Evidence and Intonational Contours: An Experimental Approach to Meaning in Intonation

Byron Ahn[1], *Stefanie Shattuck-Hufnagel*[2], *Nanette Veilleux*[3]

[1]Princeton University, USA
[2]Massachusetts Institute of Technology, USA
[3]Simmons College, USA

bta@princeton.edu, sshuf@mit.edu, veilleux@simmons.edu

Abstract

One proposed function of prosody is conveying speakers' stances on the evidence for their statements and assessments of listeners' beliefs. Testing this is challenging – specifying evidential status is difficult, and speakers vary intonationally for a given context. A novel task of reading lines from comic strips elicits relatively consistent intonation, suggesting both the method's usefulness, and the efficacy of speaker beliefs in governing prosodic contours. Preliminary results suggest H* accents are used when speakers believe they have evidence the hearer lacks, and L* accents for the flipped situation.

Index Terms: speech prosody, intonational meaning, pragmatics, semantics, experimental method

1. Introduction and Background

It is well-established that in languages like English, the meaning of a spoken utterance is signalled by the morpho-syntax (i.e., the segmentally-specified words/morphemes and syntactic structure of the sentence), and by the prosody (i.e., the grouping and prominence patterns conveyed by, e.g., the suprasegmental intonational tones/contours). The difficulty of designing rigorous experiments and defining appropriate units has hindered our understanding of the contribution of intonation to meaning. This paper presents an initial attempt at investigating the connection between formal models of meaning (semantics and pragmatics) and formal models of prosody, with the ultimate goal of creating a model that correctly predicts the appropriate prosodic contour(s) for any appropriately defined linguistic context. In particular, we test the hypothesis that the speaker's beliefs about the status of the evidence about a declaration influences the choice of intonational contour for the statement.

While past research by prosodic theorists have attempted to link elements of the prosodic inventory with meaning, most have found that their claims are limited and have been difficult to generalize. Intonationists have hypothesized mapping accent type to meanings on the basis of mutual beliefs (e.g., [1], [2]), but subsequent experimental testing has shown that these initial hypotheses require further elaboration. For example, the Low-High accent L+H* has been claimed to occur on words which explicitly contrast to an alternative. However, the distribution of L+H* accents does not fully support this prediction (e.g., [3]). We hypothesize this problem arises in part from the specificity and precision of the hypotheses. That is, accents have meanings, but their meanings have been described narrowly and incompletely. Further, we propose that advances in pragmatic and semantic theory, describing the aspects of meaning that contribute to discourse structure and how those dynamics influence the particular forms of sentences, can illuminate these issues. For example, some languages use "discourse particles" or "evidential markers" —segmental morphemes or words that indicate particular discourse structures—raising the question of how other languages (e.g. English) represent this conversational information. Recent research suggests that some of these discourse dynamics (i.e., who has what information) may be expressed intonationally in English (e.g., [4], [5]).

While the semantics/pragmatics literature suggests that intonation can signal important information ([6]), these works also tend to focus on a narrow set of prosodic features (e.g., edge tones). This approach has uncovered important findings, but has not explored potentially important generalizations about what types of meaning are carried by which intonational elements (pitch accent types and combinations of pitch accents with edge tones, i.e., 'tunes'). While some broader work on intonational meaning has been carried out ([1], [7], overview in [6]), many claims are yet to be rigorously tested experimentally, and few investigations have been framed in terms of contemporary theories of evidentiality or discourse pragmatics.

This project develops a new experimental paradigm, in which variables identified by recent advances in pragmatic theories are manipulated experimentally, building on earlier work to address the question of how meaning relates to intonation. Specifically, we bridge the conversational space between speakers and their interlocutors with an elicitation experiment with conditions defined in terms of sourcehood and evidentiality (cf. [4], [5]). Using this paradigm, we have found promising systematic results in prosodic behavior that suggest the semantic/pragmatic factors we manipulate play a role in determining the speaker's choice of intonational contour.

