contours that correlate with distinct meanings, nor will it identify where in the utterance these prosodic cues might diverge.

One long-standing question in the literature on prosody and pragmatic intent is whether particular prosodic features convey intent directly, or whether they function primarily as contrastive markers, which invite the listener to perform additional inference. For example, prosodic features may not directly convey sarcastic intent, but rather prompt listeners to integrate other multimodal, contextual information to recognize irony (Attardo et al., 2003; Bryant & Fox Tree, 2005). Our finding that number of voiced frames interacts with utterance form to predict intent (and pragmatic interpretation) suggests that certain features might serve primarily as contrastive markers, signaling a deviation from the expected interpretation and perhaps emphasizing prosodic signatures of the non-default interpretation; as mentioned above, a more fine-grained analysis of the acoustic features associated with each word might be more revealing about exactly which semantic or pragmatic features of an utterance's meaning speakers are drawing attention to.

Another question concerns the utility of prosody as a cue to speaker intent. Although speaker intent was indeed predictive of comprehenders' pragmatic interpretations in the behavioral experiment, the rate of request interpretations only differed by 10% across conditions (58% for utterances intended as requests, and 48% for utterances not intended as requests). But presumably comprehenders' success rate is higher than 55% "in the wild." Thus, what additional features do human comprehenders make use of? Here, it is instructive to contrast with the rates of request interpretations across grammatical form: 69% of modal interrogatives were interpreted as requests, compared to only 39% of declarative sentences. While it is challenging to disentangle the effect of grammatical form from other semantic or pragmatic features affecting the request prior, this does suggest that participants' decisions in the behavioral experiment were driven more by grammatical or semantic properties of the sentences themselves, rather than the prosodic features associated with the spoken utterance. Of course, comprehenders likely make use of other cues as well, such as gesture (Kelly et al., 1999), situational context (Deliens et al., 2018), and

even the speaker's likely knowledge state (Deliens et al., 2017; Trott & Bergen, 2019, 2020). How do comprehenders integrate these disparate sources of meaning in a rich, multimodal context, particularly if these cues come into conflict? Recent work (Deliens et al., 2017, 2018) suggests that at least in the case of irony processing, comprehenders exhibit a speed/accuracy trade-off in the integration of prosodic vs. contextual cues, respectively: prosody offers a rapid but less reliable cue to meaning, whereas integrating context might be more effortful but ultimately more accurate. Do comprehenders exhibit similar trade-offs when processing indirect requests as well?

The issue of context also leads into the final question: do speakers generate discriminable prosodic cues strategically to overcome ambiguity in a context-sensitive manner, or are these cues present regardless of the degree to which a given utterance might be perceived as ambiguous? In the case of syntactic ambiguity, some (Allbritton et al., 1996; Snedeker & Trueswell, 2003) have found that discriminable prosodic cues disappear when an utterance is produced in a sufficiently disambiguating context, while others (Schafer et al., 2000, 2005) have argued that the cues are produced regardless of how much information is provided by the context, provided the task is made sufficiently interactive. This connects to larger questions about the extent to which speakers engage in audience design to reduce the burden of processing for comprehenders (Ferreira, 2008, 2019). Importantly, this audience design might itself manifest in multiple ways. Speakers might produce more prominent or discriminable prosodic cues for sentences that are especially ambiguous (e.g., those without particularly strong priors, such as "Can you play the piano?"), as might be predicted by a noisy channel approach (Bergen & Goodman, 2015; Gibson et al., 2013). Alternatively, speakers might deploy prosody selectively for sentences with especially salient interpretations (e.g., "Can you open that window?"), marking a deviation from the expected interpretation. Answering these questions will illuminate how speakers and comprehenders alike recruit a diverse set of cues to coordinate on a shared understanding in a dynamic, noisy environment.

Acknowledgements

We thank Rachel Ostrand for her advice on the modeling of acoustic features, Amy Schafer for her helpful comments on an early draft, and Chigusa Kurumada for valuable pointers to relevant research on intonation and meaning (as well as ideas for future analyses). We are also grateful to both the speakers and the participants. Finally, we thank the anonymous reviewers for their helpful comments and suggestions.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Sean Trott https://orcid.org/0000-0002-6003-3731

Supplemental material

Supplemental material for this article is available online.

Notes

1. As Flöck (2016) notes, indirect forms are slightly more frequent (53.5%) in British English than American English (p. 118); see also Aijmer (2014, Chapter 4), for estimates of the frequency of indirect forms in British English.
2. Note that the original sample was 24 speakers; 6 were excluded according to pre-registered exclusion criteria (see "Participants" section for more details).

3. An alternative operationalization of duration would simply be the total number of frames (voiced and unvoiced) in the recorded utterance, that is, including pauses and voiceless obstruents. We hypothesized that speakers would mark the intent of their utterance by drawing out specific lexical items (e.g., the "can" in "Can you lift that box?"), which is more directly measured by calculating the number of voiced frames. Note that number of voiced frames is correlated with the total number of frames overall ($r = 0.63$); future work could investigate the extent to which these measures contain different information about intent.
4. Note that one limitation of this approach is that it involves more model comparisons, and thus a stricter penalty for multiple comparison corrections. Thus, although it allowed us to explore a larger number of acoustic features, the stricter penalty increases the probability of failing to detect certain marginal effects.

