



Machine Learning Facilitated Investigations of Intonational Meaning: Prosodic Cues to Epistemic Shifts in American English Utterances

Nanette Veilleux¹, Stefanie Shattuck-Hufnagel², Sunwoo Jeong³, Alejna Brugos¹, Byron Ahn⁴

¹Simmons University, ²Massachusetts Institute of Technology

³Seoul National University, ⁴Princeton University

veilleux@simmons.edu, sshuf@mit.edu, sunwooj@snu.ac.kr,
alejna.brugos@simmons.edu, bta@princeton.edu

Abstract

This work analyzes experimentally elicited speech to capture the relationship between prosody and semantic/pragmatic meanings. Production prompts were comicstrips where contexts were manipulated along axes prominently discussed in sem/prag literature. Participants were tasked with reading lines as the speaker would, uttering a target phrase communicating a proposition p (e.g., “only marble is available”) to a hearer who had epistemic authority on p . Prompts varied whether the speaker’s initial belief (prior bias) was confirmed (condition A: bias= p) or corrected (condition B: bias= $\neg p$); this meaning difference was reinforced by response particles (A: “okay so” vs. B: “oh really”) preceding the target phrase.

485 productions were annotated with phonologically-informed phonetic labels (PoLaR). To model many-to-many mappings between features (prosodic form) and classification (sem/prag meaning), Random Forests were designed on labels and derived measures (including f_0 ranges, slopes, TCoG) from 286 recordings — classifying meaning with high accuracy (>85%). RFs identified condition-distinguishing prosodic cues in both response particle and target phrases, leading to questions of how/whether functionally-overlapping lexical content might affect prosodic realization. Moreover, RFs identified phrase-final f_0 as important, leading to deeper edge-tone explorations. These highlight how explanatory ML models can help iteratively improve targeted analysis.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction and Background

Intonation systematically communicates a wide range of linguistic meanings (related to, e.g., the common ground, conversational dynamics, and information structure). This paper explores the intonational contours associated with another meaning that has been prominently discussed in the sem/prag literature (e.g., [5], [9]): a speaker’s previous beliefs/biases about the proposition in question, p , influence the utterance-final boundary tone. In evaluating whether certain types of “biased” contexts —namely evidential bias (e.g., prior discourse context supports p) and epistemic bias (e.g., the interlocutor believes p)— map onto intonational contours, we keep in mind that we expect any such mappings between the two to be many-to-many, as morphological form-meaning mappings are commonly many-to-many (in part due to homophony and polysemy).

For instance, a rising tune on a declarative radical p (‘it’s raining’) has sometimes been associated with a speaker’s prior epistemic bias towards ‘not p ’ (‘it isn’t raining’), but other times

with speakers’ prior bias towards ‘ p ’. (see e.g., [6], [12], and references therein). From the other direction, meaning which conveys something along the lines of revision of prior epistemic bias towards ‘not p ’ has sometimes been associated with /L* H* L-L%/ , often called the surprise-redundancy contour [13], but alternatively with a particular kind of rising tune, namely a steeply-rising /L* H-H%/ [4].

Based on this state of affairs, much recent work in the sem/prag literature posits that tunes do not directly encode particular contextual biases, but rather encode a more abstract level of meaning that may nevertheless systematically derive the attested correlations, mediated by pragmatic reasoning and diverse sociopragmatic input contexts (see e.g., [6], [12], among many others). But claims about what exactly this abstract meaning is for a given tune varies depending on the theory, and there is no widespread consensus on which aspects of a tune are relevant (e.g., is it the terminal contour like /H* L-L%/? something larger like a larger portion of the intonational melody? something smaller and gradient like the slopes of pitch accents or edge tones?) Part of the uncertainty stems from the fact that, to date, these theories have concentrated primarily on a few idealized tunes and contextual set-ups, while the associations between the two in actual interactional use involves complex many-to-many mappings.

1.1. The Present Study

To test and adjudicate among these theories, one can analyze large-scale production data that are nevertheless controlled in a theoretically informed way, so that we can identify all and only the correlations between tune and contextual bias that are systematic/significant, and thereby clarify the landscape of intonational variation. Put differently, it would be useful to identify the ways in which people adopt multiple alternative intonational strategies in parallel situations.