This approach will be useful for future work investigating more abstract questions about precisely how intonation carries particular meanings (e.g., tones vs. tunes; cf. [7], [1]). In addition, our results raise issues about the representational nature of discourse information in languages (like English) where such information is marked via intonation. As noted above, in many other languages, these pragmatic meanings are represented with segmental words/morphemes; is such information also syntactically represented in English-type languages? A standard view is that any pairing of phonetic/phonological information with semantic/pragmatic meaning is mediated by syntax ([8]:p.1), so that semantics/pragmatics cannot influence phonetics/phonology directly. If intonational meaning shares this characteristic, one would expect *syntactic* difference between sentences like "Are you tired↑ or sad↓?" (L* H- H* L-L%) and "Are you tired↑ or sad↑?" (L* H- L* H-H%), as in, e.g., [9], [10], [11]. In this way, our finding that the interpretive significance of intonation may track sourcehood/evidentiality is consistent with the view that (some) intonational tones/contours may be abstractly represented in the syntactic structure (also argued by, e.g., [12]). If this hypothesis is confirmed more

broadly, it would mean that intonational work is not immune to a recurrent finding in linguistics: adequate work in any one subfield requires deep understanding of all the others.

2. Methodology

Production experiments are notoriously difficult to control. In the authors' experience (and in subject self-report) speakers often adopt unintended roles or simply begin to read prompts without any particular communicative intent. For this reason, laboratory speech prompts and perception studies, rather than production experiments, predominate in the literature, or physiological methods like eye-trackers are used to ascertain the perceptual impact. Clearly this poses difficulties for researchers intending to map prosodic production onto function [13]. In this work, we devise a novel production task using comic strips to more directly explore this mapping.

We aim to explore the role of evidence present to both the Speaker and Hearer in guiding the development of the conversational 'common ground', as reflected in the intonational choices of the Speaker. Our chosen framework predicts that the choice between different prosodic contours will be influenced by who the Speaker believes holds evidence for a proposition and in the strength of this evidence. In particular, we conclude that the edge tones will vary in direction (rising for questions, falling for confirmations) and pitch accents will vary in type (High, Low or bitonal combinations such as H*, L*, L+H* in MAE_ToBI notation; [14]) when subjects/speakers respond to conditions that vary the evidence that (the speaker believes) the hearer has.

In other words, this research seeks to experimentally determine whether there is a link between degrees of evidence and shared beliefs, on the one hand, and the implementation of the pitch accent and boundary tone, on the other. Manipulating the variables identified by Gunlogson and Northrup [4, 5], we have devised five basic semantic/pragmatic contexts for a spoken interaction between S (the speaker) and H (the hearer), where each condition controls what S and H know to be true and what S reasonably believes about what H believes to be true. Using controlled narrative conditions, we manipulate what evidence the speaker (S) believes the hearer (H) has about whether a proposition is true (e.g., H is a reliable source, H has direct, indirect, no or contradictory evidence; see Table 1). The five conditions create a partially ordered list, ranging from expected confirmation of shared beliefs (A) through declaring new information (D), to contradiction (E).

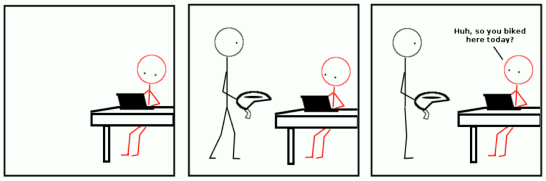
The novel methodology developed to address these issues uses dialogs to represent each of the five conditions (A-E in Table 1) for two different scenarios (raining and biking). The circumstances illustrated in 3-panel comic strips (See Figure 1) described the specific context. Ten subjects were recruited from a sample of convenience from the adult native English speaking population at the home institution of one of the authors. Subjects were instructed to read the comic strip silently, associating themselves with the character in the last panel (drawn in red) and then read that character's dialog aloud. Each of two scenarios, with prompts in all of the five contexts, were repeated a total of 3 times in non-consecutive but otherwise random order, for a total of 300 recorded responses. The analysis presented here covers Conditions A, B, and D (180 responses).

Our hypothesis is that the use of boundary (final) tones will be consistent with published accounts ([4]). As for pitch accents, literature does not (as far as we know) make any specific predictions about the relationship between pitch accents and evidentiality/sourcehood. We therefore cautiously hypothesize that the pitch accent of target words (e.g., "rain" or "bike" in the two scenarios) may depend on condition, as well.

