Methods for PoLaR Explorations with Machine Learning: Grammatical Analysis of Intonation without Grammatical Labels

Nanette Veilleux, Byron Ahn, Alejna Brugos, Sunwoo Jeong, & Stefanie Shattuck-Hufnagel

Simmons University, Princeton University, Simmons University, Seoul National University, MIT

ABSTRACT

This study provides a proof-of-concept for a new method for analyzing intonational form and meaning, demonstrated by analysis of mirative utterances in American English. Here, K-means clustering using measures derived from PoLaR labels (i.e., TCoG) revealed emergent clusters of pitch accents that are suggestive of familiar phonological categories (e.g., MAE_ToBI H* and L+H*). A Random Forest analysis then classified utterance-level meaning based on measures from both smaller granularity (about clusters and acoustics) was subsequently (related to individual pitch accents) and larger granularity (e.g., global f0 information), showing >85% correct categorization of exclamative vs filler sentences.

This work has implications for how to model mappings between prosody and meaning, especially where existing phonological categories alone don't identify semantic/pragmatic categories.

Keywords: intonation, methodology, phonetics phonology interface, form-meaning mapping, machine learning

I. INTRODUCTION

Intonation is known to convey a wide range of meanings, but exploring intonational form-meaning mapping has been challenging (e.g., [1] and [2] for some recent overviews, and [3], [4], [5], and [6] for some critical junctures in the development of the theoretical landscape). This challenge stems in part from the persistent indeterminacy regarding the relevant units of analysis, both on the form side (i.e., which phonetic and phonological aspects of tunes signal systematic meaning differences?) and on the meaning side (i.e., what types of meanings are conventionally encoded by tunes?)

On the form side, using phonological categories alone can miss important details, such as gradient variation in f0 slope that may generate incremental shifts in meaning. But unpacking these categories into global acoustic measures and treating them indiscriminately (in, e.g., machine learning), may miss the key generalization that linguistically meaningful prosodic features are often localized.

We address this methodological hurdle, with a case study on the intonation of mirativity in mainstream American English (henceforth MAE). Mirativity can be defined intuitively as an expression of speaker surprise and a perceived violation of speaker expectations (regarding a proposition). Manifesting in exclamatives like (1), it can be marked by certain particles (e.g., *'Wow!'*), or by designated syntactic configurations (e.g., the wh-fronting *without* subj-aux inversion in (1); compare this to (2), a non-exclamative), and most relevant here, intonation.

(1) (Wow!) How believable Theodore is! [*exclamative, conveying mirativity*]

(2) How is Theodore believable? [*non-exclamative, not conveying mirativity*]

Regarding the intonational correlates of mirativity, previous work [7] identifies certain (phonologically defined) pitch accents as its primary prosodic cue, but also points out that additional gradient cues may be at play. Building on this, we have developed a method to clarify the aspects of intonation associated with meanings of mirativity, by annotating the corpus data from that paper for some of its acoustic characteristics, and submitting that phonologically-informed acoustic information to machine learning.

More specifically, we use PoLaR ([8], [9]) to identify relevant acoustic cues in phonologically-defined regions (e.g., pitch accents), and submit the resulting labels and related measures (e.g., tonal center of gravity; TCoG [12]) to k-means clustering, thereby bundling accent-related measures in a form that can be converted to utterance-level information (in the form of, e.g., each accent cluster's rate of occurrence).

Utterance level features (including information about labels, clusters and acoustics) were subsequently submitted to a Random Forest. The results produced over 85% correct categorization of a balanced sample of over 250 exclamative vs. filler sentences. This approach has more general implications for analyzing intonational meaning, and establishes a method that is extendable to other prosody-meaning mappings where existing phonological categories alone do not distinguish semantic/pragmatic categories.