Investigating complex empirical data of this type can be done through the use of automatic algorithms to find relationships in speech between prosody-related input parameters and seg/prag conditions elicited under controlled, theoretically informed contexts. While standard ANOVA and regression models have been used historically to validate relationships, other Machine Learning models may be more useful for the kinds of complex exploration and analysis required to deal with multiple, often overlapping, dependent features. Models that can accommodate categorical (e.g., the locations of prominences and phrase boundaries, the identity of particular words) and numeric parameters (TCoG, f_0 range, relative duration) are best suited to initial investigation of such data. In addition, explanatory classification models suggest

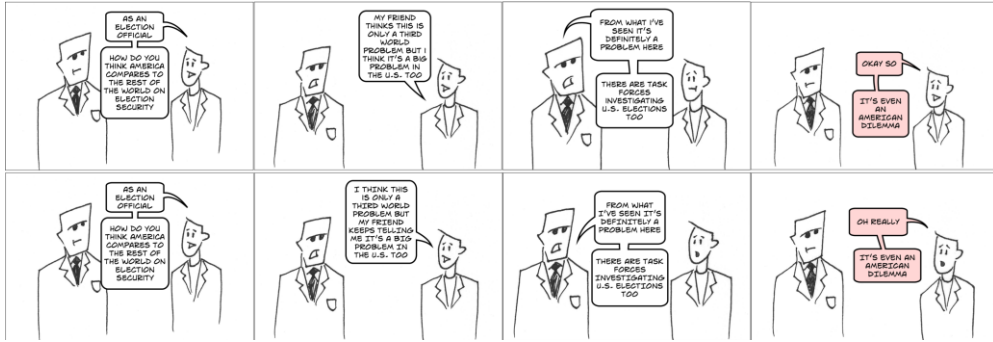


Figure 1: Two conditions for the same target sentence; condition A (above; bias = p) and condition B (above; bias = $\neg p$)

insights that can be iteratively used by linguists to refine parameter extraction.

For these reasons, Random Forests were used to investigate which prosodic attributes were useful in classifying the two pragmatic contexts ($p/\neg p$ bias) investigated in this project. A further useful characteristic of Random Forests is the ability to estimate the importance of each input parameter based on its contribution to decreasing the Gini Index/Impurity. The Gini Index measures the relative heterogeneity (impurity or entropy) of a data set with respect to classification. If a feature (say, phrase final lengthening) is used to partition the data into two subsets that are, collectively, more homogeneous than the entire set (e.g. a greater proportion of +bias in one subset and a greater proportion of -bias in the other), the Gini Impurity decreases. The amount of this decrease can be used as a measure of the importance of the feature. The standard R random forest algorithm [7] attempts all possible feature splits and calculates the benefit of each. As shown below, this featural “importance” measure is a useful byproduct that can be used to rigorously investigate the features and combinations of features.

2. Methods

2.1. Production Task

This work is part of a larger study where conditions differed along two dimensions: speaker’s previous belief (bias about p) and whether the speaker or the hearer is the epistemic authority on p ; see Table 1. The data discussed here involves only the epistemic authority dimension.

Table 1: Experimental design for the elicitation task:
4 conditions \times 24 target sentences = 96 stimuli.
The present work analyzes conditions A and B.

| | Speaker’s previous belief: p | Speaker’s previous belief: $\neg p$ |
|---------------------------------------|--------------------------------|-------------------------------------|
| Epistemic authority Hearer | A | B |
| Epistemic authority Speaker | C | D |

To date, 85 native speakers of American English have participated in an internet-based production task using PCIBex [15], recruited via Prolific, and paid \$15 for their participation. Participants were cis, gender balanced, aged 24-40, with no reported speech or reading impairments, and of a variety of racial backgrounds (69% White; 13% Black; 11% Latinx; 5% mixed, 2% Native American). The prompts for production were

comic strips with dialogs conveying contexts that reflected the experimental manipulations. 24 target items were designed such that the same target sentence (with identical text) was produced in each of the experimental contexts. (Fig.1). Target sentences were preceded by a set of response particles (“okay so” for condition A, “oh really” for condition B), chosen to reinforce the pragmatic context. A Latin square design was used, such that subjects saw each of the 24 target items only once. Items were randomized along with 48 filler comics of comparable format to the target items. Subjects were presented with and asked to read the full comic before reading aloud (and recording) the target sentence and response particles, which were highlighted in red in the comic. Participants were asked to respond to comprehension questions to confirm that they were reading the entire comic, and were attending to the contexts.

2.2. Annotation and data extraction

Each recorded utterance was force-aligned with the Montreal Forced Aligner [8] and was then independently annotated in Praat [2] with advanced PoLaR labels [1], by pairs of trained labellers. PoLaR labels (as in Fig.2) are phonologically-informed descriptive labels on multiple tiers, that encode the location and relative strength of prominences and boundaries (Pr(osodic) Str(ucture) tier), salient turning points in the f_0 contour (Points tier), local ranges (Ranges tier), and scaled pitch levels for each Points label within those ranges (Levels tier). Pairs of labellers then arrived at a single consensus label for each file, via a process of comparison and discussion.

