Enhancement of intonational contrasts in American English

Jennifer Cole[1], Jeremy Steffman[1,2]

[1]Northwestern University, [2]The University of Edinburgh
jennifer.cole1@northwestern.edu, jeremy.steffman@ed.ac.uk

ABSTRACT

We test a subset of intonational contrasts proposed in the Autosegmental-Metrical model for American English for evidence of contrast enhancement in phonologically and phonetically longer vs. shorter intervals. F0 trajectories were assessed from 32 speakers' imitated productions of six tonally distinct tunes, e.g., HHH, HHL. Maximally three tune shapes emerge from clustering analyses of imitated f0 trajectories, each cluster comprising imitations of two phonetically similar but phonologically distinct tunes. We find enhancement of tune contrasts *between* the emergent clusters in measures of f0 differences (RMSD, end f0, center of gravity). There is no evidence of enhancement for phonetically similar tunes grouped *within* the same cluster, though fine-grained phonetic distinctions are detected for these "lost" tune contrasts, suggesting a reanalysis as within-category variation.

Keywords: intonation, contrast, nuclear tune, English

1. INTRODUCTION

The intonation system of English presents a rich variety of pitch patterns in the final region of an intonational phrase. From the rightmost word with 'nuclear' phrasal stress to the end of the phrase, we find monotonically rising or falling patterns that vary in their initial or final f0 values, and more complex patterns with slope that changes e.g., rise-plateau, rise-fall, or rise-fall-rise [1-3] There is a long history of research that relates variation in these nuclear pitch patterns to variation in pragmatic meaning [4, 5], and though there is general agreement on the type of meaning distinctions encoded through intonation in English (e.g., referential alternatives, givenness, epistemic knowledge), there is not yet a consensus about the categorical status of pitch patterns—the number and type of discrete phonological contrasts that are encoded, and their associated meaning functions. This paper investigates the distinctive status of phrase-final pitch patterns in Mainstream American English (MAE) through a phonological lens, to identify distinctions that are phonetically enhanced in phonological contexts that support the full expression of pitch patterns. Informed by work in the segmental domain [6-8], we view enhancement as a phenomenon that serves to increase the perceptual distinctiveness of phonological contrasts. Applying this notion to intonation, we consider the phonetic enhancement of a distinction between two pitch patterns as indicating a representational distinction between phonological categories. This study aims to identify phonological contrasts in the phonetic distinctions among phrase-final pitch patterns, which may inform later work on the discrete and/or gradient meaning functions of intonation.

This investigation focuses on the Autosegmental-Metrical (AM) account of MAE intonation [9, 10], codified in the ToBI annotation system [11], which models phrase-final pitch patterns in terms of three features: the pitch accent marking the word with rightmost ('nuclear') phrasal stress, followed by the phrase accent and boundary tone marking the end of phrasal domains at two levels of prosodic phrasing. These intonational features are specified in terms of the tonal primitives H(igh) and L(ow), and considering only monotonal pitch accents, the system generates eight phonologically distinct "tunes", each comprised of three tone features: HHH, HHL, HLH, HLL, LLL, LLH, LHL, LHH (inclusion of a downstepped High and three bitonal pitch accents further extends the inventory to as many as 24 phonologically distinct nuclear tunes). Each tune generates a distinct pitch trajectory, and while some of these trajectories are robustly distinct (HHH, LLL), others are less so. We focus here on three tune pairs with very similar pitch trajectories that differ only in the final region: {HHH, HHL}, {HLL, HLH} and {LLL, LLH}, as shown in the schematized trajectories of Figure 1, adapted from [10].

Using a tune imitation paradigm, we examine the phonetic implementation of these six tunes to assess the shape and magnitude of f0 distinctions between tune pairs as produced over words that differ in the number of stressed and unstressed syllables in the nuclear region. Prior work with MAE speakers shows poor perceptual discrimination of these tune pairs when presented in 3-syllable words [12], and little or no distinction in the f0 trajectories produced for each pair, again in 3-syllable words [13]. The goal of the present study is to determine whether the predicted contrasts among this set of tunes are enhanced when produced on longer words, and more broadly, how phonetic distinctions in the f0 trajectories of these tunes increase or decrease as a function of metrical

structure—the syllable count and stress pattern in the region of the nuclear tune.