References

Aijmer, K. (2014). *Conversational routines in English: Convention and creativity*. Routledge.

Allbritton, D. W., McKoon, G., & Ratcliff, R. (1996). Reliability of prosodic cues for resolving syntactic ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(3), 714–735.

Attardo, S., Eisterhold, J., Hay, J., & Poggi, I. (2003). Multimodal markers of irony and sarcasm. *Humor*, *16*(2), 243–260.

Banuazizi, A., & Creswell, C. (1999). Is that a real question? final rises, final falls, and discourse function in yes-no question intonation. *CLS*, *35*, 1–14.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. arXiv preprint arXiv:1506.04967.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Beach, C. M. (1991). The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, *30*(6), 644–663.

Beckman, M. E., Hirschberg, J. B., & Shattuck-Hufnagel, S. (2004). The original ToBI system and the evolution of the ToBI framework. In S. Jun (Ed.), *Prosodic models and transcription: Towards prosodic typology* (pp. 9–54).

Bergen, L., & Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science*, *7*(2), 336–350.

Blum-Kulka, S., House, J., & Kasper, G. (1989). *Cross-cultural pragmatics: Requests and apologies* (Vol. 31). Ablex Pub.

Briggs, G., Williams, T., & Scheutz, M. (2017). Enabling robots to understand indirect speech acts in task-based interactions. *Journal of Human-Robot Interaction*, *6*(1), 64–94.

Bryant, G. A., & Fox Tree, J. E. (2002). Recognizing verbal irony in spontaneous speech. *Metaphor & Symbol*, *17*(2), 99–119.

Bryant, G. A., & Fox Tree, J. E. (2005). Is there an ironic tone of voice? *Language and Speech*, *48*(3), 257–277.

Caballero, J. A., Vergis, N., Jiang, X., & Pell, M. D. (2018). The sound of im/politeness. *Speech Communication*, *102*, 39–53.

Cheang, H. S., & Pell, M. D. (2008). The sound of sarcasm. *Speech Communication*, *50*(5), 366–381.

Culpeper, J. (2011). "It's not what you said, it's how you said it!": Prosody and impoliteness. *Discursive Approaches to Politeness*, *8*, 57–83.

Culpeper, J., Bousfield, D., & Wichmann, A. (2003). Impoliteness revisited: With special reference to dynamic and prosodic aspects. *Journal of Pragmatics*, *35*(10–11), 1545–1579.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Deliens, G., Antoniou, K., Clin, E., & Kissine, M. (2017). Perspective-taking and frugal strategies: Evidence from sarcasm detection. *Journal of Pragmatics*, *119*, 33–45.

Deliens, G., Antoniou, K., Clin, E., Ostashchenko, E., & Kissine, M. (2018). Context, facial expression and prosody in irony processing. *Journal of Memory and Language*, *99*, 35–48.

D'Imperio, M., & House, D. (1997). *Perception of questions and statements in Neapolitan Italian* [Conference session]. Fifth European Conference on Speech Communication and Technology, Rhodes, Greece, September 1997.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation*, *49*, 209–246.

Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology*, *70*, 29–51.

Flöck, I. (2016). *Requests in American and British English: A contrastive multi-method analysis* (Vol. 265). John Benjamins.

Geluykens, R. (1988). On the myth of rising intonation in polar questions. *Journal of Pragmatics*, *12*(4), 467–485.

Gibbs, R. W. Jr. (1979). Contextual effects in understanding indirect requests. *Discourse Processes*, *2*(1), 1–10.

Gibbs, R. W. Jr. (1981). Your wish is my command: Convention and context in interpreting indirect requests. *Journal of Verbal Learning and Verbal Behavior*, *20*(4), 431–444.

Gibbs, R. W. Jr. (1986). What makes some indirect speech acts conventional? *Journal of Memory and Language*, *25*(2), 181–196.

Gibbs, R. W. Jr. (1987). Mutual knowledge and the psychology of conversational inference. *Journal of Pragmatics*, *11*(5), 561–588.

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, *110*(20), 8051–8056.

Goldschmidt, M. M. (1998). Do me a favor: A descriptive analysis of favor asking sequences in American English. *Journal of Pragmatics*, *29*(2), 129–153.

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

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, *88*, 70–86.

Hirschberg, J. (2017). Pragmatics and prosody. In Y. Huang (Ed.), *The Oxford handbook of pragmatics* (pp. 532–549). Oxford University Press.

Hirst, W., LeDoux, J., & Stein, S. (1984). Constraints on the processing of indirect speech acts: Evidence from aphasiology. *Brain and Language*, *23*(1), 26–33.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70.

Holtgraves, T. (1994). Communication in context: Effects of speaker status on the comprehension of indirect requests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1205–1218.

Holtgraves, T. (1997). Styles of language use: Individual and cultural variability in conversational indirectness. *Journal of Personality and Social Psychology*, *73*(3), 624–637.

