Testing the distinctiveness of intonational tunes: Evidence from imitative productions in American English

Eleanor Chodroff[1,2] *and Jennifer Cole*[2]

[1]Department of Language and Linguistic Science, University of York
[2]Department of Linguistics, Northwestern University

eleanor.chodroff@york.ac.uk, jennifer.cole@northwestern.edu

Abstract

Understanding the structure of intonational variation is a longstanding issue in prosodic research. A given utterance can be realized with countless intonational contours, and while variation in prosodic meaning is also large, listeners nevertheless converge on relatively consistent form-function mappings. While this suggests the existence of abstract intonational representations, it has been unclear how exactly these are defined. The present study examines the validity of a well-defined set of phonological representations for the generation of intonation in the nuclear region of an intonational phrase in American English: namely, the combination of binary pitch accents (H*/L*), phrase accents (H-/L-), and boundary tones (H%/L%) proposed in Pierrehumbert (1980). In an exploratory study, we examined whether speakers maintained the eight-way distinction among intonational contours posited to exist in this representational system. We created eight synthesized contours according to Pierrehumbert (1980) and examined whether listeners generalized these contours to novel productions. Speakers largely distinguished rising from non-rising contours in production, but few other distinctions were maintained. While this does not rule out the existence of additional contours in production, these findings do suggest that the representation of rising and non-rising contours may be privileged and more readily accessible in the intonational grammar.

Index Terms: speech production, prosody, intonation, nuclear tunes, ToBI

1. Introduction

In English, sentences can be realized with an astonishing variety of intonation patterns. Distinctions in intonation convey distinctions along linguistic dimensions of meaning, including but not limited to illocutionary force, information status (givenness), and focus on semantic alternatives [1], and along extralinguistic dimensions related to speaker affect and emotion [2-3]. Varying the intonational contour can change the meaning of a sentence, though in a highly context-dependent manner. For example, in canonical usage, a globally falling contour indicates a statement. Other meanings of these contours are possible in specific contexts, as in the famously discussed example of the high rising contour in *"My name is Mark Liberman."* as used to address a receptionist in a doctor's office, where it expresses speaker uncertainty (here, about the receptionist's ability to confirm the appointment) [4], or as a sociolinguistic variable as in 'uptalk' [5]

Inferring the set of representations underlying prosodic realization is a well-known and longstanding issue in the area. While segmental research can use lexical distinctions as evidence for deriving abstract representations, such clear-cut linguistic distinctions are much harder to come by in prosodic research. Xu [6] refers to this as the 'lack of reference problem' for prosodic analyses. A complete account of intonation involves at least two parts: identifying the inventory of perceptually and meaningfully distinct intonational contours, and specifying the meaning contrasts associated with those contours. As an argument for an approach that starts with the intonational form, utterances with similar meaning may not have similar intonational patterns, so establishing equivalence classes based on meaning similarity could be misleading. Rather, assigning any two intonational contours to the same phonological category (i.e., sharing the same tonal specification) requires similarity in form (conditioned on phonological context), together with similarity in the contribution of the contour to utterance meaning [7].

Pierrehumbert [7–8; further developed in 9] proposes a concrete theory of intonation that focuses on phonetic form and variation: a set of intonational features define local pitch targets on words within a prosodic phrase and combine to define a set of phonologically distinct phrasal pitch melodies (p. 29, ex. 14). Here we focus on the obligatory intonation features in the intonational phrase (IP) in that system: the sequence of pitch accent (specifically, H* or L*) followed by a phrase accent (H-/L-) and boundary tone (H%/L%), which we refer to as the "nuclear tune". Bitonal pitch accents (e.g., L*+H) and (prenuclear) pitch accents that optionally occur earlier in the prosodic phrase are set aside for the purpose of our study. Allowing for minor contextual modifications due to local prosodic and segmental context, this sequence of high- and low-tone pitch accents, phrase accents, and boundary tones should predict a set of eight phonetically distinct pitch melodies that characterize the final region of the intonational phrase, and which are available for encoding linguistic meaning. These pitch melodies, schematized in straight-line f0 approximations following Pierrehumbert [7] (pp. 391-401), are illustrated in Figure 1. More generally, this set of intonational features with additional bitonal pitch accents, forms the basis for the well-known ToBI system in wide use throughout the prosodic literature [10].