2. METHODS

Stimuli. Participants heard model utterances with resynthesized f0 trajectories representing the six tunes, and imitated the heard tune on each trial, reproducing it in a new sentence. Model utterances were naturally produced by two speakers (one male, one female) in two sentences ("Her name is Marilyn"/ "He answered Jeremy"). The six nuclear tunes were implemented with f0 resynthesis using PSOLA in Praat [14,15], based on straight-line approximations with five target f0 values located in each model speaker's pitch range (Figure 1). The scaling and alignment of resynthesized tunes were based on examples from [10, pp. 391- 401] and online training materials [16].

Participants and procedure. 32 self-reported native speakers of American English were recruited via Prolific and completed the experiment remotely. In a given trial, participants heard the same tune on two stimuli separated by one second. Participants were prompted to reproduce the heard tune on a new target sentence presented orthographically. Target sentences were in one of five metrical conditions, all with initial stress, that varied in the syllable count (1-4) of the final word, with an additional secondary stress in 4b as a potential anchor for the phrase accent [17] (Table 1). There were 120 trials: 6 tunes x 5 conditions x 2 sentences x 2 repetitions. Model gender order (male/female or female/male) and model sentence order were evenly distributed across metrical conditions and tunes and each appeared an equal number of times in the experiment. We measured f0 using STRAIGHT in VoiceSauce [18,19] and computed time-normalized trajectories with 30 samples over each nuclear word. Files containing f0 tracking errors were flagged and removed using an automated algorithm [20]. Differences in the phonetic implementation of tunes were assessed as follows.

Clustering analyses using k-means clustering for longitudinal data [21] were performed on unlabeled

Condition	Sentence
1	He ran with *Moe*
	She lived with *Neil*
2	Her roommate *Nóra*
	His neighbor *Mánny*
3	She gathered *lávender*
	They honored *Mélanie*
4a	He went there *críminally*
	She travelled *mínimally*
4b	They saw the *nóminàtor*
	He was a *lúminàry*

Table 1: Stimuli sentences for the experiment.

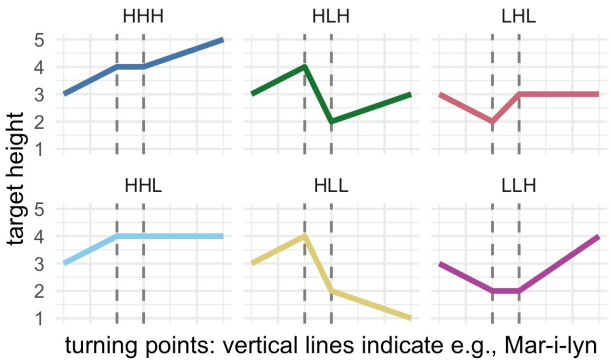

Figure 1: Schema for the model stimuli

imitated f0 trajectories to identify the number and type of distinct trajectories and their relation to the tune labels of the imitated stimuli. We tested clustering solutions with 2-10 clusters, using the Calinski-Harabatz criterion [22] to select the optimal partition of the data as the one with the highest ratio of between- to within-cluster variance. The analysis was carried out on participant mean trajectories—a single mean trajectory for imitations of each of the 6 model tunes, from each participant.