Holtgraves, T., & Joong-Nam, Y. (1990). Politeness as universal: Cross-cultural perceptions of request strategies and inferences based on their use. *Journal of Personality and Social Psychology*, *59*(4), 719–729.

Holtgraves, T., & Yang, J. N. (1992). Interpersonal underpinnings of request strategies: General principles and differences due to culture and gender. *Journal of Personality and Social Psychology*, *62*(2), 246–256.

House, D. (2003). Hesitation and interrogative Swedish intonation. *Phonum (Reports in Phonetics, University of Umeå)*, *9*, 185–188.

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15.

Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, *88*, 106–126.

Kelly, S. D. (2001). Broadening the units of analysis in communication: Speech and nonverbal behaviours in pragmatic comprehension. *Journal of Child Language*, *28*(2), 325–349.

Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, *40*(4), 577–592.

Ladd, D. R., Silverman, K. E., Tolkmitt, F., Bergmann, G., & Scherer, K. R. (1985). Evidence for the independent function of intonation contour type, voice quality, and F0 range in signaling speaker affect. *The Journal of the Acoustical Society of America*, *78*(2), 435–444.

Le Pair, R. (1996). Spanish request strategies: A cross-cultural analysis from an intercultural perspective. *Language Sciences*, *18*(3–4), 651–670.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107–120.

Pell, M. D., Vergis, N., Caballero, J., Mauchand, M., & Jiang, X. (2018). Prosody as a window into speaker attitudes and interpersonal stance. *The Journal of the Acoustical Society of America*, *144*(3), 1840–1840.

Perrault, C. R., & Allen, J. F. (1980). A plan-based analysis of indirect speech acts. *Computational Linguistics*, *6*(3–4), 167–182.

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

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, *90*(6), 2956–2970.

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Roche, J. M., Morgan, S. D., Fissel Brannick, S., & Bryndel, K. (2019). Acoustic correlates of female confidence: A production and comprehension study. *The Journal of the Acoustical Society of America*, *145*(6), 3410–3426.

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

Schafer, A. J., Speer, S. R., & Warren, P. (2005). Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task. *Approaches to Studying World-Situated Language Use*, 209–225.

Searle, J. R. (1975). Indirect speech acts. In *Speech acts* (pp. 59–82). Brill.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., & Van Ess-Dykema, C. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech*, *341*(4), 443–492.

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

Speer, S. R., Warren, P., & Schafer, A. J. (2011). Situationally independent prosodic phrasing. *Laboratory Phonology*, *2*(1), 35–98.

Sridhar, V. K. R., Bangalore, S., & Narayanan, S. (2009). Combining lexical, syntactic and prosodic cues for improved online dialog act tagging. *Computer Speech & Language*, *23*(4), 407–422.

Tromp, J., Hagoort, P., & Meyer, A. S. (2016). Pupillometry reveals increased pupil size during indirect request comprehension. *Quarterly Journal of Experimental Psychology*, *69*(6), 1093–1108.

Trott, S., & Bergen, B. (2019). Individual differences in mentalizing capacity predict indirect request comprehension. *Discourse Processes*, *56*(8), 675–707.

Trott, S., & Bergen, B. (2020). When do comprehenders mentalize for pragmatic inference? *Discourse Processes*, *57*(10), 900–920.

Trott, S., Reed, S., Ferreira, V., & Bergen, B. (2019). Prosodic cues signal the intent of potential indirect requests. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society, Montreal, Canada* (pp. 1142–1148).

Vaissière, J. (1983). Language-independent prosodic features. In G. Brown (Ed.), *Prosody: Models and measurements* (pp. 53–66). Springer.

Ward, G., & Hirschberg, J. (1985). Implicating uncertainty: The pragmatics of fall-rise intonation. *Language*, *61*, 747–776.

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

Ward, N. G., Carlson, J. C., & Fuentes, O. (2018). Inferring stance in news broadcasts from prosodic-feature configurations. *Computer Speech & Language*, *50*, 85–104.

Ward, N. G., Carlson, J. C., Fuentes, O., Castan, D., Shriberg, E., & Tsiartas, A. (2017, August). Inferring Stance from Prosody. In *INTERSPEECH* (pp. 1447–1451). Stockholm, Sweden.

Wichmann, A. (2000, September). The attitudinal effects of prosody, and how they relate to emotion. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Island.

Wichmann, A. (2002, April). *Attitudinal intonation and the inferential process* [Conference session]. Speech Prosody 2002, International Conference, Aix-en-Provence, France.

Wichmann, A. (2004). The intonation of please-requests: A corpus-based study. *Journal of Pragmatics*, *36*(9), 1521–1549.

Williams, T., Thames, D., Novakoff, J., & Scheutz, M. (2018, March). Thank you for sharing that interesting fact!: Effects of capability and context on indirect speech act use in task-based human-robot dialogue. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, (pp. 298–306). ACM, Chicago, IL.

Yaeger-Dror, M. (2002). Register and prosodic variation, a cross language comparison. *Journal of Pragmatics*, *34*(10–11), 1495–1536.