interspeech 2008 incorporating SST 08

# Prosodic Manifestations of Confidence and Uncertainty in Spoken Language

*Heather Pon-Barry*

School of Engineering and Applied Sciences, Harvard University, USA

ponbarry@eecs.harvard.edu

## Abstract

We present a project aimed at understanding the acoustic and prosodic correlates of confidence and uncertainty in spoken language. We elicited speech produced under varying levels of certainty and performed perceptual and statistical analyses on the speech data to determine which prosodic features (e.g., pitch, energy, timing) are associated with a speaker's level of certainty and where these prosodic manifestations occur relative to the location of the word or phrase that the speaker is confident or uncertain about. Our findings suggest that prosodic manifestations of confidence and uncertainty occur both in the local region that causes the uncertainty as well as in its surrounding context.

**Index Terms**: prosody, human speech perception, emotion detection, paralinguistic cues

## 1. Introduction

Prosody is a fundamental part of human-to-human spoken communication. For example, it has been repeatedly shown that prosody can affect the semantic interpretation of an utterance [1]. Prosody is also used to convey the emotional state of the speaker. In recent years, there has been work on the automatic detection of emotions such as annoyance and frustration [2] and on distinguishing positive and negative emotional states [3, 4].

In this paper, we address the problem of detecting confidence and uncertainty in spoken language. Specifically, we examine how prosodic features correlated with confidence and uncertainty are manifested relative to the source of uncertainty. Because existing speech corpora are not adequate for comparing pairs of utterances that are lexically identical but differ in their level of certainty, we present a methodology for eliciting such utterances. Our approach enables us to analyze both the prosodic features associated with perceived levels of certainty and the location of these prosodic cues relative to the word or phrase that is the source of the uncertainty. We view this as an initial step towards answering the following question, after classifying an utterance as certain or uncertain, can we determine which part of the utterance the speaker is confident or uncertain about?

This work goes beyond existing research in two ways. First, we present a novel method for eliciting lexically identical phrases uttered under varying levels of certainty that allows us to evaluate regularities across speakers as well as for one particular speaker. Our method also facilitates the analysis of how perceived levels of certainty differ from a speaker's actual level of certainty. Secondly, we identify prosodic cues associated with perceived level of certainty and compare their relative strengths in (a) the word or phrase causing the confidence or uncertainty and (b) the surrounding context.

This paper is organized as follows: Section 2 discusses past work on characterizing the prosody of uncertainty and on automatically detecting emotions in speech. Our approach to eliciting uncertain speech and to annotating it is outlined in Section 3. Section 4 presents the results of the perceptual labeling, an analysis of which prosodic features are correlated with a speaker's level of certainty, and comparisons of where these prosodic manifestations occur relative to the source of uncertainty. We discuss the implications of our results in Section 5, and Section 6 outlines multiple directions for future work.

## 2. Previous Work

The topic of how uncertainty is manifested in speech has been examined in the psycholinguistics community. In a setting where an experimenter asked participants general-knowledge questions, the participants produced hedges, filled pauses, and rising intonation contours when they had a lower 'feeling-of-knowing' [5]. A follow-up study demonstrated that listeners were sensitive to these lexical and prosodic phenomena [6]. These findings suggest that prosody is one channel through which speakers convey their level of certainty, but they do not tell us whether level of certainty can be detected solely through prosodic cues.

In the area of emotion detection, researchers have examined certainty in spoken language using data from tutorial dialogue systems [7]. In this work, they train a classifier on prosodic and contextual features to distinguish certain, uncertain, and neutral utterances. They achieved 76% accuracy, compared to a baseline of 66% accuracy (where the baseline was to choose the most common class). These results suggest that speakers' level of certainty is indeed reflected in their prosody.

Past studies have investigated whether adapting to a speaker's level of certainty in spoken tutorial dialogue systems has a positive impact on learning [8, 9]. The results have been promising but inconclusive. Adapting to a speaker's level of certainty may prove more beneficial if we had the ability to hone in on the source of the uncertainty. Our work extends the existing work by collecting a more controlled corpus in order to address the question of whether we can determine which part of an utterance a speaker is confident or uncertain about.

## 3. Method

### 3.1. Eliciting Speech

We collected speech data in two domains: (1) answering questions about using public transportation in Boston, and (2) choosing appropriate words to complete partial sentences. In both domains, we gave participants a written sentence containing one or more gaps with multiple options for filling in the gap. We then asked them to read the sentence aloud with the gap filled in according to domain-specific criteria.

The decision to collect read (i.e., non-spontaneous) as opposed to spontaneous speech stemmed from piloting this data collection experiment with both read and spontaneous answers.


Accepted after peer review of full paper

We found that all but one of the acoustic features that were significantly correlated with perceived level of certainty in spontaneous speech were significantly correlated in the read speech as well. The method of collecting read rather than spontaneous utterances allows us to control the number of words per utterance and to collect multiple instances of the same word spoken under varying degrees of certainty, both between speakers and for one particular speaker.

