INTERSPEECH 2009 BRIGHTON

Identifying Uncertain Words within an Utterance via Prosodic Features

Heather Pon-Barry, Stuart Shieber

School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

ponbarry@eecs.harvard.edu, shieber@seas.harvard.edu

Abstract

We describe an experiment that investigates whether sub-utterance prosodic features can be used to detect uncertainty at the word-level. That is, given an utterance that is classified as uncertain, we want to determine which word or phrase the speaker is uncertain about. We have a corpus of utterances spoken under varying degrees of certainty. Using combinations of sub-utterance prosodic features we train models to predict the level of certainty of an utterance. On a set of utterances that were perceived to be uncertain, we compare the predictions of our models for two candidate 'target word' segmentations: (a) one with the actual word causing uncertainty as the proposed target word, and (b) one with a control word as the proposed target word. Our best model correctly identifies the word causing the uncertainty rather than the control word 91% of the time.

Index Terms: prosody, spoken language understanding, uncertainty, emotion detection.

1. Introduction

This paper addresses the problem of detecting uncertainty in spoken language at the word-level. Existing work on emotion detection has focused on utterance-level classifications. Prosodic information has proven useful for tasks such as automatically detecting annoyance and frustration [1] and distinguishing positive and negative emotional states [2]. In the area of uncertainty detection, models trained on prosodic features have classified utterances as certain, uncertain, or neutral with accuracies of roughly 75% [3] [4]. In addition to prosodic cues, visual cues play an important role in conveying level of certainty among humans [5]. Since we are able to automatically extract prosodic features but not visual features, this paper examines models trained solely on prosodic features.

We go beyond previous work by using models trained on sub-utterance prosodic features to determine, in an utterance classified as uncertain, which word or phrase within the utterance a speaker is uncertain about. Utterance-level classifications of degree of certainty are sufficient if a system expects to be presented with utterances that contain exactly one central idea. However, in natural language, utterances are often long and rarely well-formed. The ability to determine which part of an utterance a speaker is uncertain about will enable the development of spoken language applications that understand and respond to people closer to the way that humans do. For example, this ability would be useful in spoken tutorial dialogue systems [6] [7], voice search applications [8], and second language learning and literacy systems [9].

Our approach is to build level-of-certainty prediction models that take combinations of utterance and sub-utterance prosodic features as input. We look at two types of prediction models: linear regression and support vector machine regression. The utterances our models are trained on are from a corpus containing speech of varying levels of certainty, where the degree of certainty can be attributed to a single word or phrase (called the 'target word'). We extract prosodic features from the target word, from the context (see example in Section 2), and from the whole utterance.

We compare the predicted level of certainty using the correct target word segmentation to the predicted level using an alternate segmentation where a control word is proposed as the target word. On the task of choosing between the actual target word and the control word our best model achieves an accuracy of 91%, a 71% error reduction over a baseline model trained on only non-prosodic features. This result suggests that our method of using sub-utterance prosodic features to detect uncertainty at the word-level is well-suited for this task.

2. Uncertainty Corpus

We have collected a corpus of utterances spoken under varying levels of certainty [10]. The utterances were elicited by presenting adult native English speakers with 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) answering questions about using public transportation in Boston, and (2) choosing vocabulary words to complete a sentence. An example from each domain is shown below.

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

(2) Mahler's revolutionary music, abrasive personality and _____ writings about art and life divided the city into warring factions.
   a. officious
   b. trenchant
   c. spoffish
   d. pugnacious

The term 'context' refers to the fixed part of the response (*"Take the red line to"* in example (1)) 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. The utterances display varying levels of certainty, as evidenced by the speaker's self-ratings as well as ratings from a group of human judges [10]. 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, which is on par with past work in emotion detection [2] [3]. We refer to the average of the five ratings as the 'perceived level of certainty.'

We extracted prosodic features 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. We measured correlations between each prosodic feature and the perceived level of certainty and found 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.

3. Method

We train linear regression and support vector machine (SVM) models to predict an utterance's perceived level of certainty using speech from the corpus described in Section 2 as training data. For the linear regression models we use the Weka data mining software toolkit[1] with M5 attribute selection. For the support vector machine regression models, we use the SVM-Light toolkit [11].