In an exploratory study, we examined whether speakers maintained the eight-way distinction among nuclear intonational contours posited to exist in this representational system. We created eight synthesized contours according to Pierrehumbert [7] (pp. 391–401) and the ToBI straight–line approximations in the MIT OpenCourseWare course "Transcribing Prosodic Structure of Spoken Utterances with ToBI" [11] and examined whether on hearing these contours, listeners would generalize them in subsequent production of novel sentences (see also [12] for imitation of prosodic contours). Speakers were told that the stimuli were computer-

generated speech, and asked to produce natural versions of the intonational contour. Speakers largely distinguished rising from non-rising contours in production, but few other distinctions were observed. While this does not rule out the existence of additional contours in a speaker's inventory of nuclear tunes, these findings suggest that the representation of the distinction between rising and non-rising contours may be privileged and more readily accessible in the intonational grammar than the proposed distinctions within the classes of rising and non-rising contours.

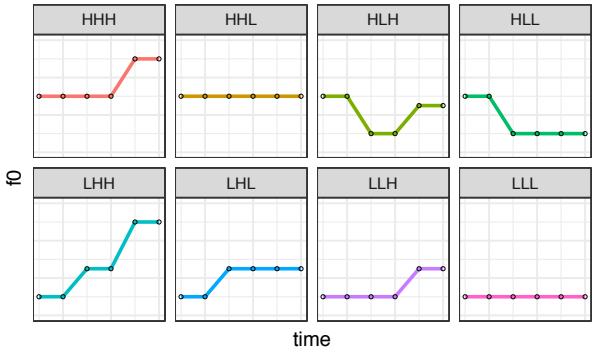

Figure 1: *Simple nuclear tune templates based on patterns presented in Pierrehumbert [7] and the ToBI straight-line approximations [11].*

2. Methods

2.1. Participants

32 native speakers of American English were recruited from the Northwestern University undergraduate community for the experiment (19 female, 13 male). Ages ranged from 18 to 25 years old. An additional 10 participants completed the experiment, but were non-native speakers of English. As such, their data were not retained for analysis.

2.2. Stimuli

The eight nuclear tune contours were synthesized using a custom Praat script [13] according to the nuclear tune templates provided in Pierrehumbert [7] and *the ToBI straight-line approximations in the MIT OpenCourseWare course "Transcribing Prosodic Structure of Spoken Utterances with ToBI"* [11]. The templates indicate four critical pitch points across all simple tunes: a high f0, mid-high f0, mid-low f0, and low f0. To derive these points, two native speakers of American English, one female and one male, first recorded all eight nuclear tunes naturally on three model sentences, designed to have a simple syntactic structure and end in a final, three-syllable, stress-initial proper name with no medial voiceless obstruents. The model sentences were: "She quoted Helena", "Her name is Marilyn", and "He answered Jeremy." The nuclear tune was always produced on the final noun phrase, and no pitch accents were produced on the earlier preamble portion. The four critical pitch points were based on the natural ranges calculated from the original productions. We first identified values (in Hz) for the female speaker and then ensured that the ratio between the f0 points in ERB was matched for the male speaker, but in his natural f0 range. The interquartile range and summary statistics for each speaker's f0 is provided in Table 1, and the selected f0 pitch points based on this distribution are shown in Table 2. Because the model speakers produced all eight nuclear tunes, we observed a wide range of f0 values, including some that contained falsetto voice.

The resynthesis involved setting the pitch at six points across a baseline sentence. The base files were selected from the natural productions for each speaker. We limited the selection to the flat intonational contours. For the female speaker, we used the natural H*H-L% productions and for the male speaker, the L*L-L% productions. The preamble region was first divided into thirds. The pitch started 20 Hz above the preamble end value, dropped 10 Hz at the second point, and reached the target value at the end of the preamble. The pre-specified nuclear tune was then overlaid on the nuclear region. The three points were placed after the start of the nuclear region 25% of the way through, 40% of the way through, and at the end of the phrase. The motivation for using 40% of the duration instead of the midpoint was to account for the slight phrase-final lengthening and to target the second syllable of the word; this also resulted in a more natural pitch contour. The shapes of the resulting contours can be seen in Figure 2. All stimuli were normalized to 70 dB for presentation.