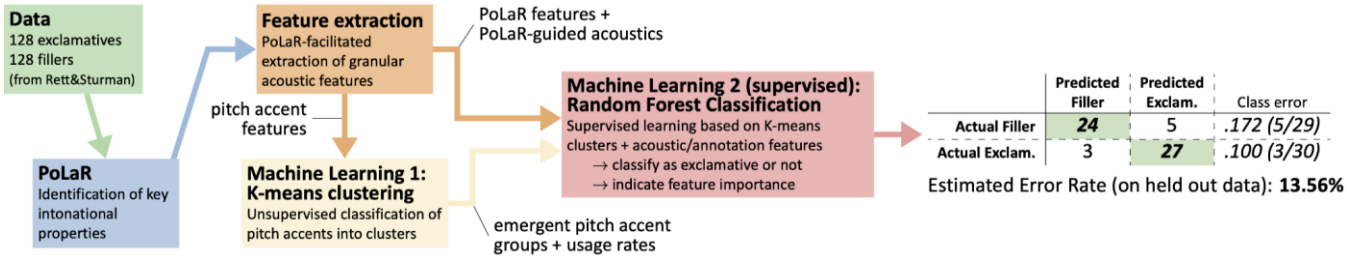

Figure 1. Overall methodology and final results of analysis

II. METHODS

The dataset used in this study is a corpus of 256 utterances collected by [7], in which two MAE speakers read scripts, half of which were exclamatives (e.g., '*Wow, is that nice!*') and the other half not (e.g., '*Is that nice?*'), occurring in four different syntactic frames (declarative, subject-aux inversion, fronted WH-phrase, definite nominal). This dataset was analyzed following the flowchart in Fig.1. The recordings were first force-aligned ([10]) and then PoLaR-labelled in Praat ([11]), as illustrated in the lower half of Fig.2.

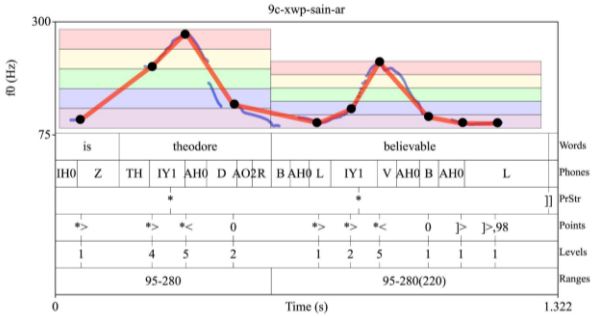

Figure 2: An annotated recording from the corpus

The f0 visualization in the top half of Fig.2 has been marked up to show some key aspects of PoLaR labelling, described more fully in [8, 9]. The dots are at coordinates of (time, f0), where the time value depends on the timing of labels on the Points tier, and the f0 coordinate is either taken from special Points tier labels ("comma override values") or (as is default) from the f0 value calculated by Praat (shown in blue). Interpolating between these dots creates a straight line approximation of the f0 (shown as a red line). PoLaR labels also include a Ranges tier, which defines the *local* f0 floor/ceiling, which can change over an utterance (or even within a phrase). Each range defined by the floor/ceiling is divided into evenly-spaced quintiles (shown as colored bands); the quintile in which the (time, f0) coordinates of a Points label occur is translated into a numerical value (1-5), which is transcribed on the Levels tier. Levels labels thus encode scaled pitch that is normalized relative to the local intonational context. In addition to these three tiers annotating f0 properties, the PrStr (Prosodic Structure) tier contains minimal phonological labels that indicate perceived prominences (*) and boundaries (]).

The annotation process involved three pairs of labellers. The two members of a pair each labeled alone, then compared their labels, and discussed disagreements to generate consensus labels. The first pair labeled according to the basic PoLaR annotation guidelines [9], and the second two pairs used the advanced annotation guidelines, which (among other things) augment basic Points labels to indicate whether Points-defined f0 movements are related to prominences and/or boundaries as labeled on the PrStr tier.

For each of these labeled recordings, a variety of features were subsequently extracted for use in analysis. Certain f0 attributes (such as maximum, minimum, average) were calculated both in raw values and z-score normalized by speaker. PoLaR-labelled TextGrids facilitated extraction and calculation of additional features, including (1) Features based on PoLaR labels themselves (e.g., counts of phrasal prominences and boundaries), (2) direct measures such as timing and (normalized) pitch values, and (3) derived measures such as slope between certain PrStr-associated Points labels and Tonal Center of Gravity ([12]) relative to local f0 ranges (i.e. PoLaR Ranges labels).