Root-mean squared difference (RMSD). We assessed differences among tunes in phonetic space using RMSD as a measure of the phonetic distance between a pair of f0 trajectories, computed for each speaker over their mean trajectories for each of the six tunes, for all pairwise combinations of tune within metrical condition. Each speaker contributed 75 RMSD values (15 tune pairs by 5 metrical conditions). Pairwise RMSD values were analyzed using mixed-effects Bayesian regression [23]. Metrical condition was modeled as a monotonic effect [24], reflecting our expectation that RMSD would increase with the number of syllables across metrical conditions 1-4. We included random intercepts for speaker, and tune pair, and by-speaker slopes for the fixed effects and interactions. Results are reported using the median posterior estimate and 95% credible intervals (CrI) and the probability of direction metric (pd) [25]. A CrI that excludes zero

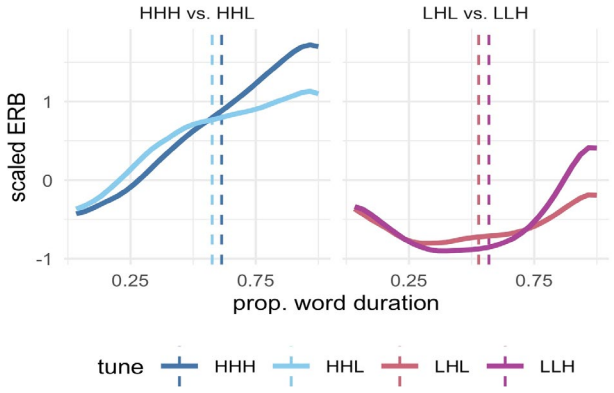

Figure 2: Mean trajectories for four tunes, with mean temporal TCoG, indicated by a dashed vertical line.

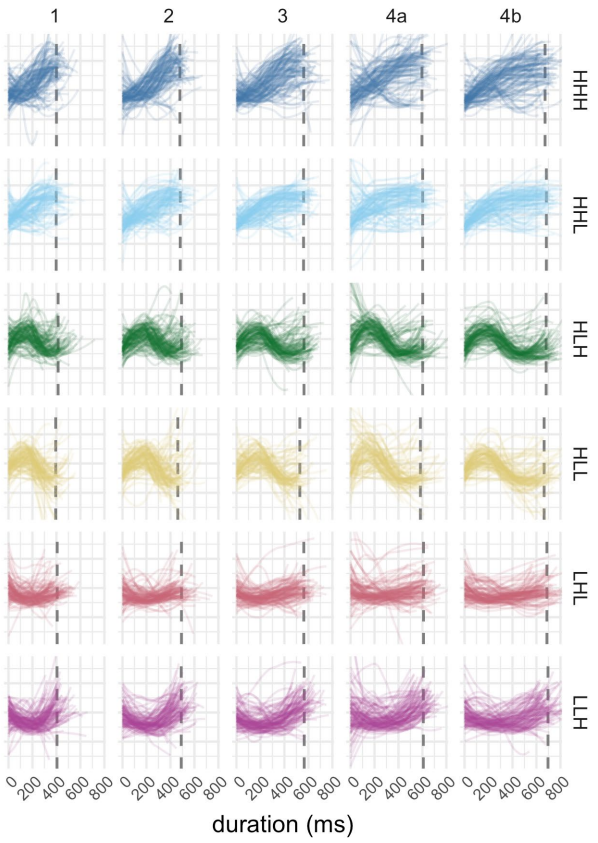

Figure 4: All trajectories for tunes split by metrical structure (columns) and tunes (rows). Vertical line indicates mean duration within each panel.

and pd > 95% are considered as compelling evidence for an effect.

Other parameters. We measured two additional parameters predicted to exhibit enhancement effects: **End f0** and **Tonal Center of Gravity** (TCoG). End f0 was measured in centered ERB values at the end of the nuclear tune interval. For tune pairs of interest, identified based on clustering and described below, we modeled variation in End f0 a s a function of tune (e.g. HHH versus HHL) and metrical condition, and their interaction, with a random intercept for speaker and by-speaker random slopes for the fixed effects and interaction. TCoG is a measure of the location in time of the bulk of the f0 mass over a specified region [26,27]. In our data, TCoG captures subtle differences in the alignment (early/late) and shape (domed/scooped) of f0 rises. We computed TCoG with reference to the end of the first syllable, in milliseconds. Figure 2 shows the mean TCoG for the four tunes we examine using this measure. For tune pairs of interest, defined by the clustering results, we modeled variation in TCoG as a function of tune, metrical structure, and their interaction, with the same random effect structure as for the end f0 model.