Table 1: *Observed speaker f0 range (Hz | ERB) in natural productions of the eight nuclear tunes.*

Measure	Female Speaker		Male Speaker	
	Hz	ERB	Hz	ERB
Mean	224	6.27	124	3.97
Minimum	85	2.90	87	2.95
1st Quartile	176	5.24	107	3.52
Median	212	6.02	120	3.87
3rd Quartile	239	6.57	129	4.10
Maximum	647	12.36	279	7.33

Table 2: *Synthesized f0 pitch points (Hz | ERB) for each model speaker.*

Synthesized Pitch Point	Female Speaker		Male Speaker	
	Hz	ERB	Hz	ERB
Preamble	200	5.77	107	3.52
High	400	9.30	262	7.01
Mid-high	225	6.29	125	4.00
Mid-low	200	5.77	107	3.52
Low	175	5.21	100	3.32

Three target sentences were created for participants to produce in the same manner as the presented synthesized stimuli. The structure of these sentences paralleled that of the model sentences with a short preamble and a final, stress-initial, three-syllable proper name without medial voiceless obstruents. The sentences were "He modeled Harmony", "They honored Melanie", and "She remained with Madelyn".

2.3. Procedure

Participants were informed that the goal of the experiment was to improve the naturalness of computer-generated speech. Each trial consisted of an auditory presentation of three resynthesized sentence stimuli in the male and female voices, as described above, all with the same intonation, which participants were told were samples of computer-generated speech. Participants were asked to listen to the melody of the sentences in a given trial, and produce a new sentence with the same melodic pattern, but "said the way you think it should sound if it were

spoken by a human English speaker, in a manner that is familiar to you."

Each trial in the experiment consisted of an auditory and visual presentation of the three model sentences with the same nuclear tune, each separated by one second, followed by a visual presentation of the target sentence. In each trial, participants were prompted to reproduce the intonation of the stimuli on a novel sentence, with the reminder presented above the target sentence: "I would say it this way". Stimuli derived from both the male and female speakers were presented on each trial, with speaker order of the model sentences randomized. All six speaker orders (FFM, FMF, MFF, MMF, MFM, FMM) were paired with all eight nuclear tunes (HHH, HHL, HLH, HLL, LHH, LHL, LLH, LLL) and all three target sentences an equal number of times throughout the experiment, resulting in 144 trials. The order of the trials was then fully randomized, and a break was offered every 24 trials. One target sentence was produced per trial.

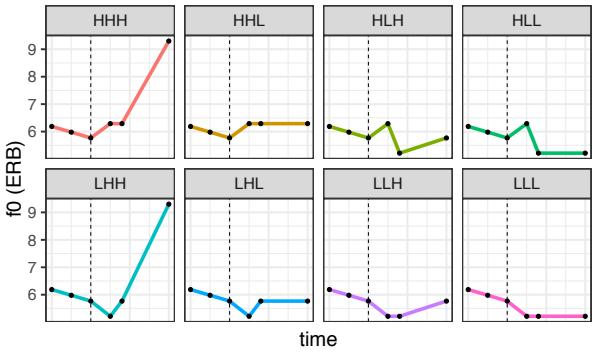

Figure 2: *Synthesized nuclear tunes for the female speaker. The preamble region occurs to the left of the dashed vertical line and has been compressed for space. The nuclear region is plotted according to the normalized word duration (pitch points at 25%, 40% and 100%).*

3. Results

As participants were asked to reproduce the exposure stimuli with a similar, but natural intonational contour, we first examined how much speakers deviated in their productions from the intonation patterns of the resynthesized stimuli. A primary goal of the experiment was to have participants access representations of intonation patterns that are hypothesized to be part of the American English intonation system, and which are therefore predicted to be familiar. A high degree of deviation could reflect any of the following: 1) the participant failed to identify the intonation pattern of the stimuli as a familiar intonation pattern, resulting in various correction strategies, 2) the stimulus pattern was recognized, but difficult for the speaker to reproduce precisely, or 3) the stimulus pattern was recognized but the nuclear tune category it represents tolerates high variability. In any case, the experiment provides a baseline understanding of how speakers generalize from their perception of the prescribed nuclear tunes to novel sentence production, and also gives us a glimpse of what kinds of modifications speakers make to bring a perceived melody in line with familiar melodies.