Table 1: *Example of five evidentiality/sourcehood-based contexts and the relationship of each to the status of the Speaker's beliefs about the Hearer's belief concerning the proposition.*

Condition	Description of Evidence/Sourcehood
A	S asserts a proposition P, deferring to H as a reliable source of information, believing that H has direct evidence that P is true.
B	S asserts a proposition P, believing that H also has direct evidence that P is true.
C	S asserts a proposition P, believing that H has indirect evidence suggesting that P is true.
D	S asserts a proposition P, believing that H has no evidence about P's truth.
E	S asserts a proposition P, believing that H has a indirect evidence that contradicts P.

Figure 1: *An example of one prompt for Scenario A. The dialog bubble reads "Huh, so you biked here today?"*



3. Results

The productions of ten participants were labeled by three experienced annotators (the authors) using MAE_ToBI (e.g., [14]). Agreement between at least 2 of the 3 labels for a given utterance was taken as the correct label; where three different labels had been generated, consensus labels were determined through discussion. This resolved issues for all but one speaker, in whose productions the target word was deaccented at roughly twice the rate of other subjects. That speaker's responses were judged to be outside of the dialect described by MAE_ToBI, and their responses were removed from the analysis. Another 15 of the 162 responses where the target word was deaccented were also discarded (This was more frequent in the "bike" scenarios where the personal pronouns 'I' or 'she' competed with the verb for accentuation). Our Phase I analysis focused on contexts A, B and D for the remaining 9 subjects.

Notably, in the 147 recordings analyzed in this Phase I analysis, labellers agreed upon labels for pitch accents in 137 tokens (93%), and labels for boundary tones in 140 tokens (95%). When considering the entire contour, labellers agreed upon labels for both pitch accents and boundary tones in 131 tokens (89%). Moreover, results of this Phase I analysis reveal strong trends in production, indicating that the manipulated contextual variables (determined by discourse-structural notions of sourcehood/evidentiality) are closely related to the meaningful contributions of intonational tones/contours.

Examining pitch accents first, Tables 3 and 4 reveal strong correlations between discourse-context condition and pitch ac-

Table 2: *Example of five contexts and a specific scenario to elicit prosodic productions.*

Condition (from Table 1)	**Description of Scenario** (Biking version)
A: S asserts a proposition P deferring to H as a reliable source of information, believing that H has direct evidence that P is true.	H walks into the room, carrying a bike helmet. S says to H "Huh, so you biked here today?"
B: S asserts a proposition P, believing that H has direct evidence that P is true.	Persons X, S and H are co-workers. Person X rides by S and H on a bicycle. S says to H "Huh, so she biked here today?"
C: S asserts a proposition P, believing that H has indirect evidence suggesting that P is true.	S and H are co-workers. H sees S walking in with a bike helmet. Later, when talking about getting home, S says to H "I biked here today, you know"
D: S asserts a proposition P, believing that H has no evidence about P's truth.	S and H are co-workers. H asks S how they got to work. S says "I biked here today."
E: S asserts a proposition P, believing that H has a belief that contradicts P.	S and H are co-workers. S has a cast on her leg and S says to H "Guess what! I biked here today."

cent choice. In both Conditions A and B, speakers used L-type (including L* and L*+H) accents most frequently (A: 37/46, 80%; B: 32/49, 65%), and in Condition D, speakers always used H-type (including H*, !H*, and all bitonal variants) accents.

Examining the contours in more detail, results that take boundary tone into account reveal that the preferred contour for A (30/46, 65%) uses an L* H-H% contour, the classic contour for a Yes-No question in American English; at the same time, the preferred contour for D (30/52, 58%) is H* L-L%, the classic contour for a neutral declarative statement. These results are given in Tables 5 and 6. (Note that in Table 6, the labels 'H*' and 'L*' do not include bitonal variants.)

Table 3: *Distribution of specific L and H-type accents across speakers for Conditions A, B, and D. Shading indicates the most common accent for the condition.*

Pitch Accent	**A**	**B**	**D**
H*	4	7	**35**
L+H*	5	10	5
!H*	0	0	1
H+!H*	0	0	11
L*	**32**	**29**	0
L*+H	5	3	0

Speakers also produced B with a preference for a L* H-H% contour (28/49, 57%) but this preference was less pronounced.