For a subset of utterances that were perceived to be uncertain (perceived level of certainty less than 2.5 where 1 means 'very uncertain' and 5 means 'very certain'), we identify a control word — a content word roughly the same length as the potential target words and if possible, the same part-of-speech. We balance the set of control words for position in the utterance relative to the position of the gap, i.e., half of the control words appear before the gap location and half appear after. After filtering utterances based on level of certainty and presence of an appropriate control word, 43 utterances remain. This is our test set. In the example items shown in Section 2, the corresponding control words are *red line* and *abrasive*.

We then compare the predicted level of certainty for two segmentations of the utterance: (a) the correct segmentation with the gap-filling word as the proposed 'target word' and (b) and alternative segmentation with the control word as the proposed 'target word.' Thus, the prosodic features extracted from the target word and from the context will be different in these two segmentations, while the features extracted from the utterance will be the same. The hypothesis we test in this experiment is that our models should predict a lower level of certainty when the prosodic features are taken from segmentation (a) rather than segmentation (b), thereby identifying the gap-filling word as the source of the speaker's uncertainty.

Before training our prediction models, we filter out utterances in the corpus that contain more than one gap. (120 of the 600 utterances have two or three gaps.) We train the models on prosodic features from only the *correct* target word/context segmentations.

To ensure that our models will make predictions for unseen speakers during testing, each 'model' that we train is actually a collection of 20 prediction models, one for each subset of 19 speakers. (The corpus contains speech data from 20 speakers.) When making predictions for an utterance in the test set, we use the corresponding model whose training data includes no utterances from that particular speaker.

3.1. Non-prosodic features

We train a prediction model on non-prosodic features to serve as a baseline for the models that we train on prosodic features. We assume that words that contain more syllables and words that are infrequent or previously unseen will generally take longer to speak aloud than words of shorter length and higher frequency or familiarity. We want to ensure that the predictions our prosodic models make are not able to be explained by these features or by part-of-speech or position features.

Our non-prosodic model has 20 features. The part-of-speech features include binary features for the possible parts-of-speech of the target word and of its immediately preceding word. Utterance position is represented as the utterance's ordinal position among the sequence of items (the order varied for each speaker). Word position features include the target word's index from the start of the utterance, index from the end, and relative position (index from start/total words in utterance). The word length features include the number of characters, phonemes, and syllables in the target word. To account for familiarity, we include a feature for how many times during the experiment the speaker has previously uttered the target word. To approximate word frequency, we use the log-probability based on British National Corpus counts where available. For words that do not appear in the British National corpus, we estimate feature values by using web-based counts (Google hits) to interpolate unigram frequencies. It has been demonstrated that using web-based counts is a reliable method for estimating unseen n-gram frequencies [12].

3.2. Prosodic features

Table 1 lists the 20 prosodic feature-types that we extract from each whole utterance, context, and target word using WaveSurfer[2] and Praat[3] (resulting in 60 prosodic features). These feature-types are comparable to those used in past level-of-certainty prediction experiments [3] [4]. The pitch and intensity features are represented as z-scores normalized by speaker; the temporal features are not normalized. The f0 contour is extracted using WaveSurfer's ESPS method. We use the ratio of voiced frames to total frames as an approximation of the speaking rate.

Pitch	min f0	relative position min f0
	max f0	relative position max f0
	mean f0	absolute slope (Hz)
	stdev f0	absolute slope (Semi)
	range f0	
Intensity	min RMS	relative position min RMS
	max RMS	relative position max RMS
	mean RMS	stdev RMS
Temporal	total silence	percent silence
	total duration	speaking duration
	speaking rate	

Table 1: *Prosodic feature-types extracted from each whole utterance, context, and target word.*

3.3. Combination Feature Set

We create a 'combination' set of 20 features based on our correlation results from previous work [10] (see Section 2). Table 2 illustrates how the combination set is created: for each prosodic feature-type (each row in the table) we choose either the whole

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://www.speech.kth.se/wavesurfer/

[3] http://www.fon.hum.uva.nl/praat/

utterance feature, the context feature, or the target word feature, whichever one has the strongest correlation with perceived level of certainty. The selected features (highlighted in Table 2) are listed below.

