Marking prominence: Towards cue-based annotation of prosodic prominence

Alejna Brugos[1], Mara Breen[2], Stefanie Shattuck-Hufnagel[3], Nanette Veilleux[1], Jonathan Barnes[4]

[1]Simmons University, [2]Mount Holyoke College, [3]Massachusetts Institute of Technology, [4]Boston University
alejna.brugos@simmons.edu, mbreen@mtholyoke.edu, sshuf@mit.edu, veilleux@simmons.edu, jabarnes@bu.edu

ABSTRACT

Phrase-level prosodic prominence in American English is understood, in the AM tradition, to be marked by pitch accents. While such prominences are characterized via tonal labels in ToBI (e.g. H*), their cues are not exclusively in the pitch domain: timing, loudness and voice quality are known to contribute to prominence perception. All of these cues occur with a wide degree of variability in naturally produced speech, and this variation may be informative. In this study, we advance towards a system of explicit labelling of individual cues to prosodic structure, here focusing on phrase-level prominence. We examine correlations between the presence of a set of 6 cues to prominence (relating to segment duration, loudness, and non-modal phonation, in addition to f0) and pitch accent labels in a corpus of ToBI-labelled American English speech. Results suggest that tokens with more cues are more likely to receive a pitch accent label.

Keywords: prosodic prominence, ToBI, phonetic cues, prosodic annotation, cue integration

1. INTRODUCTION

Phonological categories, both segmental and prosodic, are realized with great variability in the signal. Evidence from the segmental domain suggests that speakers systematically control aspects of speech below the level of contrastive linguistic constituents, and that listeners are sensitive to this variation [1,2]. Likewise, in the study of prosodic boundary and prominence events, there is well-documented variation in the cues that signal a prosodic category (see [3] for overview).

ToBI (for **To**nes and **B**reak **I**ndices) [4,5] is a phonological prosodic annotation system in which labellers mark prosodic categories of prominence (pitch accents) and grouping (boundaries). In annotating prominences with pitch-accent labels, a labeller may be uncertain as to the tonal identity of the pitch accent (i.e. H or L tonal categories), or whether a word is prominent at all. We hypothesize that this uncertainty might be related to cue density, in the sense that prosodic events evidenced by more cues may be easier to categorize than events with fewer cues.

Additional motivation for labelling individual cues to prosodic categories comes from the fact that informative phonetic variability is lost when considering categorical labels alone. Further, when there is disagreement and uncertainty, it is beneficial to know which aspects of the signal influence the labeller's category assignment. [3] advocate for "the identification of individual cues to the contrastive prosodic elements of an utterance." As argued in [3, 6], an essential goal of prosodic transcription is to identify contrastive linguistic categories of utterances, and not merely the salient aspects of the surface form. Towards this goal, we explore explicitly labelling acoustic cues to prosodic categories. Using a small corpus of ToBI-labelled American English speech we compare candidate prominence cue annotations to independently annotated ToBI pitch-accent labels.

1.1. Cues to prosodic prominence

In spoken American English, prosodic prominences are marked by a number of cues, including f0 excursions, duration changes, loudness, and voice quality (See [7] for an overview of prominence cues in English, & [8] for discussion of prominence cross-linguistically.) [9] found significant duration correlates of prominence (marked with Rapid Prosody Transcription [3]), as well as weaker influences from filtered intensity and little or no correlation with F0. Although the ToBI system focuses on F0 markers of phrase-level prominence (e.g. [5]), [10] reports that loudness predicts prominence, but F0 lends little to its perception. Thus there is disagreement about which acoustic cues correlate with perceived prominence; systematic labelling of cue patterns in conjunction with perceived phonological prominence will contribute to resolving these issues.

Phonetic cues to prominence have been extensively discussed in the literature, along with the role of cues for prominence perception (e.g. [7, 11]) and automatic prominence detection (e.g. [12]). However, we are unaware of any system designed for explicit annotation of prominence cues. The current proposal, like [13], complements categorical prosodic annotation systems.

1.2 Annotation of prosodic categories

ToBI [4,5] is a system of prosodic annotation based largely on the autosegmental-metrical model of prosodic phonology [14], with which users identify and label phonological categories of prosodic events. Tonal events indicate prominence or grouping, categorized as either pitch accents or two levels of edge tone. ToBI annotations are tier-based labels time-aligned to the speech signal. Here we focus on the *tones* tier for labelling pitch-related events of prominence (pitch accents) and phrase-level boundaries (phrase accents and boundary tones), which are placed in conjunction with orthographic text in a *words* tier.

