CREU 2016-2017 Final Report: Intonation and Evidence

Dr. Nanette Veilleux, Simmons College
Karina Bercan, Simmons College
Emily Chicklis, Simmons College
Sara Harland, Simmons College

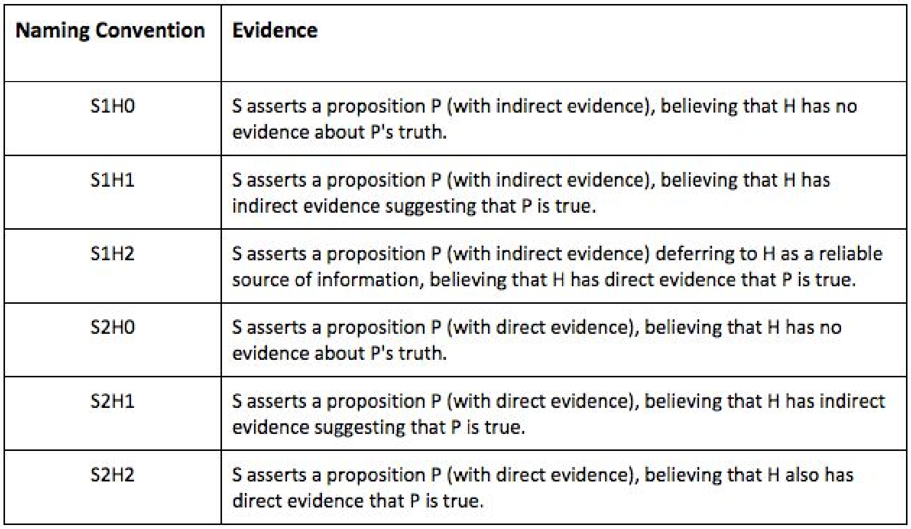
I. Goals and Purpose

This research project in Computer Science and Linguistics seeks to improve automatic speech understanding by identifying the appropriate prosody in different situations. In addition to expressing paralinguistic information, such as emotional state or gender, we propose that the prosody (or intonational contour) of an utterance conveys the speaker's certainty regarding their own beliefs, as well as their assessment of what the listener knows or believes to be true. We captured prosodic contours through production experiments designed to capture human speech in context. These findings will be used to augment automatic speech synthesis and the output will be tested for naturalness and ease of understanding.

II. Related Work

Prosodic contours are comprised of pitch accents and boundary tones [5]. A pitch accent is the variation in pitch that lends salience to a particular word or morpheme, and a boundary tone is the rise or fall at the end of a statement. These prosodic elements are categorical in nature, although they may be implemented on an acoustic continuum [4]

Past experiments have shown that systematic variation of context produces consistent boundary tones [2,3] and that the notion of "evidence" plays a predictive role. When a speaker makes a proposition, they either have indirect or direct evidence of its truth, while the hearer may have indirect evidence, direct evidence, or none at all [1]. Figure 1 outlines the permutations of evidence states for both interlocutors.



Naming Convention	Evidence
S1H0	S asserts a proposition P (with indirect evidence), believing that H has no evidence about P's truth.
S1H1	S asserts a proposition P (with indirect evidence), believing that H has indirect evidence suggesting that P is true.
S1H2	S asserts a proposition P (with indirect evidence) deferring to H as a reliable source of information, believing that H has direct evidence that P is true.
S2H0	S asserts a proposition P (with direct evidence), believing that H has no evidence about P's truth.
S2H1	S asserts a proposition P (with direct evidence), believing that H has indirect evidence suggesting that P is true.
S2H2	S asserts a proposition P (with direct evidence), believing that H also has direct evidence that P is true.

Figure 1: The six naming conventions used in the experiment, applied to one of two scenarios

Our study examines the conditions S1H2 (Speaker has only enough evidence to form the utterance, believing that the Hearer has direct evidence), where a an interrogative (H-H%)

boundary tone is expected and compares the intonational productions with those in the S2H2 condition (both speaker and hearer have direct evidence), where declarative (L-L%) boundary tone is expected.

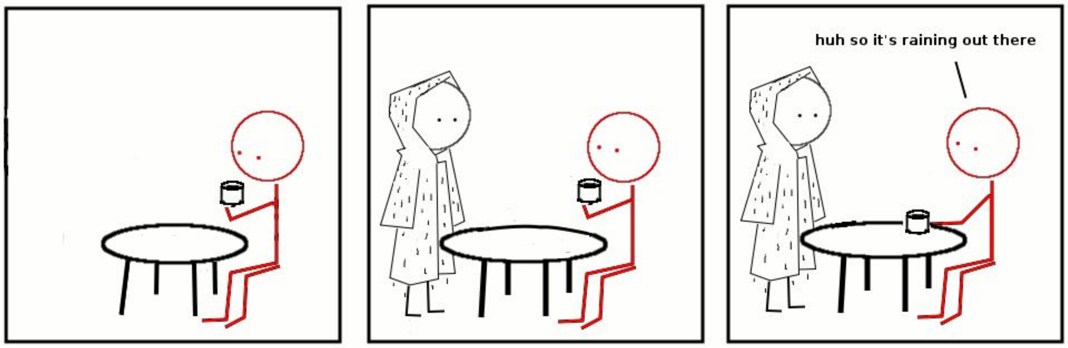
III. Process

This research used production experiments to gather data to analyze differences in prosody in situations where a speaker has direct evidence for their proposition and where a speaker has indirect evidence for their proposition. The experiment presents variations of the six evidential scenarios presented in Figure 1, with comic strips providing explicit evidence for the certainty or uncertainty of a given speaker's statement. College-aged subjects were asked to read the comics aloud, and sound file data was annotated using the standardized ToBi system of transcription.