Hypotheses and predictions. We hypothesize enhanced contrasts between phonologically distinct tunes across metrical conditions 1-4, manifest in

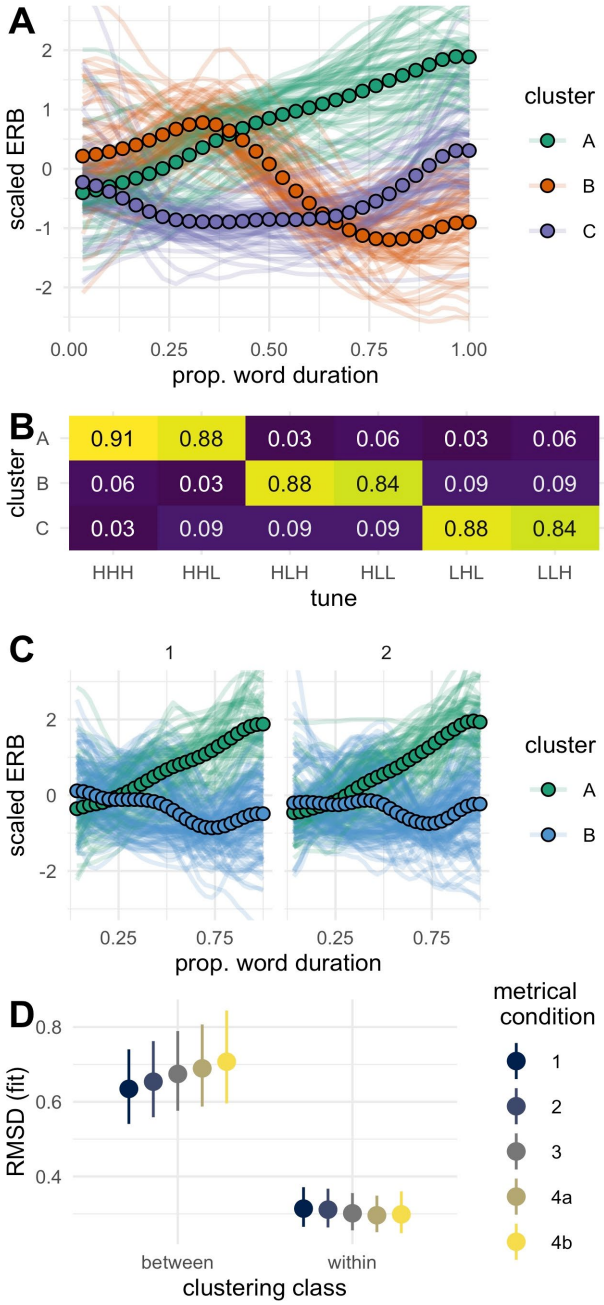

Figure 3: Panel A: Cluster means for three clusters in the partition, Panel B: the proportion of tunes (columns) in each cluster (rows). Panel C: Clustering solutions for 1 (left) and 2 (right) syllable words only. RMSD by cluster class and metrical structure.

enhanced phonetic distinctions with (i) f0 trajectories of imitated tunes that define a greater number of clusters, up to a maximum of 6 clusters in conditions 4a and 4b; (ii) increased RMSD between tune pairs; and (iii) greater differences between tunes in the measures of End f0 and TCoG.

3. RESULTS

Figure 3 shows the raw f0 trajectories from all trials in the experiment, split by metrical condition (columns) and tune (row). As expected, modeling of

the nuclear word duration showed a credible difference across metrical structures (all pairwise pd > 95). Phonetic and phonological lengthening create the potential for contrast enhancement.

Clustering results. The clustering algorithm optimally partitions the data into three clusters, merging imitations of two model tunes. The mean trajectories of the three clusters are shown in Figure 4A, while the mapping of tune to cluster is shown in Figure 4B. Cluster A shows a monotonically rising shape that is composed mainly from imitations of HHH and HHL. Cluster B has a rising-falling shape and comprises imitations of HLH and HLL. Cluster C has a low rising shape and mainly comprises imitations of LHL and LLH. These results suggest poor differentiation of the predicted tune contrasts in each cluster. Additional clustering analyses over subsets of data split by metrical condition yield similar results; conditions 3, 4a and 4b show the same three-way partition of the data as in Figure 4A, while conditions 1 and 2 show only two clusters (Figure 4C): Cluster A with imitations of HHH and HHL, and cluster B with imitations of the remaining four tunes.