Machine Learning 1: pitch accent type clusters

The machine learning modeling of this data takes place in two sequential stages: unsupervised clustering of pitch accents, and supervised random forest categorization of utterances's mirativity. The reason for this two-stage approach is two fold. First, even though ToBI-type pitch accents can signal semantic/pragmatic differences (e.g., [4]), we suspect that they may be too broad for capturing all the relevant distinctions; so, we used unsupervised learning to capture distinctions without bundling them into these umbrella phonological categories. The second issue is at the heart of the difficulties of using machine learning in prosodic form-meaning mapping: prosodic events occur at a more local level (e.g., a syllable or word) than meaning events (e.g., an utterance). This work is an example of classifying each utterance as one of two categories (exclamative

vs filler) and using the characteristics of the (usually multiple) pitch accents in each utterance. If the utterances in each category had been more parallel, then the unique pitch accent on the target words could have been used. Here, we use the percentage of pitch accents belonging to each stage-one cluster in the final random forest categorization model.

An unsupervised k-means clustering algorithm ([13], [14]) was used to model the pitch accent types, resulting in 3 clusters. Though this was a linguistically informed question (based on [7]), the k-means algorithm automatically determines the number of clusters and the feature values associated with each cluster. After systematic exploration of various combinations of intonational features (mentioned above) to the clustering algorithm, we choose the feature set that produced clusters with the best Sum of Squares characteristics. As a result, the two pitch accent measures used in the clustering algorithm were Tonal Center of Gravity (TCoG) measures, which have also been previously shown to differentiate pitch accents ([12]). Specifically, TCoG was measured over the pitch-accent's rise, and we submitted two relativized (and subsequently z-scored) values: the time of the TCoG relative to the vowel center (tcogT), and the frequency of the TCoG relative to the Range min/max (tcogF). The resulting three clusters are shown as different shapes/colors in Fig.3, with each cluster's centroid annotated in black.

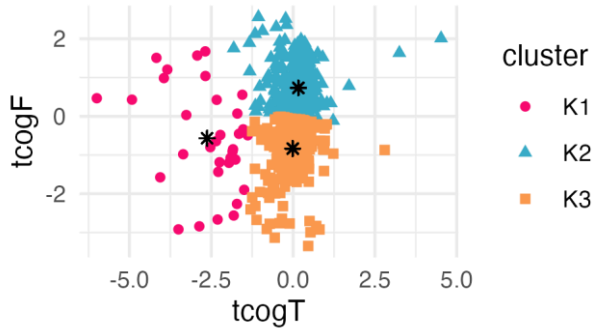

Figure 3: Results of K-Means clustering

These three clusters are suggestive of MAE_ToBI labels: K1 as H* with a preceding high target, K3 as H* without a preceding high target, and K2 as L+H*.

Machine Learning 2: classification as exclamative vs. filler

The ultimate goal of the machine learning model is to determine if exclamatives can be categorized separately from the filler sentences and what features contribute to this separation. In this stage, a supervised random forest model ([13], [15]) was used to classify exclamatives and filler sentences. The particle "wow" was excised from exclamative input utterances. In addition to the rate at which each pitch accent cluster occurred in an utterance, other acoustic and semantic features served as input to this classification, listed below. (Data analysis materials can be found at [16].)

Direct acoustic measures and derived features

- Changes in f0 (max, min, average, delta): raw and z-score normalized by speaker
- Tonal Centers of Gravity (time, frequency)

PoLaR label features

- Measures of f0 for the utterance: timing and (normalized) pitch values for turning points in f0, (local) f0 Ranges, location of prominences and boundaries
- Counts of prominences and phrase boundaries, raw and as a ratio of number of content words

Semantic features:

- Semantic type (content vs function word) for the word containing the maximum f0

III. RESULTS

A random forest was trained on a random sample of 77% of the data (197 utterances: 99 fillers and 98 exclamatives from a set of 256 equally distributed utterances.). When this model was tested on the remaining 23% of the data (59 utterances, 29 filler, 30 Excl.), the resulting classification has a 86% accuracy rate [95% CI : (0.7502, 0.9396)]. The confusion matrix shown at the end of the flowchart in Figure 1 further details this model's output. In examining the mis-classification by class, fillers tend to be misclassified as exclamatives more often. However, given the limitations of the data set size, these values are sensitive to the random selection of the test set.