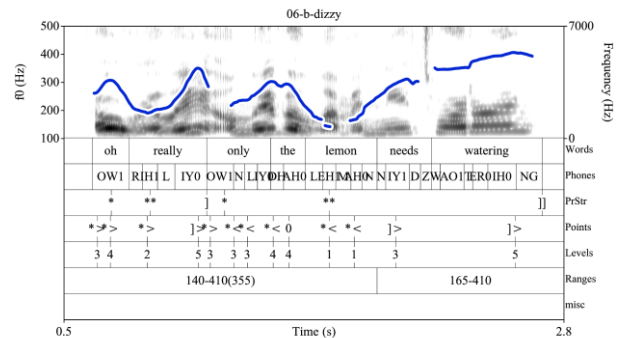


Figure 2: A sample annotated file, with consensus PoLaR labels

Scripts in Praat and R [10] using RStudio [11] were then used to extract measures from these recordings, guided and facilitated by these annotations. The most direct measures were the locations of pitch accents (PoLaR PrStr * labels). Also

rather direct were the locally defined f0 range (floor/ceiling/size; PoLaR Ranges labels) and the phrase-final f0 (PoLaR Points/Levels). (All f0 measures [in semitones] were extracted from the “straight line approximation” [14] of the f0 contour produced by Points labels, and were normed against the local f0 range defined by Ranges labels.)

We also took two measures of the phrase-final f0 movements. The first of these measures was slope and the second was (a more novel usage of) TCoG [2]; the f0 measures were drawn from the straight line approximation for each f0 contour (and normed according to local f0 range), and the time domain was the beginning of the final vowel through the end of the phrase (with values normed according to the midpoint and length of the domain). Phrase-final measures were taken at the final] (boundary label) in the PrStr tier in the response particle domain, and at the final] of the target sentence.

Additionally, TCoG measures were calculated for three pitch accents: the final PrStr * (phrase-level prominence) in the response particle domain, and the penultimate and final *s in the target sentence. For pitch accents, the time domain was defined by the midpoint of the vowel interval and extended one vowel length in each direction (with time values for TCoG calculations normed according to the midpoint and length of the domain). Again, the f0 measures were drawn from the straight line approximation (and normed according to local f0 range) and were weighted such that measures towards the center of the vowel had stronger influence (as a way of controlling for segmental effects).

Finally, we measured phrase-final lengthening, using force aligned phone intervals and PrStr labels. We measured the durations of all vowel-to-vowel (or vowel-to-phrase-edge) intervals. Of these intervals, a mean unlengthened interval was determined by averaging all intervals that were not phrase final and did not contain a pitch accent; final lengthening was then measured as a ratio of the final vowel-to-edge duration to the mean unlengthened interval.

2.3. Machine Learning

As mentioned above, Random Forests were chosen to model the prosodic features that algorithmically model implementation differences between *p*/*¬p* bias in the project data. The data was split into training data (70%, randomly selected) and test data (the remaining 30%) and checked to ensure that the proportion of positive/negative tokens were consistent across the two sets. The R randomForest package [7] was used to train two classification random forest models (see below) using 3 features / split on 500 polled trees. Accuracy was calculated based on its performance on the held out test data.

The models presented here were trained on 23 acoustically-derived attributes that include prominence- and phrasing-related cues: f0 min/max, slopes, TCoG-time, TCoG-freq on prominences and at phrase boundaries in both the response particle and target sentence. Notably, no lexical information was used in these models. (Numeric parameters were normalized, as described above, to minimize algorithmic abnormalities.) In addition, a uniformly-distributed, randomly generated parameter was included. As described in the results, this parameter serves to determine which features decrease Gini Impurity more than a random variable might.

To date, 485 condition A (*p* bias; n=239) and B (*¬p* bias; n=246) utterances have been fully PoLaR annotated. Recordings in which the target word lacked a pitch accent or was not the final pitch accent (n=199) were excluded. Of the

remaining 286 utterances, 107 were in Condition A and 179 were in Condition B.

