Prosodic cues signal the intent of potential indirect requests

Sean Trott (sttrott@ucsd.edu)
Stefanie Reed (sar046@ucsd.edu)
Department of Cognitive Science, 9500 Gilman Dr.
La Jolla, CA

Victor Ferreira (vferreira@ucsd.edu)
Department of Psychology, 9500 Gilman Dr.
La Jolla, CA

Benjamin Bergen (bkbergen@ucsd.edu)
Department of Cognitive Science, 9500 Gilman Dr.
La Jolla, CA

Abstract

Ambiguity pervades language. One prevalent kind of ambiguity is *indirect requests*. For example, "My office is really hot" could be intended not only as a complaint about the temperature, but as a request to turn on the AC. How do comprehenders determine whether a speaker is making a request? We ask whether the *prosody* of an utterance provides information about a speaker's intentions. In a behavioral experiment, we find that human listeners can identify which of two utterances a speaker intended as a request, suggesting that speakers *can* produce discriminable cues. We then show that the acoustic features associated with an utterance allow a classifier to detect the original intent of an utterance (74% accuracy). Finally, we ask which of these features predict listener accuracy on the behavioral experiment.

Keywords: indirect requests; prosody; language production; language comprehension; inference

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 open the window. Some indirect requests use a highly conventionalized form (in this example, "**Can you** X?"). But other indirect requests 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, 1990), cognitive psychology (Holtgraves, 1994), and natural language processing (Perrault & Allen, 1980; Williams et al, 2018) for several reasons. First, they're exceedingly frequent. One study eliciting requests from participants found that over 80% were indirect in some way (Gibbs, 1981). Second, successfully comprehending indirect requests requires the hearer to make inferences about the speaker's intent, using linguistic and other contextual knowledge, potentially involving diverse cognitive systems, which can pose challenges to computational implementations of language comprehension (Briggs, Williams, & Scheutz, 2017). But it still remains to be determined what information human comprehenders use to recover the intended interpretation of a potential indirect request.

Previous work suggests that successfully understanding indirect requests requires the integration of extra-linguistic contextual information. For *conventional* indirect requests, comprehenders can use the form of the utterance as a partial cue to its meaning. Consequently, conventional indirect requests are thought to be easier to understand (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 even lead listeners to *misidentify* intended questions as requests (e.g. "Can you play tennis?"), as has been reported for individuals with anterior aphasia and right-hemisphere brain damage (Hirst, LeDoux, & Stein, 1984).

Less conventional indirect requests, such as "My office is really hot", require the hearer to infer both the speech act (e.g. is it a request?) as well as the intended substance of the request, and are thus thought to incur higher processing costs than their literal, non-request counterparts (Tromp, Hagoort, and Meyer, 2016), as well as more conventional indirect requests (Gibbs, 1981). Successful disambiguation of these utterances may benefit from co-speech gesture and eye gaze (Kelly et al, 1999), as well as a representation of what is mutually known across interlocutors (Gibbs, 1987; Trott & Bergen, 2018).

Finally, indirect requests have proven challenging for machine language understanding. Wizard-of-Oz style experiments show that human speakers continue to use indirect requests when speaking to robots (Briggs, Williams, & Scheutz, 2017), even when those robots demonstrably cannot understand them (Williams et al, 2018). Current state-of-the-art solutions (Briggs, Williams, & Scheutz, 2017) use rules relating utterance forms to contexts to probabilistically derive the intended interpretation of ambiguous utterances like "Can you knock down the red tower?" While these solutions work well for established utterance-context mappings, they could still benefit from an increased understanding of precisely which disambiguating

information is available (e.g. paralinguistic or extralinguistic cues), and which is actively exploited by human comprehenders.

Specifically unexplored to date as a candidate source of disambiguating information, is **prosody**: the intonational, rhythmic, and tonal properties of how an utterance is spoken or signed.