2. METHODS

In the current study, we compare a set of acoustic cues to prosodic prominence with independently labelled MAE (Mainstream American English) ToBI pitch accent labels. We hypothesize that where prominence cues converge, there will be greater likelihood of a perceived phrase-level prominence, operationalized by ToBI pitch accent labels. In cases where only a subset of cues are present, we predict labellers will exhibit more uncertainty and disagreement about pitch accents.

The corpus [15] used to test this prosodic prominence-related cue labelling is the same used in [13] and [16]. It consists of 8 files, comprising roughly 6 minutes of speech: 178 seconds professionally read and 181 seconds of spontaneous speech. The files were produced by 7 individual speakers, roughly balanced between male and female, containing a total of 1076 words (1483 syllables). All files were labelled with MAE ToBI conventions by four expert labellers [16].

The cue annotation used 6 labels described below, taking inspiration from the cue labelling of prosodic boundaries [13], disfluencies [17] , and stuttered speech [18]. Two labels (**proc** and **prv**) are based on timing cues, and are adapted from the disfluency and stuttering annotations [17, 18]. Whereas [13] used a single label (**pr**) to capture segmental lengthening, we here expand this label to better capture lengthening in terms of segment type and position within the syllable (vowel, **prv** vs onset consonant, **proc**), taking on some of the granularity proposed in disfluency annotations of [19] and capturing that prominence-related lengthening primarily affects durations of syllable onsets and nuclei [20]. A new label is tested relating to amplitude (**ampi**), standing in for loudness. Two labels relating to voice quality are used, including irregular pitch periods of syllable beginning with a vowel (**gi**, used similarly to [13]) and a new label intended to capture other amodal phonation (**vq**) that may contribute to perceived prominence, such as localized creakiness or breathiness. All of these are labelled specifically when they are heard in conjunction with a prominence, and when their presence is interpreted as contributing to that prominence perception:

proc: Prolongation of the onset consonant
prv: Prolongation of a vowel
f0: A local f0 event (e.g. peak or valley) on or around a stressed syllable of a word
ampi: A local increase in vowel amplitude
gi: Irregular pitch periods at the beginning a vowel (or other sonorant consonant)
vq: Other amodal voice quality of a vowel

An expert prosodic labeller independently labelled the eight files for both perceived prosodic structure and cues. Labels for phrasal prominence and boundary locations were adapted from PoLaR Prosodic Structure tier labels [21], similar to RPT labels [9]. Tokens perceived as clearly prominent were labelled with a star (*); cases where the labeller was uncertain about a prominence were labelled with a star and a question mark (*?). For each perceived prominence (whether * or *?), the labeller chose from among the 6 proposed prominence cue labels.

Cue labels were placed on a point tier, time-aligned to be within the associated vowel's interval. Labels were used in sequences delimited by a period. Each syllable could hypothetically be labelled with between 0 and 6 cues from this set (e.g. no prominence labels, or only **proc** or **proc.prv.ampi.f0.vq**), but tokens marked with all 6 cues were not expected, as **gi** and **proc** are typically mutually exclusive, as **gi** only rarely occurs with sonorant consonants). All files were annotated

Cue	proc	prv	ampi	f0	gi	vq
Count	373	400	392	368	40	47

Table 1. Number of times each acoustic cue was labelled in the ToBI-annotated corpus.

using Praat TextGrids [22] and labels were automatically extracted for analysis. Of the 1483 syllables in the data set, 591 syllables were labelled as potentially prominent by the cue labeller, with 1 or more prominence cues. Each individual cue appeared between 40 and 400 times (Table 1). Because the number of tokens labelled with the two voice quality cues were relatively few, and no tokens were labelled as both gi and vq, these cues were collapsed into "vq" in the analyses.

3. ANALYSIS & RESULTS

Table 2 shows correspondence between the number of cues for a given token and pitch accent score, which is the number of the four ToBI labellers who indicated a pitch accent label (whether with tones, or the uncertainty marker). Tokens associated with an increased number of cues tend to be labelled as pitch-accented by more labellers.

	Number of labellers indicating * or *?				
	0	1	2	3	4
# of Cues					
0	749	84	19	31	9
1	19	6	16	17	30
2	14	15	26	25	84
3	7	6	9	26	114
4	1	2	2	18	144
5	0	0	0	1	9

Table 2: Number of labellers indicating a pitch accent by number of cues labelled.