Table 4: *Distribution of L-type and H-type accents across speakers for Conditions A, B, and D. Note the progression of a greater prevalence of H-type accents as S's belief about H's evidence for the proposition increases. Shading indicates the most common accent-type for the condition.*

Pitch Accent type	**A**	**B**	**D**
H-type	9	17	**52**
L-type	**37**	**32**	0

Table 5: *Distribution of contours with both L and H-type accents for four boundary tones (!H merged with H in all cases), across speakers for Conditions A, B, & D. Shading indicates the most common contour for the condition.*

Pitch Accent type	**Boundary Tone**	**A**	**B**	**D**
H-type	L-L%	9	15	**43**
	H-L%		1	6
	L-H%			3
	H-H%		1	
L-type	L-L%		1	
	H-L%	4	1	
	L-H%			
	H-H%	**33**	**30**	

Table 6: *Distribution of contours across speakers for Conditions A, B, & D, organized by boundary tone. 'H*' and 'L*' exclude bitonals. Shading indicates most common contour.*

Boundary Tone	**Pitch Accent**	**A**	**B**	**D**
L-L%	H*	4	7	**30**
	Other pitch accents	5	9	13
H-H%	L*	**30**	**28**	0
	Other pitch accents	3	3	0
Other boundary tones	Any pitch accent	4	2	9

That is, for the B condition, H-type accents appeared in about 35% of tokens (17/49), with 59% of those (10/17) appearing as L+H*. Although the data should be considered preliminary, this suggests an evolution with respect to the alignment along a scale of certainty about S's beliefs about H's evidence for the proposition at hand. Moreover, it is our impression that this shared ToBI label for conditions A and B hides a reliable difference in scaling; further analysis at the acoustic level will be required to test this hypothesis.

These results are remarkable in their high level of consistency for a given context (A, B or D) across speakers and (more notably) *within* speakers across repetitions (which were not elicited consecutively). Individual speakers produced identical pitch accents and boundary tones across the entire triad of repetitions (the 3 non-consecutive repetitions of the prompt, for an individual speaker) 19 out of 54 times (2 scenarios, 3 con-

texts, 9 speakers). Accents were identical in 24 triads – even when counting, e.g., !H* and H* as distinct; collapsing L-type and H-type accents raises this within-speaker consistency to 36 of 54 triads. Boundary tones were also produced with a high degree of consistency for entire triads; speakers produced identical boundary tones within a triad in 35 of 54 cases.

In sum, each context (defined by manipulating evidentiality-based variables) elicited consistent intonational contours from individuals. This intra-speaker consistency is valuable because a given speaker, with a consistent understanding of both the evidence and the speaker's/hearer's relationship with that evidence, produces a consistent intonational contour under these experimental conditions. (Across speakers, there may be different ways of understanding the pragmatic context, allowing for the observed limited variation in intonation.) These findings support earlier proposals from formal pragmatic research that evidence strength and reliability variables impact a speaker's intonational choices ([4], [5]). Further, we find that these variables can also impact pitch accent type/alignment.

4. Conclusions

This analysis of three of the evidentiality/sourcehood conditions (contexts A, B, and D) suggests that speakers use intonational contours to signal degrees of evidence the speaker has about the facts, and, perhaps more importantly, about the hearer's beliefs and evidence about the facts. Preliminary evidence suggests that speakers use H-type accents when the speaker believes the hearer has no evidence (D), and L-type accents when the speaker believes the hearer has direct evidence (A). However, when the speaker and hearer transparently share the evidence (B), responses more mixed: speakers use L* predominantly (as in A), but some subjects choose H* instead.

A useful outcome of this experiment is the demonstration that the method used to prompt subjects was effective for eliciting consistent intonational contours. Although not every (non-consecutive) repetition for each subject was identical to other repetitions for that prompt, there was a high degree of within-speaker consistency, with only small amounts of variation, and a noticeable degree of agreement across speakers. Moreover, speakers' prosody changed depending on the condition and they did not simply repeat a favorite intonational contour for all prompts. This suggests that speakers understand the context and then produce what, for them, is the appropriate prosody for each of these conditions, with evidentiality/sourcehood as one of the conditioning factors.