During data exploration, several other characteristics were discovered that required adjustments. Phrases with missing f0 values were extracted as having flat slope (0) and these cases were adjusted to be NA (missing values), a more felicitous description. Other missing values were handled in one of several ways: (a) the feature itself was removed if it had a significant number of missing values (e.g. penultimate prominences were not consistently present, so related parameters were removed), (b) the utterance was removed if there were missing values or (c) the R random forest na.roughfix process was used that replaces numeric missing values with the overall mean and categorical missing values by the majority value.

3. Results

Two resulting random forest models serve to demonstrate the usefulness of machine-learning aided inquiry: the first (rf-all), most inclusive model revealed prosodic features associated with the initial response particle phrases (abbreviated with rPrt.*) were more important than those associated with the target sentence (abbreviated with targS.*). Random forests were trained on roughly 75% of the data and tested on the remaining held out utterances.

Table 2: Confusion matrix for rf-all: rPrt + targS features: 84.8% accurate, N=79 (7 lost to na.omit)

| | | Predicted | |
|--------|---|-----------|----|
| | | A | B |
| Actual | A | 22 | 7 |
| | B | 5 | 45 |

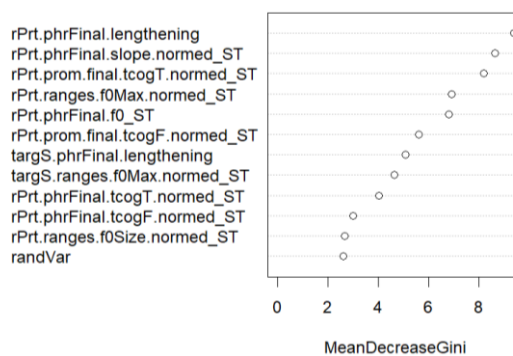


Figure 3: Relative importance of features used in rf-all. Only features that were more important than a random variable parameter are shown.

This suggested that target sentences might not need to “carry” as much of the classification burden and that prosodic cues were primarily present in the response particle phrase. As a result, a second model (rf-target) using only features associated with a target sentence was built.

The target-only model performs less well in terms of classification, although several features are demonstrated to contribute. These features are associated with phrase-final lengthening and overall ranges. This suggests that future inquiries (and parameter calculations) might be focused on phrase boundary cues. Further, other phrase related variables

were not useful despite the value of lengthening. This suggests a possible information loss when phrase-final f0 estimates are compromised by f0 inaccuracies in the final syllable, where robust, regular utterance-final f0 often fades away. This is a consistent difficulty in intonation research, and it is an advantage of this model that it serves to motivate creative solutions to such missing f0 measures.

Table 3: *Confusion matrix for rf-target: targS features: 69% accurate (just better than chance: 62%)*

| | | Predicted | |
|--------|---|-----------|----|
| | | A | B |
| Actual | A | 16 | 16 |
| | B | 10 | 43 |

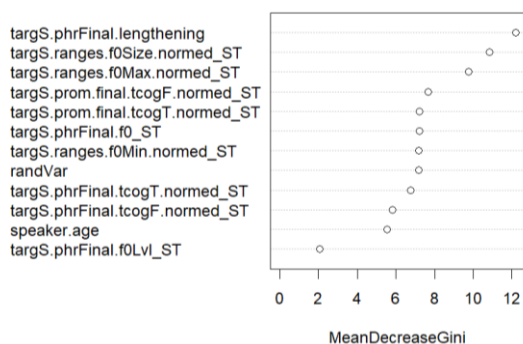


Figure 4: *Relative importance of features used in rf-target. Here features both more and less important than a random variable parameter are shown.*

4. Discussion

Speakers in this study reliably distinguished utterances along a subset of measures identified by Random Forest models: these measures relate to phrase-final prosody, as well as to pitch accent prosody and local pitch range ceiling. This RF model demonstrates a high degree of classification accuracy, indicating that speech produced in *p/¬p* contexts can be distinguished based on their prosodic attributes.

In particular, these RF models found the most influential prosodic characteristics in the response particle portion of the utterance (which is produced before the target sentence is uttered). This finding is noteworthy in at least two ways. First, it shows that the prosodic cues to this meaning are distributed across the utterance and are not restricted to any one prosodic element. Second, the response particle portion of the utterance has two sets of cues about the speaker’s epistemic shift: lexical content (the response particles themselves, which were absent from RF models) and prosodic content (the prosodic measures that come out as important in RF models). This sheds light on the interplay between prosodic and lexical content: it has been suggested that there is “cue trading” across lexical and prosodic content, but that idea is not supported in the domain of this meaning.