We used linear mixed effects regression to determine whether prominence ratings (i.e., ToBI pitch accent labels) increased with the number of acoustic cues. We assigned each word the average prominence rating of the four labellers: any pitch accent label category was coded as 1; an uncertain pitch accent (***?**) was coded as 0.5; no pitch accent was coded as 0. These scores were averaged for each token. Number of cues was a significant predictor of average prominence rating, B = 0.24, SE = 0.004, t = 57.61, p < .01 (Figure 1).

We used linear mixed effect regression to predict the average prominence rating of each token from the five cues. Each cue was a fixed effect in the model; speaker was included as a random effect. The fixed effects are shown in Table 3 and Figure 2.

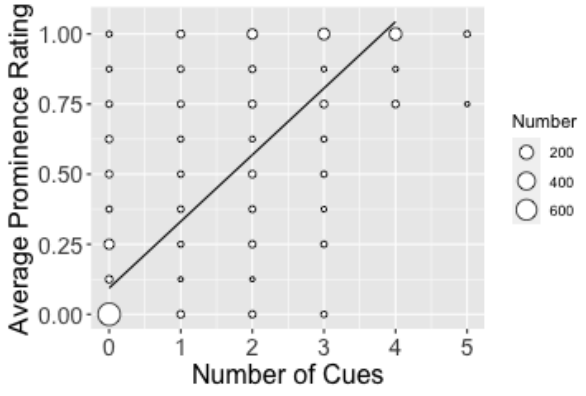

Fig. 1. Average prominence rating by number of cues per token. Dot size indicates number of tokens. For clarity of display, cases with only one token in a cell are excluded.

	Estimate	SE	t	p-value
(Intercept)	0.68	0.02	38.19	0.00
proc	0.26	0.02	13.25	0.00
prv	0.30	0.02	16.27	0.00
f0	0.28	0.02	13.93	0.00
ampi	0.13	0.02	6.83	0.00
vq	0.20	0.03	7.38	0.00

Table 3. Fixed effects in the model predicting average prominence rating from each cue.

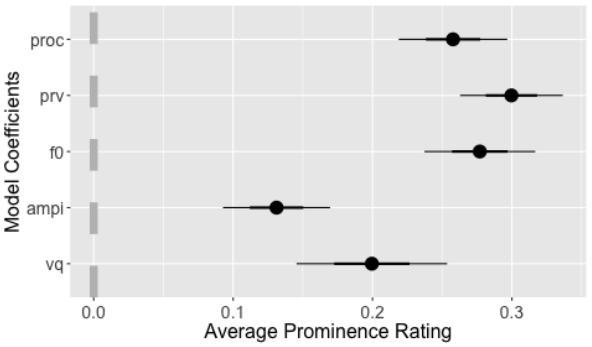

Figure 2. Coefficients for fixed effects in the model predicting average prominence rating from each cue.

The five cues (where **vq** includes the 6th cue, **gi**) independently increased the average prominence rating by ToBI labellers, demonstrating that each cue in the current study can signal pitch accents.

Finally, we assessed the relationship between number of acoustic cues and labeller uncertainty. If any of the ToBI labellers labelled the syllable with *?, we considered it uncertain. Table 4 shows the frequency of uncertainty by the number of cues labelled for each token. A chi-square analysis shows that the proportion of uncertain labels was higher for a small number of cues than for no cues, or many, $X^2(df = 5, N = 1483) = 100.32$, $p < .01$.

		Cue Number					
		0	1	2	3	4	5
Uncertain	0	782	54	101	115	120	7
	1	110	34	63	47	47	3
%		12.3	63.0	62.4	29.0	28.1	30.0

Table 4: Labeller uncertainty cases by cue number.

4. DISCUSSION

We found a strong correlation between the number of annotated acoustic cues and pitch accent labels from four independent ToBI labellers. Moreover, there was greater uncertainty in ToBI labels in cases with a small number of cues than in cases with many or no cues. However, a more informative

picture emerges by examining cases that countered predictions, namely: 1) where the cue labeller perceived a potential prominence and labelled acoustic cues to that prominence, but few or no ToBI labellers marked a pitch accent and 2) where the cue labeller did not mark a prominence, and labelled none of the cues, but one or more ToBI labellers marked a pitch accent. (See figure 3.)