In addition to exploring the degree of deviation, we also assessed whether speakers could maintain an eight-way distinction among contours, and if not, how many distinct clusters and what type of clusters speakers produced. The cluster analysis critically differs from the deviation analysis, as it does not depend on any notion of accuracy in replicating the original stimuli. The cluster analysis instead ignores the exposure stimuli while assessing variability and systematicity in the set of speaker-generated productions.

3.1. Deviation analysis

To examine deviation in production, we extracted a 30-point time-normalized f0 contour (ERB) from the nuclear region of the model speakers' and participants' productions using ProsodyPro [14]. The f0 window for female speakers was set between 75 and 600 Hz and for the male speakers between 50 and 300 Hz. Each contour was centered on the speaker's mean for comparison, and the root-mean-square error (RMSE) between target production and final exposure contour was derived for each trial. The averaged contours by participant and tune are shown in Figure 3 next to the average exposure contours, and the mean RMSE across participants and tunes is reported in Table 3. Qualitatively, speakers deviated substantially more in the rising tunes (H-H%) relative to the non-rising tunes. In particular, speakers generally imitated the drop in f0 below their mean for LHH but often produced this same contour for HHH, instead of producing the H* pitch accent above their mean f0. In addition, speakers were quite variable in their realizations of HLL. This contour, commonly associated with declarative statements, may permit greater variation in realization relative to other contours; speakers may also have shifted the H* realization to the intended preamble region.

Table 3: *Rank-ordered mean root-mean-square error (RMSE) per tune between the time-normalized intonational contour and final exposure tune within each trial.*

Tune	RMSE	Tune	RMSE
1. HHH	0.903	5. LLH	0.435
2. LHH	0.786	6. HHL	0.429
3. HLL	0.482	7. LLL	0.427
4. HLH	0.457	8. LHL	0.367

A linear mixed-effects model was used to quantitatively analyze the by-trial RMSE with tune, gender of the model speaker presented last in the trial (gender), and their interactions as fixed effects, as well as a random intercept for participant and word [15–16]. More complex random effect structures failed to converge. As observed in the qualitative assessment, significantly greater deviation from the exposure contours was observed following the rising contours (β_{HHH} = 0.38, β_{LHH} = 0.26, $ps < 0.001$) and significantly less deviation was observed following all non-rising contours (β_{HHL} = -0.11, β_{HLH} = -0.08 β_{HLL} = -0.06, β_{LLH} = 0.11, β_{LHL} = -0.17, $ps < 0.001$). The dynamic nature of the rising contours could have made these difficult to reproduce precisely. An additional linear mixed-effects model was used to analyze just the non-rising tunes to determine whether deviation differed from the average among this particular set. Relative to the average deviation, post-HLL productions deviated significantly more than average (β_{HLL} = 0.05, $p < 0.01$), while post-LHL productions deviated significantly less (β_{LHL} = -0.06, $p < 0.001$). In both models, there was significantly greater deviation from the final exposure contour produced by the female model speaker than the male model speaker (model 1: β_F = 0.10, model 2: β_F = 0.12 $ps < 0.001$).