#### *3.1.1. Participants*

Twenty members of the Harvard community participated in the data collection, 14 females and 6 males. All participants were native-English speakers.

#### *3.1.2. Transit Materials and Procedure*

We collected 10 transit question responses from each participant. The questions varied in difficulty. A question of medium difficulty is shown below.

Question: How can I get from Harvard to Faneuil Hall on the T?

Answer: Take the red line to the _____

a. green line
b. orange line

and get off at _____ .

c. Haymarket
d. Government Center

We use the term 'context' to refer to the fixed part of the response (*Take the red line to the* _____ *and get off at* _____, in this example) and the term 'target words' to refer to the options for filling in the gaps.

For each gap, we presented two possible target words for filling it in. There were eight noun phrases that occurred as a target word in multiple answers (e.g., *Government Center* appeared four times, *green line* appeared six times). This was done deliberately in order to elicit the same phrase uttered multiple times by a speaker under different levels of certainty.

Each item was presented in the following manner. First, the experimenter asked the question aloud. We instructed participants to respond as if they were talking to another person who had recently moved to the Boston area. Next, participants saw the 'context' sentence with gaps but they did not see the target words (i.e., the options for filling in the gaps). They could look at the partial sentence for as long as they wanted to. Upon a key press, they moved to the next screen where target words were displayed below the context and they began reading the sentence aloud when they heard a beep. The beep occurred 1500 ms after the target words were displayed. Finally, participants rated their own level of certainty on a 1-5 scale for the answer they gave (1 = very uncertain, 5 = very certain). This procedure is summarized below. Participants pressed a key to move from one step of the procedure to the next.

*Procedure for each item:*

1. Experimenter asks the question.
2. Participant sees context (i.e., sentence with gaps), target words are not shown.
3. Participant sees context plus target words. After 1500 ms a beep is played and participant reads sentence aloud.
4. Participant rates their level of certainty.

Participants completed two practice items to get a feel for the procedure before beginning the main body of items.

We also collected 10 'neutral' sentences in the transit domain that participants read aloud. The neutral sentences were designed to be similar to the question responses in number of syllables while being lexically distinct from any of the possible question responses. All of the target words occurred in at least one neutral sentence.

The order of presentation for the items was balanced across subjects. The neutral transit items were split into two parts, A and B. Within each part, the order of the items was randomized. Half the participants completed part A before the question responses and part B after the question responses. The other half of the participants completed part B before the question responses and part A after.

#### *3.1.3. Vocabulary Materials and Procedure*

We collected 20 vocabulary utterances from each participant. An example vocabulary item is shown below.

Only the _____ workers in the office laughed at all the manager's bad jokes.

a. pugnacious
b. craven
c. sycophantic
d. spoffish

For each gap, we presented four possible target words for filling it in. Participants were instructed to choose the word that best completed the sentence. To elicit words uttered multiple times by a particular speaker, the options for each blank were drawn from a pool of only 13 words. To facilitate varied levels of certainty, 3 of the 13 words were extremely infrequent words (e.g., *spoffish*) and 5 of the 20 vocabulary items offered four options of which none fit well in the context.

The procedure for the vocabulary items was identical to steps (2)-(4) in the procedure for the transit items. Again, the order of presentation for the items was balanced across subjects.

### 3.2. Annotating Speech

#### *3.2.1. Participants*

Five members of the Harvard community rated the recorded utterances for perceived level of certainty. All annotators were native-English speakers.

#### *3.2.2. Materials*

Each annotator rated all 600 utterances: 200 transit question responses and 400 vocabulary sentences.

#### *3.2.3. Procedure*

Utterances were presented to the annotators in a random order in twelve sections, each containing 50 utterances. We did not let the annotators see any contextual information (i.e., the questions, the options for filling the gaps, the instructions given to the speakers). We did tell the annotators that the speakers were

given a sentence containing one or more gaps, multiple options for filling in the gap, and some criteria for how to fill in the gap. We instructed the annotators to rate how certain the speaker sounded regardless of how sensible the resulting sentence was.

### 3.3. Prosodic Analysis

We extracted the following features from each utterance. All features were represented as z-scores normalized by speaker.

- Pitch (f0) features: minimum, maximum, mean, standard deviation, range, relative position in utterance of minimum pitch, relative position in utterance of maximum pitch, absolute slope
- Intensity (RMS) features: minimum, maximum, mean, standard deviation, relative position in utterance of minimum intensity, relative position in utterance of maximum intensity
- Temporal features: total silence, percent silence, total duration, speaking duration (total duration minus pauses), speaking rate

### 3.4. Context and Target Word Utterances

To assess whether prosodic cues of confidence and uncertainty occur within the target words versus in the surrounding contexts, we created 'context' and 'target word' utterances by manually removing the target words from the original recorded utterances. Pauses preceding the target word were considered part of the target word and were removed along with the target word. Because participants had unlimited time to read over the context before seeing the target words, we consider the target word region to be the *source* of the speaker's confidence or uncertainty; it corresponds to the decision that the speaker had to make.

