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, USA
ponbarry@eecs.harvard.edu

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

Abstract

We present an experiment aimed at understanding how to optimally use acoustic and prosodic information to predict a speaker's level of certainty. With a corpus of utterances where we can isolate a single word or phrase that is responsible for the speaker's level of certainty we use different sets of sub-utterance prosodic features to train models for predicting an utterance's perceived level of certainty. Our results suggest that using prosodic features of the word or phrase responsible for the level of certainty and of its surrounding context improves the prediction accuracy without increasing the total number of features when compared to using only features taken from the utterance as a whole.

1 Introduction

Prosody is a fundamental part of human-to-human spoken communication; it can affect the syntactic and semantic interpretation of an utterance (Hirschberg, 2003) and it can be used by speakers to convey their emotional state. In recent years, researchers have found prosodic features to be useful in automatically detecting emotions such as annoyance and frustration (Ang et al., 2002) and in distinguishing positive from negative emotional states (Lee and Narayanan, 2005).

In this paper, we address the problem of predicting the perceived level of certainty of a spoken utterance. Specifically, we have a corpus of utterances where it is possible to isolate a single word or phrase responsible for the speaker's level of certainty. With this corpus we investigate whether using prosodic features of the word or phrase causing uncertainty and of its surrounding context improves the prediction accuracy when compared to using features taken only from the utterance as a whole.

This work goes beyond existing research by looking at the predictive power of prosodic features extracted from salient sub-utterance segments. Previous work on uncertainty has examined the predictive power of utterance- and intonational phrase-level prosodic features (Liscombe et al., 2005) as well as the relative strengths of correlations between level of certainty and sub-utterance prosodic features (Pon-Barry, 2008). 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 knows specific words or phrases that are likely to cause uncertainty. For example, this would occur in a tutorial dialogue system when the speaker answers a direct question (Pon-Barry et al., 2006; Forbes-Riley et al., 2008), or in language (foreign or ESL) learning systems and literacy systems (Alwan et al., 2007) when new vocabulary is being introduced.

2 Previous Work

Researchers have examined certainty in spoken language using data from tutorial dialogue systems (Liscombe et al., 2005) and data from an uncertainty corpus (Pon-Barry, 2008).

Liscombe et al. (2005) trained a decision tree

Proceedings of NAACL HLT 2009: Short Papers, pages 105–108,
Boulder, Colorado, June 2009.

classifier on utterance-level and intonational phrase-level prosodic features to distinguish between certain, uncertain, and neutral utterances. They achieved 76% accuracy, compared to a 66% accuracy baseline (choosing the most common class).

We have collected a corpus of utterances spoken under varying levels of certainty (Pon-Barry, 2008). The utterances were elicited by giving adult native English speakers a written sentence containing one or more gaps, then displaying multiple options for filling in the gaps and telling the speakers to read the sentence aloud with the gaps filled in according to domain-specific criteria. We elicited utterances in two domains: (1) using public transportation in Boston, and (2) choosing vocabulary words to complete a sentence. An example is shown below.

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

The term 'context' refers to the fixed part of the response (*"Take the red line to _____",* in this example) and the term 'target word' refers to the word or phrase chosen to fill in the gap.

The corpus contains 600 utterances from 20 speakers. Each utterance was annotated for level of certainty, on a 5-point scale, by five human judges who listened to the utterances out of context. The average inter-annotator agreement (Kappa) was 0.45. We refer to the average of the five ratings as the 'perceived level of certainty' (the quantity we attempt to predict in this paper).

We computed correlations between perceived level of certainty and prosodic features extracted from the whole utterance, the context, and the target word. Pauses preceding the target word were considered part of the target word; all segmentation was done manually. Because the speakers had unlimited time to read over the context before seeing the target words, the target word is considered to be the *source* of the speaker's confidence or uncertainty; it corresponds to the decision that the speaker had to make. Our correlation results suggest that while some prosodic cues to level of certainty were strongest in the whole utterance, others were strongest in the context or the target word. In this paper, we extend this past work by testing the prediction accuracy of models trained on different subsets of these prosodic features.