These findings also suggest that the phrase boundary implementation provides cues to the two *p/¬p* bias contexts. In addition to phrase final lengthening, the normed f0 and f0 range maximum and the slope of the f0 and at the end of the phrase appear to be discriminative. It may be somewhat surprising that, in the rf-all model (Fig.3), utterance-final intonation for the

target sentence was not important, while the phrase-final (utterance-medial) intonation of the response particle phrases. One possibility is that utterance-final intonation is susceptible to issues with f0 tracking and difficulties in annotation. As a result, we are returning to the data and feature extraction with goals of including phonological analysis of edge tones (e.g., with MAE_ToBI) and of developing better measures of utterance-final f0.

Finally, the Tonal Center of Gravities (in time and in frequency) on the prominent syllable in the response phrase also served to separate the two contexts. Moreover, when response particle phrase cues are excluded from the RF model, the target sentence boundary lengthening and the TCoG-T were also influential in separating *p/¬p* bias context utterances. Since TCoG is linked to phonology [2], the importance of TCoG measures in both the response and the target sentences suggests a phonological difference between the prominences in A vs B contexts in both phrases.

5. Conclusions

These results suggest that a speaker’s prior belief (i.e. their bias) influences the intonation of a utterance produced in response to information from an epistemically authoritative interlocutor. Primary cues were found at phrase boundaries, but also in pitch accents. Further work on the annotation and data extraction is expected to uncover more precise and pointed results. These investigations will include incorporating more information about the speakers than is currently present (e.g., age, gender, race, language background; information already collected) as well as more honed measures of prosody, to which reserachers have been directed by the present Machine Learning results.

The importance of phrase-final intonation in the response phrase was identified through these ML analyses, and the lack of “important” effects from the target sentence phrase-final intonation suggested that we ought to return to our data. After returning to the data, the researchers (as speakers of English and trained expert linguists) did indeed have the impression that many recordings do differ between the two conditions in the target sentence region. In this way, Machine Learning is not used as an end-point classifier but also as input to an iterative analytic process: these results have helped identify directions for further areas of investigation.

In sum, this work demonstrates the usefulness of ML in facilitating and encouraging iterative modelling and analysis, specifically: ML can help identify which aspects of the phonology to focus on and reconsider how those features are extracted, which ought to yield improved ML models— and highlights the usefulness of both domain-specific expertise and mindful usage of modelling tools.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 2042694, 2042702, 204274.

7. References

- [1] B. Ahn, N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos, “PoLaR annotation guidelines (version 1.0).” 2021.
- [2] J. Barnes, N. Veilleux, A. Brugos, and S. Shattuck-Hufnagel, “Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology,” *Laboratory Phonology 3: Change in Phonology*, 2012.
- [3] P. Boersma and D. Weenink, “Praat: doing phonetics by computer [Computer program].” 2023.
- [4] D. Goodhue, “A unified account of inquisitive and assertive rising declaratives,” *Proceedings of the Linguistic Society of America*, vol. 6, no. 1, Art. no. 1, Apr. 2021.
- [5] C. Gunlogson, *True to form: Rising and falling declaratives as questions in English*. New York: Routledge, 2004.
- [6] S. Jeong, “Intonation and sentence-type conventions: Two types of rising declaratives,” *Journal of Semantics*, vol. 35, no. 2, pp. 305–356, 2018.
- [7] A. Liaw and M. Wiener, “Classification and regression by randomForest,” *R news*, vol. 2, no. 3, pp. 18–22, 2002.
- [8] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, “Montreal forced aligner [Computer program] (version 1.0.1).” Apr. 2019.
- [9] C. Poschmann, “All declarative questions are attributive?,” *Belgian Journal of Linguistics*, vol. 22, no. 1, pp. 247–269, 2008.
- [10] R Development Core Team, “R: A language and environment for statistical computing.” Vienna, Austria, manual, 2022. [Online]. Available: <http://www.R-project.org>
- [11] RStudio Team, “RStudio: Integrated Development for R,” Boston, MA, manual, 2020. [Online]. Available: <http://www.rstudio.com>
- [12] D. Rudin, “Intonational Commitments,” *Journal of Semantics*, vol. 39, no. 2, pp. 339–383, May 2022.
- [13] I. Sag and M. Liberman, “The intonational disambiguation of indirect speech acts,” in *Papers from the eleventh regional meeting Chicago Linguistics Society*, 1975, pp. 487–497.
- [14] J. ‘t Hart, R. Collier, and A. Cohen, *A perceptual study of intonation: an experimental-phonetic approach to speech melody*. New York: Cambridge University Press, 1990.
- [15] J. Zehr and F. Schwarz, “PennController for Internet Based Experiments (IBEX),” Mar. 2018.