## 4. Results

Assessments by participants of their own level of certainty were distributed over all 5 categories and had a mean of 2.61 (1 = very uncertain, 5 = very certain). The annotator's ratings of perceived certainty had means of 3.22, 3.35, 3.60, 3.82, and 3.30. Figure 1 shows how the distribution of self-ratings is heavily concentrated on the uncertain side whereas the annotators' ratings are more heavily concentrated on the certain side.

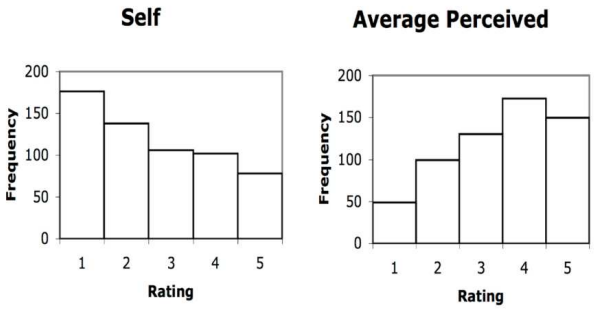


Figure 1: *Self-reported and perceived levels of certainty*

Inter-rater agreement was calculated using the Kappa statistic. The average Kappa score between annotators was 0.284. To compare our annotations with those of Liscombe et al. [7], we recoded our annotations such that uncertain = 1, neutral = 2 or 3, and certain = 4 or 5. With this recoding, the average Kappa score between annotators was 0.45 (other possible recodings resulted in average Kappa scores between 0.41-0.43), which is on par with the 0.52 average Kappa score reported by Liscombe et al. as well as scores reported in other studies of emotion detection [3].

For the following results, we use the term 'average rating' to refer to the average among the five annotators only. We intend to investigate the difference between self-assessed level of certainty and perceived level of certainty in future work, but it is beyond the scope of this paper.

Correlations between average perceived rating and prosodic features extracted from the whole utterance, from the context, and from the target word are shown in Table 1. These correlations are based on prosodic features extracted from 300 utterances drawn from both transit and vocabulary items. This set contains 15 utterances from each of the 20 participants.

## 5. Discussion

From our annotation of the perceived level of certainty, we found that the average perceived level was higher than the speaker's self-reported level for 67% of the 600 utterances. This result is of interest to researchers in emotion detection, particularly those studying uncertainty, because classifiers trained to detect and respond to perceived levels may be overlooking many instances of actual uncertainty.

The correlations between average rating and prosodic features extracted from whole utterances suggest that temporal features (i.e., total silence, percent silence, total duration, speaking duration) are the features most strongly associated with the perceived level of certainty. Other features, including absolute slope f0, range f0, and speaking rate, had statistically significant but smaller correlations with the average rating. It is likely that the perceived level of certainty is associated with a combination of these features (especially in light of past results [7]).

The comparison between prosodic features extracted from the whole utterance, from the context, and from the target word (see Table 1) has implications both for classifying the level of certainty of an utterance and for identifying the word or words that a speaker is confident or uncertain about.

First, we discuss classifying the level of certainty of an utterance. We observed that some features, such as absolute slope f0, have stronger correlations in the whole utterance than in the context or target word. For features behaving in this way (absolute slope f0, total silence, total duration, speaking duration, speaking rate), separating the context from the target word does not provide any additional information. More interestingly, we observed that features such as range f0 have stronger correlations in the context than in the whole utterance or the target word. This suggests that as a cue to uncertainty, the range f0 feature is manifested most strongly in the context. If the word or phrase causing uncertainty is known ahead of time, features behaving in this way (range f0, min f0, max f0, stdev f0, min RMS) could be computed for the context region rather than the whole utterance to improve classification accuracy.

When the word or phrase causing uncertainty is not known ahead of time, features such as percent silence might be useful in determining the source of the uncertainty. We observed that the percent silence feature had a much stronger correlation in the target word than in the context. This suggests that, as a cue to uncertainty, the percent silence feature is manifested most strongly in the target region. The opposite holds for features such as speaking duration, range f0, and min RMS. That is, these features had much stronger correlations in the context than