Prosodic Cues for Disambiguation

Previous work on other kinds of linguistic ambiguity has already demonstrated that prosodic cues can provide disambiguating information about a speaker's intent.

For one, prosodic features such as *pitch* and *pause duration* can act as "parsing instructions" for listeners. Using speech synthesis, Beach (1991) modified the pitch and duration of critical regions of sentences involving temporary ambiguity (e.g. whether a noun phrase was functioning as a sentential complement or direct object), and found that participants were able to identify the intended parse without listening to the entire utterance. Similarly, Price et al (1991) found that FM radio newscasters, naïve to the purposes of the experiment, produced marked prosodic cues that aided listeners' comprehension of parenthetical statements, apposition, and prepositional phrase attachment ambiguities. 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 substantial debates about the conditions under which *speakers* reliably produce such cues—some studies (Allbritton et al, 1996; Snedeker & Trueswell, 2003) have found that discriminating prosodic cues disappear in the presence of sufficiently disambiguating contextual information, while others (Schafer et al, 2000; Speer et al, 2011) have found that they persist, and have argued that the failure to find such cues is due to limitations on the elicitation paradigms used (e.g. being non-interactive or having low stakes). Regardless, the evidence shows that when such cues *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 (Shriberg et al, 1998) found that including prosodic features from conversational speech (including duration, pause, F0, energy, and speech rate) improved 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. More recently, Hellbernd & Sammler (2016) asked whether trained human speakers could produce cues that identified the intended speech act of one-word utterances—e.g. producing the word "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 were able to 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), suggesting that success was at least partially dependent on information contained in the speech signal (though see Bryant & Fox Tree (2005) for further discussion of whether these prosodic features are *global* or *local*, and whether they are uniquely characteristic of irony in particular). 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 vs. 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 (Pell et al, 2018; Ward et al, 2017; Ward et al, 2018). 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), 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 (Wichmann, 2000; Culpeper, Bousfield, & Wichmann, 2003).

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 reliable enough to be detectable—and useful—to both human and machine comprehenders.

However, the role of prosodic features in signaling the intended interpretation of potential indirect requests is currently unexplored. Do speakers and hearers use prosody to overcome the pragmatic ambiguity intrinsic to the most common way to make requests? We addressed this in the current work through three core questions. First, *can* speakers produce reliable cues to indicate to human listeners whether or not they are making a request? Second, *which* cues do speakers actually produce? And third, are these the same cues that listeners seem to use?

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/prosody_indirect_requests.

Experiment: Listener Judgments of Intent

In a behavioral experiment, we asked whether speakers can produce reliably discriminable prosodic cues. Specifically,

we asked whether these prosodic cues reliably aid human *listeners* in discriminating the speaker's pragmatic intent. On each trial, participants were given two recordings of the same utterance by the same speaker (e.g. "Can you open that window?", or "My soup is cold"), and were asked to select which of the two utterances was intended as a request. If speakers can produce detectable, reliable cues, then participants should be able to identify which utterance was produced as a request; but if speakers cannot produce such cues, or if the cues they produce are not usable by human listeners, then participants should perform at chance.

Methods

Participants 78 participants, all native English speakers, were recruited from Amazon Mechanical Turk. We aimed to recruit 80 participants, but Mechanical Turk under-sampled to 78 participants. The mean age of our participants was 37 (SD=11), ranging from 20 to 69. 30 identified as female, 45 as male, 2 as non-binary, and 1 declined to answer. Each participant was paid $2 for participating, and the experiment took on average 24 minutes to complete.

Materials We recorded five English speakers (2 male, 3 female). Speakers were given 12 utterances to produce (6 conventional indirect requests of the form "Can you X?", and 6 non-conventional indirect requests of the form "My X is Y"), and were instructed to say each utterance twice—once as a request, and once as a literal question or statement. They were allowed to read over the utterances before speaking. The experiment was implemented using JsPsych (de Leeuw, 2015).