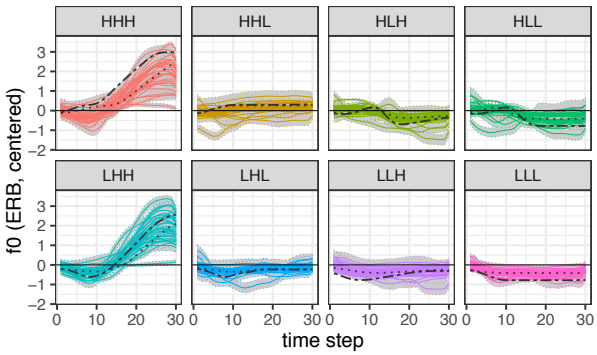

Figure 3: *Time-normalized averaged f0 (ERB) contours per participant and nuclear tune, centered on the speaker mean. The gray bands correspond to ± 1 standard error of the mean. The solid black line at y = 0 corresponds to the mean for each speaker, the dashed line reflects the averaged female model speaker contour and the dotted line the averaged male model speaker contour.*

3.2. Cluster analysis

The 30-point time-normalized f0 (ERB) contours from the nuclear region of each participant-produced utterance were again used in the cluster analysis. To examine the dispersion of the eight exposure nuclear tunes, we first conducted a k-means cluster analysis for longitudinal data (KML cluster analysis) on the mean f0 contours calculated for each participant and exposure tune [17]. The optimal number of clusters was determined using the Calinski-Harabasz criterion and interpretations of each cluster were provided after inspection of their contents.

The cluster analysis on the averaged f0 contours for each tune and participant yielded two optimal clusters: rising tunes and flat tunes. Post-HHH and post-LHH productions, the canonical rising contours, were predominantly classified as cluster A (HHH: 30/32, LHH: 31/32), whereas the remaining tunes fell into cluster B with the exception of one post-HLL which was classified as cluster A. Approximately 75% of the averaged f0 contours were grouped into a single cluster, which roughly approximated a 'flat' tune; however, there could be fine-grained distinctions within that cluster. When the productions following the canonical rising tunes HHH and LHH were removed, the algorithm yielded an additional four clusters (Table 4). The separation of tunes among these four clusters indicated a cluster for post-HHL productions (ending in a mid-level f0), one for post-HLH and post-HLL productions (falling tunes), and two clusters for productions following low pitch accents.

While the optimal number of clusters was determined to be two, we additionally examined the partitioning of the by-speaker averaged contours when eight clusters were assigned (assuming one for each intended nuclear tune). In this case, the contours did not split evenly by exposure tune. Instead, two clusters were used to account for the rising tunes, but with each cluster split between post-HHH and post-LHH productions. Post-HHL productions predominantly formed one cluster, post-HLH and post-HLL formed another cluster. Post-LHL productions also largely formed one cluster along with decent representation from post-LLH productions. There was also a cluster driven by post-LLL productions but that included some participants' average post-LLH and post-LHL productions.

Table 4: *Results of KML cluster analysis on f0 contours averaged by participant and exposure tune, excluding productions following HHH and LHH exposure tunes. Tunes are roughly ordered by cluster divisions.*

Tune	A	B	C	D
HHL	28	0	3	1
HLH	2	24	4	2
HLL	1	25	3	3
LLH	1	2	21	8
LHL	1	0	29	2
LLL	0	2	12	18

4. Discussion & Conclusion

The present study pursued a form-based analysis of prosodic realization by examining the extent to which speakers could generalize a subset of the nuclear tunes posited in Pierrehumbert [7]. This particular framing of prosodic theory suggests that prosodic *form* is just as critical for understanding the prosodic system as prosodic *function.* While previous studies have often addressed the form-function relationship in prosody, we have taken seriously the proposal to study prosodic form independent of its function (see also [18]). We found that speakers did not maintain the proposed eight-way distinction in the intonational contour of the nuclear region. It may be that the resynthesized f0 contours did not adequately tap into the intended representations, or that listeners were unable to access an appropriate meaning which hindered their ability to reproduce the target intonation. Some tunes were nevertheless more readily accessible for production than others: in particular, strong evidence was found for a clear distinction between rising and non-rising tunes. This may in part reflect the salient illocutionary distinction in American English, but it also reflects the cross-linguistic tendency to make use of this prosodic contrast [19–21]. Among the non-rising tunes, the deviation analysis revealed high variability in HLL productions, commonly associated with declarative utterances, and relatively low deviation in LHL productions. The cluster analysis on the non-rising tunes indicated some marginal distinctions between HHL, commonly associated with list intonation, high and falling tunes (HLH and HLL), and low tunes (LLH, LHL, and LLL).