Subjects were recruited from Simmons College to read aloud certain sentences in thirty-six comic strips. Each subject was set up with one of three versions of a slideshow presentation featuring the thirty-six comics in randomized order (three different orders, each order being one of the three versions of the slideshow). There are twelve distinct comics--two distinct scenarios (a copier/printer being out of order and it being rainy) presented once in the context of each evidential scenario as described in Figure 1. These twelve comics are displayed three times in a random order. The subject was instructed to take their time reading and understanding the context of each comic strip and to speak aloud the last sentence in the comic as though they were the character drawn in red. The subject's speech was recorded, and they completed the task alone in a quiet room, controlling the pace of the slideshow progression on their own.

The comics intentionally omit punctuation so as not to sway the subjects unfairly towards one intonation over another. However, this approach raises the possibility of subjects "defaulting" to a declarative boundary tone (L-L%), as though the propositions end with periods. To examine this issue, an associate expanded on the existing comics by adding thought bubbles to clarify the red character's perspective in the situation.

Upon collecting and organizing the audio data, we labeled the prosody of each audio file using the MaeToBI+ transcription system. Finally, we analyzed the data to uncover trends or indications.



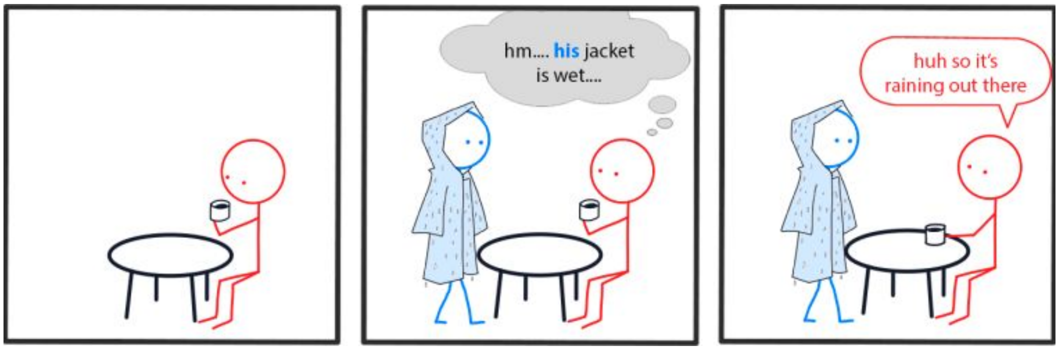


Figure 2: At top, an example of an S1H2 condition used in our first experiment; at bottom, the edited version of that same comic used in a second experiment. Note the thought bubble for clarification.

IV. Results and Discussion

In the initial round of experiments, using the set of comics without thought bubbles to add clarification, the omittance of punctuation created enough ambiguity that subjects varied from expected boundary tones. When the speaker in the comic had only indirect evidence of a proposition for which the hearer had direct evidence (S1H2), subjects had some tendency to default to the L-L% (declarative) boundary tone rather than the expected H-H% (interrogative) boundary tone. When both speaker and hearer had direct evidence, there were significant instances of "upspeak" (L-H% boundary tone), inviting further comment from the hearer.

The second round of experiments used the set of altered comics with thought bubbles for clarification, in addition to more detailed instructions. The new directions specify that although the comics do not feature punctuation, the subject should use whatever punctuation they feel is appropriate in the context of the strip. Figures 3 and 4 compare the boundary tones of the participants from both rounds or versions of the experiment. Both experiments elicited a variety of tones; however, the altered version of the experiment with thought bubbles resulted in a significant drop in the use of an L-H% boundary tone, which is considered "upspeak".

Limitations of this study primarily concern the subject pool of the production experiments. All of the participants are college-aged females, ranging from 18 to 22 years old, and an overwhelming majority are cisgender. This raises a question about how much gender and age influenced participants speech prosody habits and thus influenced the results of the production experiments. Additionally, it is possible that some participants rushed through the comics and read them aloud without fully considering the context and the situation before speaking. Finally, texting language common today may affect one's grammar and prosody, and in this experiment, potentially played a role in the participants' intonation.