Procedure After completing an audio check, participants were instructed that they would listen to a series of paired utterances. They were told that one member of each pair was always intended as a request, and the other member was not. Their task was to indicate which was the request by selecting one of two buttons (either "First" or "Second", corresponding to the first or second utterance presented).

On each trial, participants heard two utterances, containing the same words and produced by the same speaker, with 1 second of silence following each utterance. The order of the utterances (e.g. whether the request or non-request version came first) was counterbalanced within-speaker using a weighted randomization scheme (e.g. for each *speaker-block*, 6 trials contained the request version first, and 6 contained the non-request version first). After listening to both versions, participants indicated which one they thought was intended as a request via button-press.

Each participant performed 60 trials (12 utterance pairs for each of the 5 speakers), blocked by *speaker*. The order of the trials within each *speaker-block* was randomized, as was the order of *speaker-blocks*.

Results

All statistical analyses were performed in R (R Core Team, 2017), using the *lme4* package (Bates et al, 2015). Random effects structure was determined by beginning with the maximal model, then reducing as needed for model convergence (Barr et al, 2013).

Our first question was whether participants could successfully determine which utterance was intended as the request. To test this, we built a generalized linear mixed effects model, with *response* (First or Second) as the dependent variable, and *correct answer* (First or Second) as a fixed effect, as well as random slopes for the effect of *correct answer* for both subjects and items (as well as random intercepts for both). We compared this full model to a reduced model omitting the fixed effect of *correct answer*, and found that the full model explained significantly more variance [$X^2(1)=24.97$, $p=5.8*10^{-7}$]. In other words, participants were able to discriminate request and non-request utterances at a rate above chance.

We were also interested in which characteristics predicted accuracy on particular items—were participants better at identifying pragmatic intent for certain *forms* (conventional vs. non-conventional), or for certain *speakers*? We used nested model comparisons, with *correct* (Yes or No) as a dependent variable, by-item random slopes for *speaker*, by-subject random slopes for *form*, and random intercepts for both items and subjects, to determine whether *form*, *speaker*, and their interaction explained independent sources of variance in participant accuracy. A model with fixed effects for both *form* and *speaker* explained more variance than a model with *form* alone [$X^2(4)=11.5$, $p=.02$], as well as a model with *speaker* alone [$X^2(1)=5.2$, $p=.02$]. Adding an interaction between *form* and *speaker* explained additional variance [$X^2(4)=14.1$, $p=.007$]. In other words, certain speakers produced more discriminable signals overall, and *conventional* requests were generally easier to identify than non-conventional requests, except in the case of one speaker, "S2" (see *Figure 1*).

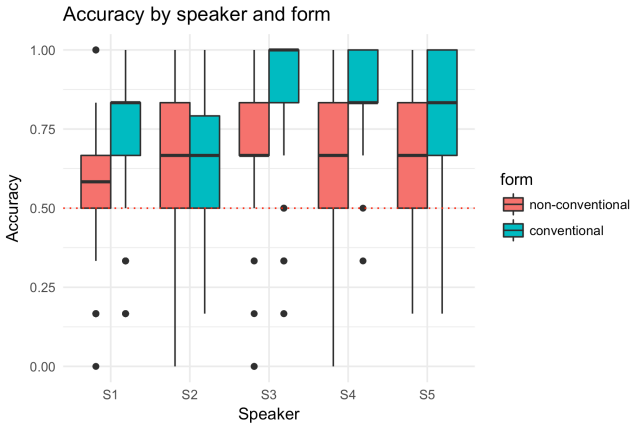

Figure 1: Human accuracy was above chance for all speakers and forms. Accuracy was higher for some speakers (e.g. S3, S4) and some forms (e.g. conventional requests). Dotted red line signifies chance (50%).