In addition, Random Forest models allow the predictors to be ranked according to importance (which tends not to vary significantly between the random selections of training/test sets), as shown in Fig.4 for the present model.

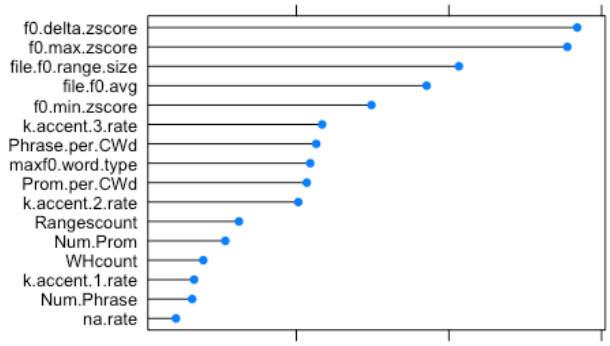

Figure 4: Relative importance of factors in the Random Forest classification of Exclamatives vs Fillers.

Each utterances' z-scored f0 max-min and f0 max were ranked highest. Accent types still play a role despite the use of related acoustic measures.

IV. DISCUSSION

First, we discuss the broad findings of this study. This work reaffirms the idea that semantic mirativity is marked intonationally. Moreover, the specific findings are strikingly similar to [7]'s, suggesting that their analysis does not depend on the use of MAE_ToBI labels or adoption of its grammar. While the rate of occurrence of a particular pitch accent type (here: k-means cluster) is not as important as other acoustic characteristics for classification in this model (see Fig.4), the features that do rank high in importance (e.g., f0.delta.zscore) may be the very cues that identify L+H* apart from other accents — again consistent with [7]'s analysis. On the other hand, the meaning of mirativity may arise *not just* from (cues to) L+H* alone, but rather from some constellation of intonational characteristics, including categories of intonational phonology (such as L+H*) and other gradient components of intonation.

Methodologically, the main contribution here is a proof-of-concept machine learning analysis that is on par with established qualitative methods for intonational meaning, since our results are broadly consistent with the findings in [7]. Moreover, unlike a model that uses only global acoustics, the use of PoLaR labels has the advantage of enabling the targeting of acoustics from smaller, *phonologically-relevant* domains, thereby allowing for a clearer characterization of the intonational form of MAE exclamatives.

Combining PoLaR annotation with ML techniques opens many new avenues for pursuing form-meaning mapping research, in areas where it's not (yet) clear what the categories of form are. This is facilitated by PoLaR's capacity to capture linguistically-informed acoustic measures without presupposing a particular set of phonological categories. In fact, this methodology can help identify intonational categories by revealing which aspects of form map onto particular meanings. In addition, for cases where there are no discrete phonological categories (i.e., if the meanings and forms are not grammatically structured; e.g., for so-called "paralinguistic" uses of intonation), PoLaR labels can potentially identify which dimensions of acoustic form are relevant for signalling particular meanings.

Beyond using ML techniques alongside PoLaR, using PoLaR for intonational research is itself advantageous, as it does not require the same extensive training or experience that other labelling systems might. Instead, labelers are able to identify appropriate regions for collecting the salient acoustic measurements that feed into the statistical and machine learning analyses. Utterance-level acoustic measures (e.g., f0 min/max/average) are insufficiently targeted, as the most critical acoustic values are often localized in specific phonologically-relevant regions. On the other hand, strictly phonological annotation systems run the risk of ignoring key patterns in the acoustics that may be of interest in conveying meaning (for discussion on this point, see [8]). In contrast, PoLaR identifies phonologically-informed acoustic measures which can be input into ML models of intonational form-meaning relationships, to discover potential meaning-bearing aspects of the f0 signal that are not directly related to phonological contrasts.