3 Prediction Experiments

In our experiments we used 480 of the 600 utterances in the corpus, those which contained exactly one gap. (Some had two or three gaps.) We extracted the following 20 prosodic feature-types from each whole utterance, context, and target word (a total of 60 features) using WaveSurfer[1] and Praat[2].

Pitch: minf0, maxf0, meanf0, stdevf0, rangef0, relative position minf0, relative position maxf0, absolute slope (Hz), absolute slope (semitones)

Intensity: minRMS, maxRMS, meanRMS, stdevRMS, relative position minRMS, relative position maxRMS

Temporal: total silence, percent silence, total duration, speaking duration, speaking rate

These features are comparable to those used in Liscombe et al.'s (2005) prediction experiments. The pitch and intensity features were represented as z-scores normalized by speaker; the temporal features were not normalized.

Next, we created a 'combination' set of 20 features based on our correlation results. Figure 1 illustrates how the combination set was created: for each prosodic feature-type (each row in the table) we chose either the whole utterance feature, the context feature, or the target word feature, whichever one had the strongest correlation with perceived level of certainty. The selected features (highlighted in Figure 1) are listed below.

Whole Utterance: total silence, total duration, speaking duration, relative position maxf0, relative position maxRMS, absolute slope (Hz), absolute slope (semitones)

Context: minf0, maxf0, meanf0, stdevf0, rangef0, minRMS, maxRMS, meanRMS, relative position minRMS

Target Word: percent silence, speaking rate, relative position minf0, stdevRMS

[1] http://www.speech.kth.se/wavesurfer/
[2] http://www.fon.hum.uva.nl/praat/

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**

Figure 1: *The* Combination *feature set (highlighted in table) was produced by selecting either the whole utterance feature, the context feature, or the target word feature for each prosodic feature-type, whichever one was most strongly correlated with perceived level of certainty.*

To compare the prediction accuracies of different subsets of features, we fit five linear regression models to the feature sets. The five subsets are: (A) whole utterance features only, (B) target word features only, (C) context features only, (D) all features, and (E) the combination feature set. We divided the data into 20 folds (one fold per speaker) and performed a 20-fold cross-validation for each set of features. Each experiment fits a model using data from 19 speakers and tests on the remaining speaker. Thus, when we test our models, we are testing the ability to classify utterances of an unseen speaker.

Table 1 shows the accuracies of the models trained on the five subsets of features. The numbers reported are averages of the 20 cross-validation accuracies. We report results for two cases: 5 prediction classes and 3 prediction classes. We first computed the prediction accuracy over five classes (the regression output was rounded to the nearest integer). Next, in order to compare our results to those of Liscombe et al. (2005), we recoded the 5-class results into 3-class results, following Pon-Barry (2008), in the way that maximized inter-annotator agreement. The naive baseline numbers are the accuracies that would be achieved by always choosing the most common class.

4 Discussion

Assuming that the target word is responsible for the speaker's level of certainty, it is not surprising that the target word feature set (B) yields higher accuracies than the context feature set (C). It is also not surprising that the set of all features (D) yields higher accuracies than sets (A), (B), and (C).

The key comparison to notice is that the combination feature set (E), with only 20 features, yields higher average accuracies than the utterance feature set (A): a difference of 6.42% for 5 classes and 5.83% for 3 classes. This suggests that using a combination of features from the context and target word in addition to features from the whole utterance leads to better prediction of the perceived level of certainty than using features from only the whole utterance.

One might argue that these differences are just due to noise. To address this issue, we compared the prediction accuracies of sets (A) and (E) per fold. This is illustrated in Figure 2. Each fold in our cross-validation corresponds to a different speaker, so the folds are *not* identically distributed and we do not expect each fold to yield the same prediction accuracy. That means that we should compare predictions of the two feature sets within folds rather than between folds. Figure 2 shows the correlations between the predicted and perceived levels of certainty for the models trained on sets (A) and (E). The combination set (E) predictions were more strongly correlated than whole utterance set (A) predictions in 16 out of 20 folds. This result supports our claim that using a combination of features from the context and target word in addition to features from the whole utterance leads to better prediction of level of certainty.