We consider several possible explanations for such divergences. First, the method of capturing cues for only syllables where a prominence was perceived by the labeller entails that tokens not heard as prominent were not labelled for cues. Some cues may have been arguably present but did not meet a threshold of prominence for this listener, but may have been sufficient to cue prominence for another. Further, cues were labelled in a binary way: present or absent. For cases with high agreement of the presence of a pitch accent, but only one or two labelled cues, tokens may have evidenced those cues with a stronger magnitude (e.g. an extra long or loud vowel, or a more extreme f0 excursion) and therefore a single cue led to perceived prominence.

Why label cues at all and not just extract acoustic data? Like [7], we believe that the acoustic characteristics are always interpreted in the context of the prosodic structure: the listener "interprets most fine-grained phonetic variation only after having performed a parse of the signal into a coarser, more discrete sequence of categorical events [7]." Specifically, acoustic features are always context- and speaker-dependent, and must be interpreted with reference to this context. Not only are acoustic values (e.g. duration and amplitude) variable across discourses, they also vary by contextual factors such as speech rate, loudness and other paralinguistic factors. In addition, listeners perceive subtle acoustic variability across a range of listening conditions, including in noise where feature extraction is less reliable. Ultimately, prosodic categories are likely integrated perceptually [23], as suggested for category perception in a wide range of domains, both linguistic [24, 25] and auditory [26].

In this study we focus on the binary presence (or absence) of a given cue perceived by a trained labeller, but that these cues can and do also vary in their magnitude. The lengthening of onset consonants and vowels vary, as well as the size of f0 excursions of pitch accents. Further, directly annotating the cues to prosodic events may shed light on additional sources of ambiguity, such as when a labeller is uncertain as to whether a given cue indicates a prominence event (pitch accent) or a boundary event, or some combination of the two. Still another source of contextual ambiguity is in the case of neighboring strong syllables: labellers must decide which one is prominent–one, the other or both. (And all of these factors likely interact with signal-extrinsic factors [7,9]). Reliability of cues may vary; for example, speakers with creakier voices may not signal boundaries with irregular pitch periods, or such cues may be less salient.

A future direction of investigation would be to combine labelling of cues to prominence and boundary. For example, the voice quality cue **gi** can signal that a vowel-initial word is either phrase-initial or that is pitch accented, or in some cases both [27]. Further, the overlapping of cues of prominence and boundary may itself be informative, highlighting the prominence of certain structural positions [8]. Explicit labelling of prosodic cues can give insight to the relative strength of the instantiations of those categories, reflecting suggested prominence hierarchies [7].

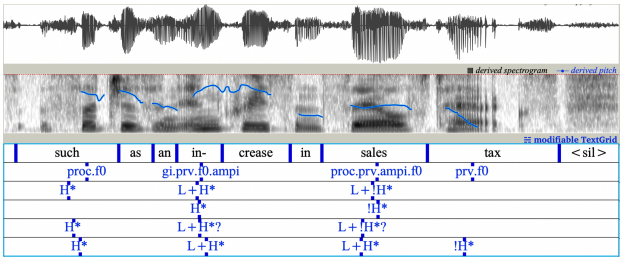

Figure 3: Labelling example, with two tokens ("in-" and "sales") where 4 cue labels were marked and all labellers labelled a pitch accent, and two tokens ("such" and "tax") where 2 cues were marked, but only a subset of labellers labelled a pitch accent. Tiers are syllables, cues, and pitch accent labels from each of four ToBI labellers.

5. CONCLUSIONS

Speakers control a set of acoustic cues with which they may signal prosodic structure. However, not all cues are used equally, and not all speakers use the cues in the same way. Listeners (and prosodic labellers) are sensitive to aspects of the speech signal that are cues to phonological categories, signaling prosodic structure events of phrasing and prominence. This study shows promising support for the meaningful relations between labels intended to reflect the presence of specific cues in the acoustic signal, and categorical/phonological labels of pitch accents as used in ToBI labelling.

6. ACKNOWLEDGEMENTS

This material is based partly upon work supported by the National Science Foundation under Grant No. 2042694, 2042702, and 2042748. We are also grateful to several undergraduate research assistants, whose insights contributed to the development of these labelling methods: Kallie Dimaris, Laena Tieng, and Elaine Wang.

7. REFERENCES

[1] Cole, J., Shattuck-Hufnagel, S. 2018. Quantifying phonetic variation: landmark labelling of imitated utterances. In Cangemi et. al (eds) *Rethinking Reduction*. De Gruyter Mouton, 164-204.