One possibility is that participants improved in accuracy over the course of the experiment, perhaps learning which prosodic features signaled intent. We compared a model

with *correct* (Yes/No) as a dependent variable, *Order* (1-60) as a fixed effect, and random intercepts for subjects and items, to a model omitting the fixed effect of *Order*, and found that the full model did not explain significantly more variance [$X^2(1)=.7$, $p=.4$]. Thus, there is no evidence that participants improved over the course of experiment. However, it is also possible that participants improved within each *speaker-block*, but that this adaptation did not carry over across blocks. To test this, we replaced *Order* with *Order-within-block* (1-12) as a fixed effect; a model including *Order-within-block* explained marginally more variance than a model omitting this term [$X^2(1)=3.1$, $p=.08$]. This explanatory power was independent from the variability explained by *speaker*, as determined by comparison of a model including fixed effects of both *speaker* and *Order-within-block* to a model with only *speaker* [$X^2(1)=3.3$, $p=.07$]. Adding an interaction between these factors did not increase explanatory power [$X^2(4)=3.3$, $p=.5$]. This provides weak evidence for within-block adaptation or learning, but requires further analysis and experimentation.

Analysis of Acoustic Features

Listener judgments of pragmatic intent in the behavioral experiment described above demonstrated that speakers produced signals that increased communicative success. However, this analysis does not indicate *which* acoustic features predict a speaker's intended pragmatic interpretation. Here, we asked whether seven acoustic features reliably predicted a speaker's *intent*. Predictive power was assessed in two ways. First, we asked about the explanatory power of each variable in turn using nested model comparisons. Second, we used leave-one-out cross-validation to determine how the combination of *all* features improved the ability of a classifier to identify *intent*.

Data Processing

For each of the 120 recordings (5 speakers producing 12 utterances with two versions each), we used Parselmouth (Jadoul et al, 2018), a Python interface to Praat, to extract the following acoustic features: mean F0, range F0 (max F0 – min F0), standard deviation of F0, duration (number of voiced frames), mean intensity, standard deviation of intensity, and slope of F0 (slope of regressing *F0 ~ time*). We then *z-scored* each of these variables with respect to each speaker's mean and SD, to account for considerable variability in speakers overall.

Results

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, fixed effects for each of the seven acoustic features, and random intercepts for each utterance. We adjusted for multiple comparisons using Holm-Bonferroni corrections (Holm, 1979). In each case, a positive coefficient represents a higher likelihood of a *Non-Request*, while a negative coefficient represents a higher likelihood of a *Request*.

For a logistic regression model predicting *intent* of all items (e.g. both *conventional* and *non-conventional* utterances), model fit was improved by including *mean intensity* [$X^2(1)=8.7$, $p=.003$, $p_{adj}=.02$] and *SD intensity* [$X^2(1)=7.8$, $p=.005$, $p_{adj}=.03$]. Higher-intensity utterances were more likely to be Requests [$\beta=-.69$, $SE=.25$, $p=.006$], as were utterances were with greater variation in intensity [$\beta=-1.1$, $SE=.4$, $p=.01$]. No other acoustic features significantly improved model fit after correcting for multiple comparisons.

Because human listener accuracy differed significantly as a function of *form* (see the behavioral experiment), it is possible that distinct prosodic features predict intent for *conventional* and *non-conventional* requests. Thus, we ran the same analysis as above twice: once on only *conventional* and once on only *non-conventional* requests.

For a model predicting *intent* of only *conventional* requests, model fit was improved by including *F0 slope* [$X^2(1)=7.8$, $p=.005$, $p_{adj}=.03$], *SD intensity* [$X^2(1)=7.7$, $p=.005$, $p_{adj}=.03$], and *F0 duration* [$X^2(1)=8.8$, $p=.003$, $p_{adj}=.02$]. More positive slopes were associated with Non-Requests, e.g. literal questions [$\beta=1.1$, $SE=.5$, $p=.01$], as were longer utterances [$\beta=1.1$, $SE=.4$, $p=.01$] and less variation in intensity [$\beta=-1.1$, $SE=.4$, $p=.01$].