Our best prediction accuracy for the 3 class case, 74.79%, was slightly lower than the accuracy reported by Liscombe et al. (2005), 76.42%. However, our difference from the naive baseline was 18.54% where Liscombe et al.'s was 10.42%. Liscombe et al. randomly divided their data into training and test sets, so it is unclear whether they tested on seen or unseen speakers. Further, they ran one experiment rather than a cross-validation, so their reported accuracy may not be indicative of the entire data set.

We also trained support vector models on these subsets of features. The main result was the same:

Table 1: *Average prediction accuracies for the linear regression models trained on five subsets of prosodic features. The models trained on the* Combination *feature set and the* All *feature set perform better than the other three models in both the 3- and 5-class settings.*

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%

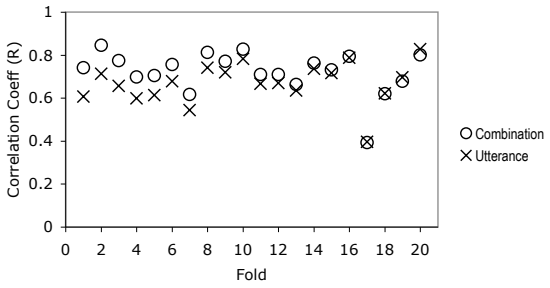

Figure 2: *Correlations with perceived level of certainty per fold for the* Combination (O) *and the* Utterance (X) *feature set predictions, sorted by the size of the difference. In 16 of the 20 experiments, the correlation coefficients for the* Combination *feature set are greater than those of the* Utterance *feature set.*

the set of all features (D) and the combination set (E) had better prediction accuracies than the utterance feature set (A). In addition, the combination set (E) had the best prediction accuracies (of all models) in both the 3- and 5-class settings. The raw accuracies were approximately 5% lower than those of the linear regression models.

5 Conclusion and Future Work

The results of our experiments suggest a better predictive model of level of certainty for systems where words or phrases likely to cause uncertainty are known ahead of time. Without increasing the total number of features, combining select prosodic features from the target word, the surrounding context and the whole utterance leads to better prediction of level of certainty than using features from the whole utterance only. In the near future, we plan to experiment with prediction models of the speaker's self-reported level of certainty.

Acknowledgments

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

References

Abeer Alwan, Yijian Bai, Matthew Black, et al. 2007. A system for technology based assessment of language and literacy in young children: the role of multiple information sources. *Proc. of IEEE International Workshop on Multimedia Signal Processing*, pp. 26–30, Chania, Greece.

Jeremy Ang, Rajdip Dhillon, Ashley Krupski, et al. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. *Proc. of ICSLP 2002*, pp. 2037–2040, Denver, CO.

Kate Forbes-Riley, Diane Litman, and Mihai Rotaru. 2008. Responding to student uncertainty during computer tutoring: a preliminary evaluation. *Proc. of the 9th International Conference on Intelligent Tutoring Systems*, Montreal, Canada.

Julia Hirschberg. 2003. Intonation and pragmatics. In L. Horn and G. Ward (ed.), *Handbook of Pragmatics*, Blackwell.

Chul Min Lee and Shrikanth Narayanan. 2005. Towards detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303.

Jackson Liscombe, Julia Hirschberg, and Jennifer Venditti. 2005. Detecting certainness in spoken tutorial dialogues. *Proceedings of Eurospeech 2005*, Lisbon, Portugal.

Heather Pon-Barry, Karl Schultz, Elizabeth Bratt, Brady Clark, and Stanley Peters. 2006. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education* 16:171-194.

Heather Pon-Barry. 2008. Prosodic manifestations of confidence and uncertainty in spoken language. *Proc. of Interspeech 2008*, pp. 74–77, Brisbane, Australia.