[2] Turk, A., Shattuck-Hufnagel, S. 2014. Timing in talking: what is it used for, and how is it controlled?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1658), 20130395.

[3] Cole, J., Shattuck-Hufnagel, S. 2016. New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1) 8, 1–29.

[4] Beckman, M., Ayers Elam, G. 1993/1997. Guidelines for ToBI Labelling (Version 3, March 1997) Copyright (1993) the Ohio State University Research Foundation.

[5] Beckman, M., Hirschberg, J., Shattuck Hufnagel, S. 2005. The Original ToBI System and the Evolution of the ToBI Framework. In S. A. Jun (ed.) *Prosodic Typology: the Phonology of Intonation and Phrasing*, 9–54.

[6] Frota, S. 2016. Surface and Structure: Transcribing Intonation within and across Languages. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 7(1): 7. 1–19.

[7] Bishop, J., Kuo, G., Kim, B. 2020. Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody Transcription. *Journal of Phonetics*, 82, 100977.

[8] Grice, M., Kügler, F. 2021. Prosodic prominence–a cross-linguistic perspective. *Language and Speech* 64.2 (2021), 253-260.

[9] Cole, J., Mo, Y., Hasegawa-Johnson, M. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology*, 1(2), 425-452.

[10] Kochanski, G., Grabe, E., Coleman, J., Rosner, B. 2005. Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118(2), 1038-1054.

[11] Ludusan, B., Wagner, P., Wlodarczak, M. 2021. Cue Interaction in the Perception of Prosodic Prominence: The Role of Voice Quality. *Proc. Interspeech 2021*, 1006-1010.

[12] Rosenberg, A. 2010. AutoBI-a tool for automatic ToBI annotation. *Proc Interspeech 2010*, 146-149.

[13] Brugos, A., Breen, M., Veilleux, N., Barnes, J., Shattuck-Hufnagel, S. 2018. Cue-based annotation and analysis of prosodic boundary events. *Proc. Speech Prosody 2018*, 245-249.

[14] Pierrehumbert, J. 1980. The phonology and phonetics of English intonation (Unpublished Ph.D. thesis). Department of Linguistics and Philosophy, Massachusetts Institute of Technology.

[15] Breen, M., Dilley, L. C., Kraemer, J., Gibson, E. 2012. Inter-transcriber agreement for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory* 8, 277–312.

[16] Dilley, L. C., Breen, M., Brown, M., Gibson, E. A. F. Rhythm and Pitch. LDC2018S04. Web Download. Philadelphia: Linguistic Data Consortium, 2018.

[17] Brugos, A., Langston, A., Shattuck-Hufnagel, S., Veilleux, N. 2019. A cue-based approach to prosodic disfluency annotation. *Proc. of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 3413-3417.

[18] Arbisi-Kelm, T. R. 2006. An intonational analysis of disfluency patterns in stuttering (Dissertation, UCLA).

[19] McDougall, K., Duckworth, M. 2017. Profiling fluency: An analysis of individual variation in disfluencies in adult males. *Speech Communication* 9, 16-27.

[20] Turk, A. E., White, L. 1999. Structural influences on accentual lengthening in English. *Journal of Phonetics,* 27(2), 171-206.

[21] Ahn, B., Veilleux, N., Shattuck-Hufnagel, S., Brugos, A. 2021. PoLaR Annotation Guidelines (version 1.0). Available at https://osf.io/usbx5/. doi: 10.17605/OSF.IO/USBX5.

[22] Boersma, P., Weenink, D. 2022. Praat: doing phonetics by computer [Computer program]. Version 6.3.03, retrieved 17 December 2022 from www.praat.org

[23] Brugos, A.M., 2015. *The interaction of pitch and timing in the perception of prosodic grouping* (Doctoral dissertation, Boston University).

[24] Martin, A. E. 2016. Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology. *Frontiers in Psychology*, 7, 120.

[25] Brown, M., Tanenhaus, M. K., Dilley, L. 2021. Syllable inference as a mechanism for spoken language understanding. *Topics in Cognitive Science*, 13(2), 351-398.

[26] Bregman, A. S. 1994. *Auditory scene analysis: The perceptual organization of sound*. MIT press.

[27] Dilley, L., Shattuck Hufnagel, S., Ostendorf, M. 1996. Glottalization of Word Initial Vowels as a Function of Prosodic Structure. *Journal of Phonetics* 24 (November 5, 1996): 423–444.