For a model predicting *intent* of only *non-conventional* requests, model fit was significantly improved by including *F0 duration* [$X^2(1)=19.6$, $p=9.7*10^{-6}$, $p_{adj}=.00004$], with longer utterances having a higher probability of being Requests [$\beta=-2.5$, $SE=.94$, $p=.008$].

In sum, we identified several acoustic features that predict pragmatic intent. Overall, intent was predicted by *mean intensity* and *SD intensity*. For *conventional* requests in particular, intent was predicted by *F0 slope*, *F0 duration*, and *SD intensity*; for *non-conventional* requests, intent was predicted by *F0 duration*. These results suggest that those features *could*, in principle, be used to identify the intent of an ambiguous utterance.

To determine whether the combination of all seven acoustic features could improve a classifier's ability to detect *intent*, we used leave-one-out cross-validation (LOOCV). A model including all seven acoustic features (as well as their interactions with *form*) accurately predicted *intent* on 74% of the held-out items, a rate substantially above chance (50%).

Predicting Accuracy from Acoustic Features

By regressing pragmatic intent against extracted acoustic features, we isolated multiple features that appear to indicate intent of either conventionally or non-conventionally formatted utterances: F0 slope, F0 duration, mean intensity, and SD intensity. However, this does not entail that listeners actively exploit differences in these features to infer intent. It could be that these features are *statistically* reliable, but

not *psychologically* valid. Which, if any, of these features actually benefit listeners?

One way to test this is to ask: do by-item *differences* in any of the acoustic features explain independent sources of variance in listener *accuracy*, above and beyond the full model specified above in the behavioral experiment (containing an interaction between *form* and *speaker*)? If larger differences from a given dimension (e.g. *F0 slope*) consistently predict accuracy, this suggests that listeners are actively benefitting from those differences, and are thus consistently sampling and deploying information about that particular dimension.

Data Processing

For each utterance pair, we computed the *difference* of each z-scored feature between the Request version and the Non-request version. Thus, a positive value for *F0 slope difference* indicates that the Request version had a larger slope than the Non-request version, while a negative value indicates that the Non-request version had a larger slope. We repeated this procedure for each acoustic feature.

Results

We asked about the informativeness of each acoustic feature (as well as its interaction with *form*) using nested model comparisons. The explanatory power of a given variable was determined by comparing a model including that term to a model without it. We adjusted for multiple comparisons using Holm-Bonferroni corrections (Holm, 1979).

The full model included the terms from the maximal model specified in the behavioral experiment, with *correct* (Yes/No) as a dependent variable, an interaction between *form* and *speaker*, fixed effects for both *form* and *speaker*, and random intercepts for subjects and items. It also included each of the seven acoustic features, as well as their interaction with *form*.

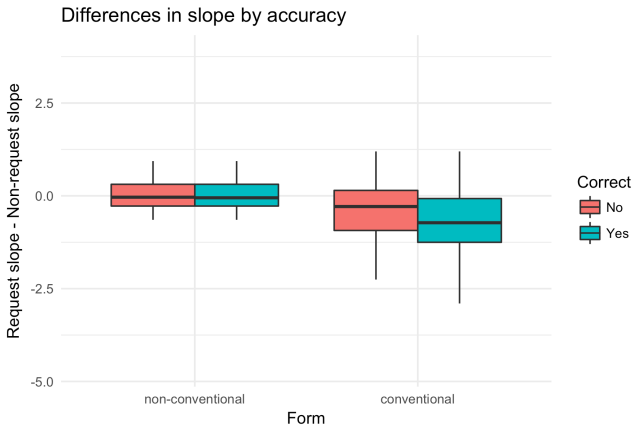

Figure 2: Differences in z-scored F0 slope by *form* and *accuracy*. Conventional items with a larger difference between the Request and Non-Request version (specifically, where the slope on the *Non-Request* version was more positive than the slope on the *Request* version) were more likely to be answered correctly.