Table 1: *Correlations between mean perceived rating and prosodic features for whole utterances, contexts, and target words, N=300 (note: * indicates significant at $p < 0.05$; ** indicates significant at $p < 0.01$)*

| Feature | Whole Utterance | Context | Target Word |
|---|---|---|---|
| min f0 | 0.074 | $0.176^{**}$ | $-0.022$ |
| max f0 | $-0.102$ | $-0.166^{**}$ | $-0.025$ |
| mean f0 | $-0.039$ | 0.080 | $-0.048$ |
| stdev f0 | $-0.078$ | $-0.147^{*}$ | 0.011 |
| range f0 | $-0.136^{*}$ | $-0.247^{**}$ | 0.005 |
| rel. position min f0 | 0.002 | $-0.010$ | 0.099 |
| rel. position max f0 | 0.073 | 0.056 | 0.079 |
| absolute slope f0 | $0.312^{**}$ | $0.226^{**}$ | $0.171^{**}$ |
| min RMS | 0.085 | $0.216^{**}$ | $-0.007$ |
| max RMS | $-0.076$ | 0.068 | $-0.028$ |
| mean RMS | 0.008 | 0.090 | $-0.052$ |
| stdev RMS | $-0.015$ | 0.010 | 0.001 |
| rel. position min RMS | 0.039 | $-0.108$ | $0.135^{*}$ |
| rel. position max RMS | $-0.085$ | $-0.083$ | $-0.035$ |
| total silence | $-0.644^{**}$ | $-0.497^{**}$ | $-0.525^{**}$ |
| percent silence | $-0.459^{**}$ | $-0.198^{**}$ | $-0.568^{**}$ |
| total duration | -0.653** | -0.568** | -0.600** |
| speaking duration | $-0.515^{**}$ | $-0.480^{**}$ | $-0.281^{**}$ |
| speaking rate | $0.134^{*}$ | 0.088 | 0.089 |

in the target word. By considering all possible target words and comparing context and target word correlations for these features, we may be able to determine the word or phrase causing the speaker's confidence or uncertainty.

## 6. Conclusions and Future Work

Our findings lead us to conclude that certain prosodic cues regarding uncertainty are localized in the target region (i.e., the word or words that a speaker is uncertain about) while other prosodic cues are manifested in the surrounding context. This result will help to answer the broad question of how to determine which part of an utterance a speaker is uncertain about.

We collected a corpus of utterances, uttered under varying levels of certainty, in a controlled fashion to allow for sub-utterance level prosodic analysis. The comparison of features extracted from the whole utterance, from the context, and from the target word is one of many experiments that are possible. In the future, we plan to compare features from target words occurring in utterances of different levels of certainty both between speakers and within a particular speaker. We also plan to compare classification accuracies for confidence and uncertainty using various combinations of features.

Another direction for future work is to extract and analyze prosodic features in utterances for which there was a significant difference between the speaker's level of certainty and the perceived level of certainty. For example, it would be useful to know how to characterize utterances where the speaker is uncertain but human listeners do not detect this uncertainty. Such information would be useful to researchers working on automatic emotion detection as well as to cognitive scientists, teachers, and researchers developing educational technology.

Finally, because of the tradeoff between collecting spontaneous speech and controlling the lexical content of utterances, in order to extend the conclusions of this and future analyses more broadly, our future work includes plans to test on spontaneous speech.

## 7. Acknowledgements

The author wishes to thank Stuart Shieber, Barbara Grosz, Abeer Alwan, and Enrique Henestroza for their help, guidance, and feedback. This research was supported in part by an NSF Graduate Research Fellowship.

## 8. References

[1] Hirschberg, J., "Intonation and pragmatics," in L. Horn and G. Ward, ed., *Handbook of Pragmatics*, Blackwell, 2003.

[2] Ang, J., Dhillon, R., Krupski, A., Shriberg, E., and Stolcke, A., "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proceedings of ICSLP 2002*, pp. 2037–2040, Denver, Colorado, 2002.

[3] Litman, D. and Forbes-Riley, K., "Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors," *Speech Communication 48*:559–590, 2006.

[4] Lee., C. and Narayanan, S., "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing 13(2)*:293–303, 2005.

[5] Smith, V. and Clark, H., "On the course of answering questions," *Journal of Memory and Language 32*:25–38, 1993.

[6] Brennan, S. and Williams, M., "The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers," *Journal of Memory and Language 34*:383–398, 1995.

[7] Liscombe, J., Hirschberg, J. and Venditti, J., "Detecting certainness in spoken tutorial dialogues," in *Proceedings of Eurospeech'05*, Lisbon, Portugal, 2005.

[8] Pon-Barry, H., Schultz, K., Bratt, E., Clark, B. and Peters, S., "Responding to student uncertainty in spoken tutorial dialogue systems," *International Journal of Artificial Intelligence in Education 16*, 171-194, 2006.

[9] Forbes-Riley, K., Litman, D. and Rotaru, M., "Responding to student uncertainty during computer tutoring: a preliminary evaluation," in *Proceedings of the 9th International Conference on Intelligent Tutoring Systems (ITS)*, Montreal, Canada, June 2008.