RMSD Results. We focus our analysis of RMSD based on results from the clustering analysis, to compare the phonetic distinction between tunes that clustered together with those that clustered apart. Within-cluster pairs are {HHL, HHH}, {HLH, HLL} and {LHL, LLH}; all other pairs are between-cluster (see Figs. 4A,B). This cluster class variable was used as a predictor in the model, interacting with metrical structure. Results show an unsurprising main effect of cluster class: within-cluster tune pairs have a smaller RMSD (β = -0.70, 95%CrI = [-1.01,-0.39], pd = 100). There was no credible main effect of metrical structure, but notably, there was an interaction between clustering class and metrical condition (pd = 99), showing that there is enhancement (larger RMSD across conditions 1-4), *only for tune pairs that are grouped into different clusters,* as shown in Figure 4D: RMSD ascends left to right across conditions, *only* for between-cluster pairs. Pairwise comparisons between metrical conditions confirm credible differences for between-cluster pairs: {1,2} < 3 < {4a,4b} with pd > 95, while within-cluster pairs show only a slightly *smaller* RMSD for 3 versus 2 syllables (pd = 96), an effect opposite to enhancement.

Tune-end f0 results. For each pair of tunes that cluster together, we observe small differences in the End f0 as predicted by the model tunes (see Figure 1): HHL < HHH (β = -0.35, 95%CrI = [-0.50,-0.20], pd = 100), HLL < HLH (β = -0.36, 95%CrI = [-0.54,-0.19], pd = 100), and LHL < LLH (β = 0.27, 95%CrI = [0.08,0.45], pd = 100). Importantly, there was no interaction between tune and metrical structure in any of the three models (pds = 69, 93 and 89 respectively), indicating no evident enhancement effects across metrical conditions. In line with the RMSD results, the End f0 data show a *lack of within-cluster enhancement*. **TCoG results** show a similar pattern. TCoG in HHL is credibly earlier than that in HHH (β = -21, 95% CrI = [-33,-9], pd = 100), and TCoG in LLH is credibly later than LHL (β = 23, 95%CrI = [7,40], pd = 100). See Figure 2 for reference. Crucially however, there was not an interaction between tune and metrical structure (pd for HHH vs. HHL = 55, pd for LHL vs. LLH = 60). In other words, the distinction in TCoG in these tunes did not change systematically across metrical conditions, again showing no within-cluster enhancement.

4. DISCUSSION & CONCLUSION

We tested the hypothesis that phonological contrasts between nuclear tunes are enhanced when produced over longer vs. shorter intervals. Imitated productions of six tunes were assessed for evidence of enhanced contrasts, across five metrical conditions of increasing syllable count and in the presence of an additional, secondary stressed syllable following the nuclear pitch accent. Evidence from the clustering analysis shows maximally three distinct tune shapes, each comprising imitations of two phonetically similar tunes that differ in their tonal specification, with no further increase in the number of tune shapes in longer nuclear intervals. RMSD findings show enhancement of tune contrasts between, but not within, the three emergent tune clusters, with greater RMSD in nuclear intervals of increasing length, but with no extra enhancement in the presence of an additional secondary stress as a potential anchor for the tune-medial phrase accent. Two additional f0 measures, ending f0 and the temporal Tonal Center of Gravity, show expected differences between tunes, but critically, these effects are uniform across metrical conditions, and thus provide no evidence of enhanced contrasts between tunes in the phonetically similar pairs. For the six tunes tested, our findings support an analysis of a three-way phonological contrast in tune shape, with phonetic enhancement in words with longer nuclear intervals. The "lost" tune contrasts involve distinctions in phrase accent and boundary tones: {HHL vs. HHH}, {HLH vs. HLL}, {LHL vs. LLH}. The same tunes are also poorly discriminated in perception [28]. Tunes that cluster together may be better understood as within-category variation. This conclusion calls for reconsideration of categories and gradience in intonational phonology (see [29]), across speech communities and styles.