Model fit was significantly improved by the interaction between *F0 slope difference* and *form* [$X^2(1)=16.98$, $p=3.78*10^{-4}$, $p_{adj}=.0005$], but was not significantly improved by *F0 slope difference* alone ($p_{adj}>.1$). The direction of this interaction is illustrated in *Figure 2*: accuracy on non-conventional items was not significantly impacted by the difference in *F0 slope* between the Request and Non-Request differences, whereas a larger difference for conventional items predicted more accurate responses. Specifically, conventional items on which the Non-Request version had a more positive slope than the Request version (and thus their difference was more *negative*) were more likely to be answered correctly [$\beta=-.4$, $SE=.1$, $p=5.5*10^{-5}$].

Model fit was also improved by the interaction between *mean F0* and *form* [$X^2(1)=10.6$, $p=.001$, $p_{adj}=.01$], as well as the main effect of *mean F0* [$X^2(1)=15.03$, $p=.0001$, $p_{adj}=.001$]. Specifically, conventional items on which the Request version had a lower *mean F0* than the Non-Request version were more likely to be answered correctly [$\beta=-.47$, $SE=.14$, $p=.001$]. Because these comparisons included a term for *F0 slope*, this does not appear to be due simply to conventional Non-Request items exhibiting a sharper final rise (e.g. more positive slope). Differences in *mean F0* explained independent sources of variance from *F0 slope*.

A model including an interaction between *mean intensity* and *form* did not explain more variance than a model omitting that term, but the fixed effect of *mean intensity* did improve model fit [$X^2(1)=9.7$, $p=.002$, $p_{adj}=.02$]. Specifically, items on which the Request version had a higher overall *mean intensity* than the Non-Request version were marginally more likely to be answered correctly [$\beta=.1$, $SE=.05$, $p=.06$].

Model fit was also improved by the interaction between *SD intensity* and *form* [$X^2(1)=7.12$, $p=.008$, $p_{adj}=02$], though not the fixed effect of *SD intensity* alone ($p_{adj}>.1$). Conventional items on which the Request version exhibited greater variation in intensity than the Non-Request version were more likely to be answered correctly [$\beta=.29$, $SE= .11$, $p=.007$].

In summary, four of the acoustic features we extracted predicted listener accuracy—*F0 slope*, *mean F0*, *mean intensity*, and *SD intensity*. *F0 slope* appeared to be useful primarily for conventional requests (with more positive slopes indicating the literal, Non-Request interpretation). *Mean F0* was helpful for both, though again, appeared to be particularly predictive of accuracy on the conventional items (with higher mean F0 on the Non-Request versions predicting higher accuracy). *Mean intensity* was predictive of accuracy on both kinds of items; items on which the Request version exhibited higher overall intensity than the Non-Request version were more likely to be answered correctly. Finally, *SD intensity* was particularly helpful for conventional items—Request versions with more variability in intensity than their Non-Request counterpart were more likely to be correctly identified.

General Discussion

Human listeners were able to discriminate the pragmatic intent of potential indirect requests, indicating that speakers *can* produce discriminable cues, at least when made aware of an utterance's different interpretations. We extracted seven acoustic features from each recorded utterance, and found that four of these features were predictive of listener accuracy in the behavioral experiment: F0 slope, mean F0, mean intensity, and SD intensity. Specifically, larger differences in each of these features were associated with more accurate responses; some were primarily helpful for conventional items (F0 slope, SD intensity, mean F0), while others were helpful for both (mean intensity).

Additionally, using leave-one-out cross-validation, a machine learning classifier trained on these features (and their interaction with utterance *form*) successfully identified the *intent* of potential request utterances 74% of the time (where chance is 50%). Thus, prosodic features are not only useful to human comprehenders attempting to discriminate a speaker's pragmatic intent—they are also informative to machines, suggesting that they could perhaps be integrated into existing natural language understanding architectures (Briggs et al, 2017).