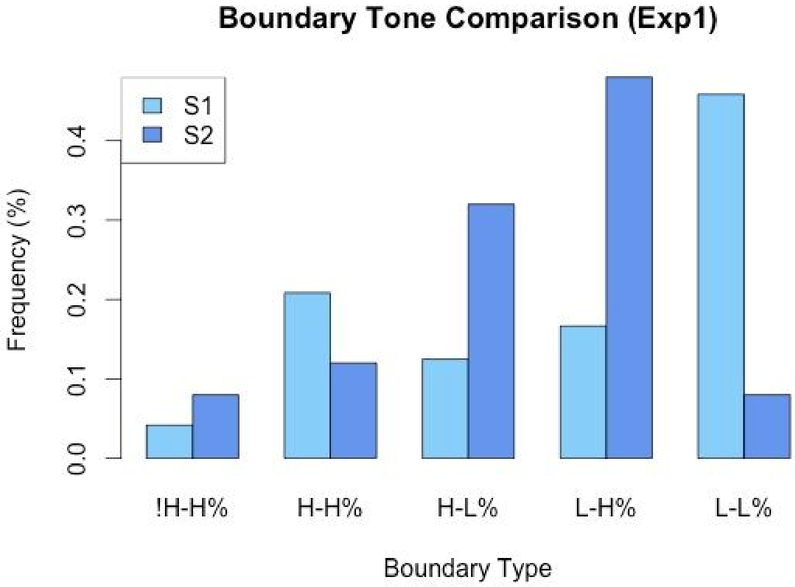


Figure 3: A summary of our preliminary results that compares the frequency of boundary tones between the S1H2 and S2H2 conditions. Note the variety of tones elicited.

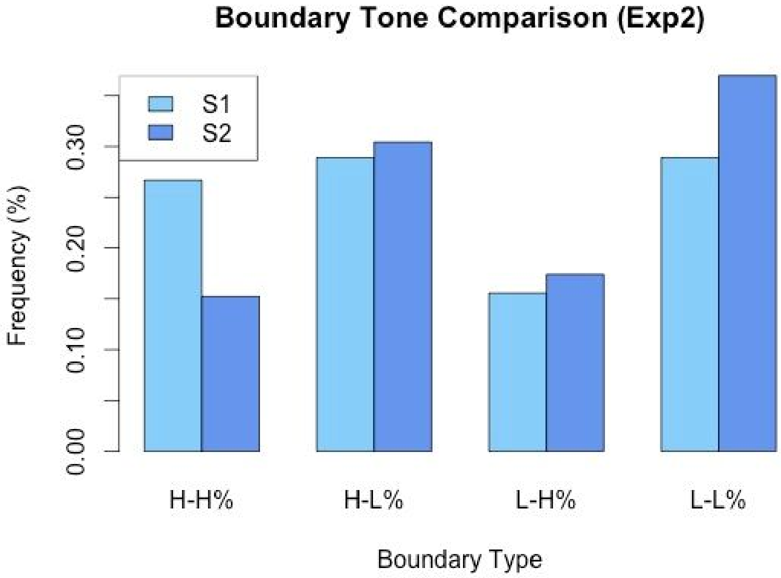


Figure 4: A summary of our second round of results using edited comics (with thought bubbles for clarification). Note the sharp decrease in instances of upspeak among the S2 files.

V. Future Work

The next steps of this research into prosody's relationship with evidential scenarios include further analyzing the collected data from the experiments described in this paper and designing and implementing perception experiments. There remains audio data to annotate and to compare, and more statistical work can be done on the data that was collected. Primarily, comparing results of the labeled data by region where participants grew up, by musical fluency, by foreign language skills, and other factors may contribute to mapping relationships among certain prosodic contours and various situations and contexts. Moreover, the results should be analyzed for significance. Finally, the fully-analyzed results from the production experiments will contribute to the design and implementation of perception experiments, where subjects will match artificially produced or altered speech with a comic featuring the matching context.

Other future work includes expanding further on the production experiment described in section III. It would be advantageous to include male participants in production experiments to identify potential differences in prosody of males and females. Additionally, prosody may differ across age groups, so it would be beneficial to include subjects older than college-aged and subjects who may not be influenced by texting habits, such as implying a period at the end of a sentence instead of actually including it.

VI. Web Links

Project Description and Blog:
http://anita.simmons.edu/~creu/IntonationAndEvidence/index.html

VII. Presentations and Publications

Poster Presentations:

1) New England Celebration of Women in Computing (NECWIC) Conference
2) Simmons College Undergraduate Symposium

VIII. References

1. Ahn, B., Shattuck-Hufnagel, S., Veilleux, N. (2016) Evidence and Intonational Contours: An Experimental Approach to Meaning in Intonation. Proceedings of the Speech Science and Technology, 16th Australasian International Conference, Parramatta, Australia, December, 2016
2. Hirschberg, J., 2004. Pragmatics and intonation. In The Handbook of Pragmatics.
3. Gunlogson, C., 2008. A Question of Commitment. Belgian Journal of Linguistics, 22, 101–136.
4. Barnes, J., Veilleux, N., Brugos, A. and Shattuck-Hufnagel, S. (2010). "Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology." *Laboratory Phonology*.
5. Veilleux, N., Shattuck-Hufnagel, S., Brugos, A., ToBI for Prosodic Transcription of American English, MIT Opencourseware 6.911, 2006.