Acknowledgments: Thanks to Chun Chan and the Prosody & Speech Dynamics Lab at Northwestern U. This project was supported by NSF BCS-1944773.

5. REFERENCES

[1] Cruttenden, A. (1997). *Intonation*. Cambridge University Press.

[2] Palmer, H. E. (1924). *English intonation with systematic exercises*. W. Heffer & Sons Limited.

[3] Pike, K. (1945). *The intonation of American English.* Ann Arbor, MI: University of Michigan Press.

[4] Prieto, P. (2015). Intonational meaning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(4), 371-381.

[5] Westera, M., Goodhue, D., & Gussenhoven, C. (2020). Meanings of Tones and Tunes. In C. Gussenhoven & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 442–453).

[6] Keyser, S. J., & Stevens, K. N. (2006). Enhancement and overlap in the speech chain. *Language*, 33-63.

[7] Stevens, K. N., & Keyser, S. J. (1989). Primary features and their enhancement in consonants. *Language*, 81-106.

[8] Stevens, K. N., & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics*, *38*(1), 10-19.

[9] Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.

[10] Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (Doctoral dissertation, Massachusetts Institute of Technology).

[11] Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. Im S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing*.

[12] Cole, J., Steffman, J., & Tilsen, S. (2022). Shape matters: Machine classification and listeners' perceptual discrimination of American English intonational tunes. *Proceedings of Speech Prosody 2022*, 23-26.

[13] Cole, J., & Steffman, J. The primacy of the rising/non-rising dichotomy in American English intonational tunes. In *Proc. 1st International Conference on Tone and Intonation (TAI)* (pp. 122-126).

[14] Boersma, P. & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.1., retrieved from http://www.praat.org/.

[15] Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, *9*(5-6), 453-467.

[16] Veilleux, N., Shattuck-Hufnagel S. & Brugos A. *6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI.* January IAP 2006. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu.
License: Creative Commons BY-NC-SA.

[17] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2010). Turning points, tonal targets, and the English L-phrase accent. *Language and Cognitive Processes*, 25(7-9), 982-1023.

[18] Kawahara, H., Cheveigné, A. D., Banno, H., Takahashi, T. & Irino, T. (2005). Nearly defect-free f0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Ninth European Conference on Speech Communication and Technology*.

[19] Shue, Y.-L., Keating, P., Vicenik, C., Yu, K. (2011) VoiceSauce: A program for voice analysis. In *Proceedings of ICPhS XVII*, 1846-1849.

[20] Steffman, J., & Cole, J. (2022). An automated method for detecting F 0 measurement jumps based on sample-to-sample differences. *JASA Express Letters*, *2*(11), 115201.

[21] Genolini C., Alacoque, X., Sentenac, M., & Arnaud, C. (2015). kml and kml3d: R Packages to Cluster Longitudinal Data. *Journal of Statistical Software*, 65(4), 1-34.

[22] Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1-27.

[23] Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software*, *80*, 1-28.

[24] Bürkner, P. C., & Charpentier, E. (2020). Modelling monotonic effects of ordinal predictors in Bayesian regression models. *British Journal of Mathematical and Statistical Psychology*, *73*(3), 420-451.

[25] Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *Journal of Open Source Software*, *4*(40), 1541.

[26] Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology*, *3*(2), 337-383.

[27] Barnes, J., Brugos, A., Veilleux, N., & Shattuck-Hufnagel, S. (2021). On (and off) ramps in intonational phonology: Rises, falls, and the Tonal Center of Gravity. *Journal of Phonetics*, *85*, 101020.

[28] Ladd, D. R. (2022). The trouble with ToBI. In Barnes, J., & Shattuck-Hufnagel, S. (Eds.). *Prosodic Theory and Practice*, 247-258. MIT Press.