Open questions remain. First, we noted a weak effect of *Order-within-block*, but not *Order* overall, on accuracy. That is, there is no evidence that listeners improved over the course of the entire experiment, but they might have improved while listening to each speaker. If true, this provides weak evidence for *adaptation* to each speaker, which may not successfully carry over across speakers. The effect was marginally significant. Since it arose during exploratory data analysis, it requires further investigation.

Second, a limitation of the behavioral experiment is that participants were asked to explicitly discriminate between two versions of the same utterance (e.g. "which was the request?"), rather than *classifying* an individual utterance (e.g. "is that a request?"). The latter design is clearly more applicable to real-world scenarios, in which comprehenders do not have immediate access to alternative versions of an utterance. We are designing a new set of studies to ask whether comprehenders can identify whether a given utterance was intended as a request, and whether the same acoustic features—e.g. F0 slope, mean intensity, etc.—predict their response. This task design will also allow more direct comparison to the classifier's results, so that we can determine whether the classifier is using similar features (and making similar errors) as human comprehenders.

Third, a long-standing question in the literature on prosody and pragmatic intent is whether particular prosodic features convey direct information about the intended speech act, 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, Eisterhold, Hay, & Poggi, 2003; Bryant & Fox Tree, 2005). Our experiment was not designed to adjudicate between these two possibilities, but our results do suggest that the answer is nuanced, and likely falls somewhere in between. Certain features, such as *F0 slope*, were predictive only of accuracy for conventional forms (E.g. "Can you open that window?"), and thus might be more aptly described as "marking" a deviation from the default interpretation of modal interrogatives as requests (Gibbs, 1986). But other features, such as *mean intensity*, predicted accuracy across forms; in both cases, items with higher intensity on the Request version (vs. the Non-Request version) were more likely to be answered correctly.

Finally, perhaps the most obvious question is whether, or under what conditions, these kinds of prosodic cues would actually be produced. Speakers in our experiment were made aware of the two interpretations of each utterance, and were explicitly asked to produce utterances consistent with those interpretations. While our results indicate that speakers *can* produce discriminable cues, they do not demonstrate that speakers actually *do*. A similar issue arises in the study of prosodic cues for syntactic disambiguation—some (Allbritton, 1996; Snedeker & Trueswell, 2003) have found that these cues are no longer present when the utterance is produced in a disambiguating context, while others (Schafer et al, 2000; Schafer et al, 2005) have argued that the cues are produced regardless of how much information is provided by the context. Thus, the question becomes: are the discriminable prosodic features we observed automatically and conventionally associated with pragmatic intent, or are they deployed strategically for a particular audience in a particular context?

Acknowledgments

We thank Rachel Ostrand for her helpful advice on the modeling of acoustic features. We are also grateful to both the speakers and the participants, and to the reviewers for their suggestions.

References

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.

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

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*.

Beach, C. M. (1991). The Interpretation of Prosodic Patterns at Points of Syntactic Structure Ambiguity: Evidence for Cue Trading Relations. *J. of memory and language*, *30*(6), 644.

Briggs, G., Williams, T., & Scheutz, M. (2017). Enabling robots to understand indirect speech acts in task-based interactions. *J. 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.

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. doi:10.3758/s13428-014-0458-y

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.

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

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

Gibbs, R. W. (1986). What makes some indirect speech acts conventional?. *J. of memory and language*, *25*(2), 181.

Gibbs, R. W. (1987). Mutual knowledge and the psychology of conversational inference. *J. of pragmatics*, *11*(5), 561-588.

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

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*, 65-70.

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

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.

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

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.

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, Vienna, Austria. URL https://www.R-project.org/.

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

Searle, J. R. (1990). Indirect Speech Acts 12. *The philosophy of language*, 161.

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.

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

Trott, S., & Bergen, B. (2018). Individual Differences in Mentalizing Capacity Predict Indirect Request Comprehension. *Discourse Processes*, *00*(00), 1–33.

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

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

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

Williams, T., Thames, D., Novakoff, J., & Scheutz, M. (2018, February). 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).