Although this report involves only 10 subjects over three conditions, the results are promising for eliciting consistent results within speaker and condition. Moreover, since the stimulus manipulations were between contexts defined in terms of evidentiality and sourcehood, it finds evidence for a reliable connection between formal aspects of the semantic/pragmatic context on the one hand and specific intonational tones/contours on the other. More intuitive notions like sentence function (e.g., "question" vs. "statement") are too coarse-grained to be useful in predicting the intonational contour; in particular, such notions could not be used to predict that context A and context B differ in the proportion of H* accents. (More broadly, there is no universal "question" or "statement" intonation; consider, e.g., the difference between WH-questions and Yes/No questions.)

However, one should be cautious in concluding that evidentiality/sourcehood are the primary factors at play in our results. For example, it should be noted that both A and B stimuli were punctuated with question marks, unlike those in D which were punctuated with periods. This may have influenced speakers to use certain contours in A/B differently from D. However, this does not nullify our conclusions; question marks do not always go with L* H-H% contours (e.g., WH questions and polar alternative questions), and additionally the intonational differences between A and B cannot be due to punctuation. To address this issue, future experiments will not include punctuation.

To further investigate this connection, we are in the process of increasing the size of the subject pool and labeling the remaining conditions (C and E). Additionally, we plan to investigate whether the naturalness of these contours depends on evidentiality/sourcehood in a perceptual experiment, using these cartoon prompts with audio recordings. Finally, further methods of describing the distinct contours will be investigated. For example, labellers had difficulty with labels that are less frequently encountered in laboratory speech (e.g., !H-L% vs. H-L%). Other methods for categorizing accents, such as the Tonal Center of Gravity, might shed more light on both alignment as well as scaling differences ([15]). Finally, we would like to highlight that this research will have an impact on further development of speech interface technologies. As automatic conversational agents become more widespread, users will begin to expect a nearly human experience. Without a clear understanding about what level of meaning/function is reflected by which set of prosodic categories, and which prosodic categories may map to a particular underlying meaning, developing algorithms to provide such an experience is impossible.

5. References

[1] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, 1990.

[2] C. Gussenhoven, *On the Grammar and Semantics of Sentence Accent*. Dordrecht: Foris, 1983.

[3] D. Watson, M. Tanenhaus, and C. Gunlogson, "Interpreting pitch accents in online comprehension: H* vs. L+H*," *Cognitive Science*, vol. 32, pp. 1232–1244, 2008.

[4] C. Gunlogson, "A question of commitment," *Belgian Journal of Linguistics*, vol. 22, pp. 101–136, 2008.

[5] O. Northrup, "Grounds for commitment," Ph.D. dissertation, UC Santa Cruz, 2014.

[6] J. Hirschberg, "Pragmatics and intonation," in *The Handbook of Pragmatics*, 2004.

[7] C. Bartels, *The intonation of English statements and questions: A compositional interpretation*, 2014.

[8] E. Selkirk, *Phonology and syntax: the relationship between sound and structure*, 1984.

[9] K. Pruitt and F. Roelofsen, "The interpretation of prosody in disjunctive questions," *Linguistic Inquiry*, vol. 44, no. 4, 2013.

[10] D. Farkas and F. Roelofsen, "Polarity particle responses as a window onto the interpretation of questions and assertions," *Language*, vol. 91, no. 2, pp. 359–414, 2015.

[11] L. Winans, "Disjunction and alternatives in Egyptian Arabic," 2015, ms., UCLA.

[12] E. O. Aboh, "Information structuring begins with the numeration," *Iberia*, vol. 2, no. 1, pp. 12–42, 2010.

[13] G. Garding and A. Arvaniti, "Dialectal variation in the rising accents of American English," *Laboratory Phonology 9*, 2004.

[14] M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original ToBI system and the evolution of the ToBI framework," in *Prosodic Typology*, S.-A. Jun, Ed., 2005.

[15] 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," *LabPhon 3*, 2012.