These findings do not discount the existence of all eight (or more) nuclear tunes, but they nevertheless provide important insight into the structure and accessibility of intonational forms in American English. The study shows promise for exploring prosodic representation through imitation and generalization. A follow-up experiment is currently underway which focuses on distinctions in imitative productions following the non-rising tunes alone. The large f0 contrast present in the rising contours may have obscured more fine-grained contrasts in imitative production. More generally, these findings offer a fruitful starting point for positing novel hypotheses regarding the inventory of intonational contours and prosodic form.

5. Acknowledgements

We thank Stefanie Shattuck-Hufnagel and Alejna Brugos for helpful discussion, questions and commentary, as well as members of the Prosody and Speech Dynamics Lab at Northwestern University.

6. References

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

[2] T. Bänziger and K.R. Scherer. The role of intonation in emotional expressions. *Speech Communication*, *46*(3–4), 252–267, 2005.

[3] D. Grandjean, T. Bänziger, and K.R. Scherer. Intonation as an interface between language and affect. *Progress in Brain Research*, *156*(06), 235–247, 2006.

[4] C. Bartels. *The intonation of English statements and questions: a compositional interpretation.* New York & London: Garland Publishing, 2006.

[5] P. Warren. *Uptalk: The phenomenon of rising intonation*. Cambridge University Press, 2016.

[6] Y. Xu. Speech prosody: A methodological review. *Journal of Speech Sciences*, *1*(1), 85–115, 2011.

[7] J.B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD Dissertation. MIT, 1980.

[8] J.B. Pierrehumbert. Tonal elements and their alignment. In M. Horne (Ed.), *Prosody: Theory and Experiment* (pp. 11–36). Kluwer Academic Publishers, The Netherlands, 2000.

[9] M.E. Beckman and J.B. Pierrehumbert. Intonational structure in Japanese andEnglish. Phonology Yearbook 3, 255–310, 1986.

[10] M.E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel. "The Original ToBi System and the Evolution of the ToBi Framework" in S.-A. Jun (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 1–37). Oxford University Press, 2005.

[11] N. Veilleux, S. Shattuck-Hufnagel, and A. Brugos. *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, 2006.

[12] J.B. Pierrehumbert and S.A. Steele. Categories of tonal alignment in English. *Phonetica*, Vol. 46, pp. 181–196, 1989.

[13] P. Boersma and D. Weenink. Praat: Doing phonetics by computer (Version 6.0.19), 2016.

[14] Y. Xu. "ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis" in *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)* (pp. 7–10). Aix-en-Provence, France, 2013.

[15] D. Bates, M. Maechler, B. Bolker, and S. Walker, "lme4: Linear mixed-effects models using Eigen and S4," *R package*, version 1, no. 7, pp. 1-23, 2014.

[16] A. Kuznetsova, P.B. Brockhoff, and R.H.B. Christensen, "Package 'lmerTest'," *R package*, version 2.0, 2015.

[17] C. Genolini, X. Alacoque, M. Sentenac, and C. Arnaud. kml and kml3d: R packages to cluster longitudinal data. Journal of Statistical Software, 65(4), 2015.

[18] C. Gussenhoven and A.C.M. Rietveld. An experimental evaluation of two nuclear-tone taxonomies. *Linguistics*, *29*(3), 423–450, 1991.

[19] D. Bolinger. Intonation across languages. In Greenberg, J. H., Ferguson, C. A., and Moravcik, E. A. (eds.) Universals of Human Language, Phonology, 2: 471–524, 1978.

[20] A. Cruttenden. Falls and rises: Meanings and universals. Journal of Linguistics, 17(1), 77–91, 1981.

[21] C. Gussenhoven and A. Chen. Universal and language-specific effects in the perception of question intonation. In Proceedings of the 6th International Conference on Spoken Language and Processing (ICSLP). Beijing, 2000.

[22] H. Hirst and A. Di Cristo. A survey of intonation systems. In D. Hirst & A. Di Cristo (Eds.), Intonation Systems: A Survey of Twenty Languages. Cambridge: Cambridge University Press, 1998.