The Importance of Sub-Utterance Prosody in Predicting Level of Certainty

Heather Pon-Barry
School of Engineering and Applied Sciences, Harvard University
Cambridge, MA 02138, ponbarry@eecs.harvard.edu

Stuart Shieber
School of Engineering and Applied Sciences, Harvard University
Cambridge, MA 02138, shieber@seas.harvard.edu

Overview

We address the problem of predicting the perceived level of certainty of a spoken utterance. We have a corpus of utterances spoken under varying levels of certainty. In each utterance, a single word or phrase is responsible for the speaker's level of certainty. We investigate whether using prosodic features of this word or phrase and of its surrounding context improves the prediction accuracy when compared to using features taken only from the utterance as a whole.

We go beyond previous work by looking at the predictive power of prosodic features extracted from salient sub-utterance segments. Previous research on uncertainty has examined the predictive power of utterance- and intonational phrase-level prosodic features (Liscombe et al., 2005). Our results suggest that we can do a better job at predicting an utterance's perceived level of certainty by using prosodic features extracted from the whole utterance plus ones extracted from salient pieces of the utterance, without increasing the total number of features, than by using only features from the whole utterance.

This work is relevant to spoken language applications in which the system can identify locations likely to cause uncertainty. Examples of such systems include tutorial dialogue systems (Pon-Barry et al., 2006; Forbes-Riley et al., 2008) and second language learning and literacy systems (Alwan et al., 2007).

Uncertainty Corpus

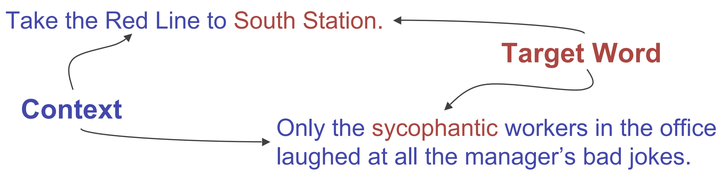
- 20 speakers
- 600 utterances
- Method of elicitation:
  1. Speakers are presented with a sentence containing one or more gaps
  2. Options for filling in the gap are displayed
  3. Upon hearing a beep the speaker read the sentence aloud

Transportation

Q: How can I get from Harvard to the Silver Line?
A: Take the Red Line to _____.
a. South Station
b. Downtown Crossing

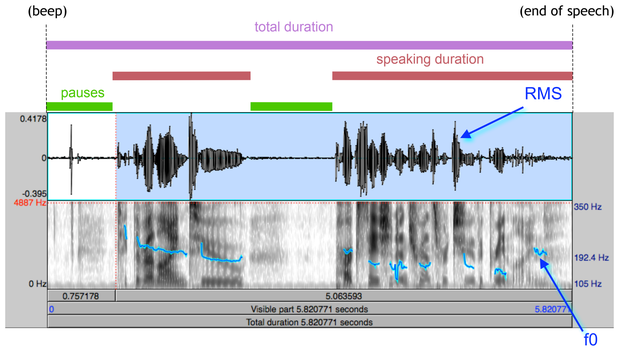
Vocabulary

Only the _____ workers in the office laughed at all the manager's bad jokes.
a. pugnacious
b. craven
c. sycophantic
d. spoffish



- Five annotators rated the perceived level of certainty on a 5-point scale
- Speakers rated their own level of certainty on the same 5-point scale

Prosodic Features



"Take the red line to Park Station and transfer to the green line."

- The *Combination* feature set (shaded in table below) is created by selecting either the whole utterance feature, the context feature, or the target word feature, whichever one is most strongly correlated with perceived level of certainty

Correlations with Perceived Level of Certainty

Feature-type	Whole Utterance	Context	Target Word
min f0	0.107	**0.119**	0.041
max f0	−0.073	**−0.153**	−0.045
mean f0	0.033	**0.070**	−0.004
stdev f0	−0.035	**−0.047**	−0.043
range f0	−0.128	**−0.211**	−0.075
rel. position min f0	0.042	0.022	**0.046**
rel. position max f0	**0.015**	0.008	0.001
abs. slope f0 (Hz)	**0.275**	0.180	0.191
abs. slope f0 (Semi)	**0.160**	0.147	0.002
min RMS	0.101	**0.172**	0.027
max RMS	−0.091	**−0.110**	−0.034
mean RMS	−0.012	**0.039**	−0.031
stdev RMS	−0.002	−0.003	**−0.019**
rel. position min RMS	0.101	**0.172**	0.027
rel. position max RMS	**−0.039**	−0.028	−0.007
total silence	**−0.643**	−0.507	−0.495
percent silence	−0.455	−0.225	**−0.532**
total duration	**−0.592**	−0.502	−0.590
speaking duration	**−0.430**	−0.390	−0.386
speaking rate	0.090	0.014	**0.136**

Acknowledgements

This work was supported in part by a National Defense Science and Engineering Graduate Fellowship.

Results

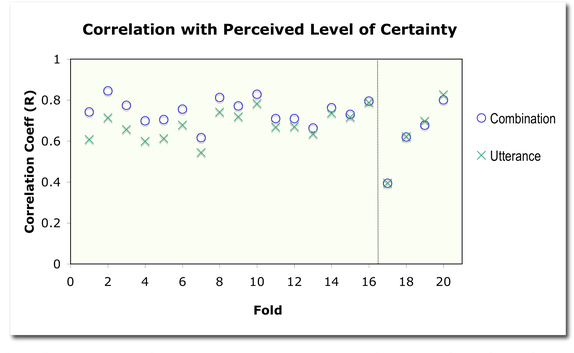
Linear Regression model accuracies

- Value to predict: perceived level of certainty
- Results shown: 20-fold 'leave one speaker out' cross-validation averages

Feature Set	Num Features	Accuracy (5 classes)	Accuracy (3 classes)
Naive Baseline	N/A	31.46%	56.25%
(A) Utterance	20	39.00%	68.96%
(B) Target Word	20	43.13%	68.96%
(C) Context	20	37.71%	67.50%
(D) All	60	48.54%	74.58%
(E) Combination	20	45.42%	74.79%

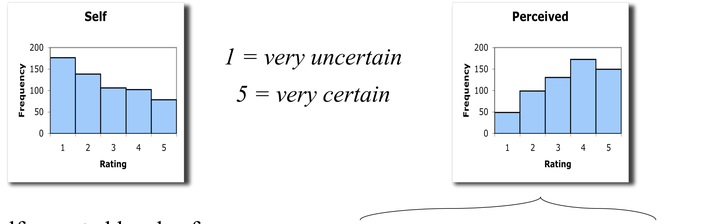
- **Combination** feature set, with only 20 features, yields higher average accuracies than **Utterance** feature set
- Similar behavior for SVM prediction models

Are the differences due to noise?



- **Combination** set predictions are more strongly correlated with perceived level of certainty than **Utterance** set predictions in 16 out of 20 folds

Self vs. Perceived Level of Certainty



- Self-reported levels of certainty were consistently lower than perceived levels of certainty