Turning now to an analysis of how to formally model intonational meaning, these results are consistent with a model in which L+H* is the marker of mirativity. Despite this, caution should be exercised in modelling this relationship with a conventional and categorical one-to-one mapping between L+H* form and mirative meaning. Instead, we speculate that it may be more fitting to propose a many-to-many mapping between form and meaning, with multiple intonational cues (related to or partly consisting of L+H*) marking multiple possible interpretations. This is especially plausible since L+H* has been argued to mark other meanings, such as contrastive focus. Additionally, determining whether these ML algorithms reflect the intonational factors that matter for interpretation of utterances by human listeners, will require extensive study of human intonation perception.

V. CONCLUSIONS

The key contribution of this paper is its demonstration of a new methodology for exploring intonational form and meaning. We are confident in the applicability/usefulness of this methodology, but due to practical limitations such as the size of the corpus, we do not yet draw strong conclusions about the phonology of exclamatives in MAE and what in the semantics maps onto the relevant phonology. Coupling this methodology with a deeper investigation into the production and (human) perception of these intonational variables (e.g., with minimal pairs that differ in terms of the intonational dimensions identified here) may be able to illuminate form-meaning relationships *as well as* the phonetics-phonology interface.

VI. REFERENCES

[1] H. Truckenbrodt, "Semantics of intonation," in *Semantics: An international handbook of natural language meaning*, vol. 3, K. V. Heusinger, C. Maienborn, and P. Portner, Eds. De Gruyter, 2012, pp. 2039–2069. doi: 10.1515/9783110253382.2039.

[2] D. Büring, *Intonation and meaning*. Oxford: Oxford University Press, 2016.

[3] G. Ward and J. Hirschberg, "Implicating uncertainty: The pragmatics of fall-rise intonation," *Language*, vol. 61, pp. 747–776, 1985.

[4] J. Pierrehumbert and J. B. Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," pp. 271–311, 1990, doi: 10.7916/D8KD24FP.

[5] M. Steedman, "Information structure and syntax-phonology interface," *Linguistic Inquiry*, vol. 31, no. 4, pp. 649–689, 2000.

[6] N. Constant, "English rise-fall-rise: A study in the semantics and pragmatics of intonation," *Linguistics and Philosophy*, vol. 35, no. 5, pp. 407–442, 2012.

[7] J. Rett and B. Sturman, "Prosodically marked mirativity," in *Proceedings of the 37th West Coast Conference on Formal Linguistics*, D. K. E. Reisinger and M. Huijsmans, Eds. Somerville, MA: Cascadilla Proceedings Project, 2021, pp. 1–20.

[8] B. Ahn, N. Veilleux, and S. Shattuck-Hufnagel, "Annotating prosody with PoLaR: Conventions for a decompositional annotation system," in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Canberra, Australia: Australasian Speech Science and Technology Association Inc., 2019, pp. 1302–1306.

[9] B. Ahn, N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos, "PoLaR annotation guidelines (version 1.0)." 2021. doi: 10.17605/OSF.IO/USBX5.

[10] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner [Computer program] (version 1.0.1)." Apr. 2019. doi: 10.5281/zenodo.2630943.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]." Feb. 2022.

[12] 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*, vol. 3, no. 2, pp. 337–383, 2012, doi: 10.1515/lp-2012-0017.

[13] R Development Core Team, "R: A language and environment for statistical computing," Vienna, Austria, manual, 2022. [Online]. Available: http://www.R-project.org

[14] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, "NbClust: An R package for determining the relevant number of clusters in a data set," *Journal of Statistical Software*, vol. 61, no. 6, pp. 1–36, 2014.

[15] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[16] Ahn, B, Brugos, A., Jeong, S., Shattuck-Hufnagel, S. Veilleux,N, "A Case Study for PoLaR Explorations with Machine Learning", 2023. doi: 10.17605/OSF.IO/H8N6F.

VII. ACKNOWLEDGEMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2042694, 2042702, and 2042748. We would like to thank Jessica Rett and Beth Sturman for consulting and sharing their corpus. We are also indebted to the undergraduate members of our research team: Ismah Ahmed, Samin Charepoo, Julia Hartnett, Michael Malvone, Tal Schaeffer, Sreeniketh Vogoti, and Megan Willis, who did a bulk of the PoLaR labelling, as well as Shirley Fong, Madeline Guettler, and Sofia Hirschman, who were invaluable in the development of the machine learning models.