1. **Whole Utterance:** total silence, total duration, speaking duration, relative position max f0, relative position max RMS, absolute slope (Hz), absolute slope (semitones)
2. **Context:** min f0, max f0, mean f0, stdev f0, range f0, min RMS, max RMS, mean RMS, relative position min RMS
3. **Target Word:** percent silence, speaking rate, relative position min f0, stdev RMS

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**

Table 2: *The* Combination *feature set (highlighted in table) is produced by selecting either the whole utterance feature, the context feature, or the target word feature for each prosodic feature-type, whichever one is most strongly correlated with perceived level of certainty.*

3.4. Experiment Feature Sets

In our experiment, we train models on six sets of prosodic features. The *Target Word* set contains the 20 prosodic feature-types (see Table 1) extracted from the target word region. Likewise, the *Context* set contains the 20 prosodic feature-types extracted from the context region. We do not have an *Utterance* feature set because the prosodic features from the utterance have the same values in the correct segmentation and the control segmentation. The *Target Word, Context, Utterance* set contains all 60 prosodic features. The *Target Word, Utterance* set is the union the 20 target word features and the 20 utterance features. The *Combination (Target Word, Context, Utterance)* set contains 20 prosodic features: a mixture of context, target word, and utterance features (see Section 3.3). The *Combination (Target Word)* set contains only the four target word features from the combination set.

We also train models on two feature sets containing both prosodic and non-prosodic features. The *Target Word, Non-prosodic* feature set is the union of the 20 target word features and the 20 non-prosodic features. The *Target Word, Context, Utterance, Non-prosodic* feature set is the union of all 60 prosodic features and the 20 non-prosodic features.

4. Results

Our models yield accuracies as high as 91% on the task of identifying the word or phrase causing uncertainty when choosing between the actual word and a control word. The experiments were run using models trained on subsets of the whole corpus (see Section 3) and tested on the 43 utterances that have a perceived level of certainty less than 2.5 and contain a suitable control word.

Table 3 shows the linear regression and support vector machine detection accuracies. The models trained on the non-prosodic features provide a baseline from which to compare the performance of the models trained on prosodic features. This baseline accuracy is 67% for both the linear regression and SVM models.

The linear regression model trained on the target word feature set had the highest accuracy among all the prosodic models, 86%. The highest overall accuracy, 91%, was achieved on the linear regression model trained on the target word features plus the non-prosodic features from the baseline set.

The support vector machine model trained on the target word feature set had an accuracy of 79%. This was the second highest accuracy among the SVM models. The SVM model with the highest accuracy, 81%, was the one trained on all of the prosodic features (all the context, target word, and utterance features). For these two SVM models, the addition of the non-prosodic features from the baseline set had no effect on the accuracy.

5. Discussion

This experiment shows that sub-utterance prosodic features are useful in detecting uncertainty at the word-level. Our best model, the linear regression model that uses target word prosodic features plus the non-prosodic features from the baseline set, identifies the correct word 91% of the time whereas the linear regression baseline model using only non-prosodic features is accurate just 67% of the time. This is an absolute difference of 23% and an error reduction of 71%. The best SVM model, trained on all the prosodic features, has an accuracy 14% above the SVM baseline, an error reduction of 43%. These large improvements over the non-prosodic baseline models imply that sub-utterance prosodic features are crucial in word-level uncertainty detection.

In creating the non-prosodic feature set for this experiment we wanted to account for the most obvious differences between the target words and the control words. The baseline model's low accuracy on this task is to be expected because the non-prosodic features are not good at explaining the variance in the response variable (perceived level of certainty): the correlation coefficient for the baseline linear regression model is only 0.27 (as a comparison, the coefficient for the target word linear regression model is 0.67).

The combination feature set, which in our past work had high accuracy in classifying an utterance's overall level of certainty [4], did not perform as well as the other feature sets for this detection task. We speculate that this may have to do with the context features. While the prosodic features we extracted from the context are beneficial in classifying an utterance's overall level of certainty, the low accuracies for the con-

Table 3: *Accuracies on the task of identifying the word or phrase causing uncertainty when choosing between the actual word and a control word. Experiments were run on both linear regression and support vector machine models. The linear regression model that was trained on the set of target word features and non-prosodic features achieves* 91% *accuracy.*

Feature Set	Number of Features	Linear Regression Detection Accuracy	Support Vector Machine Detection Accuracy
Non-prosodic (baseline)	20	67.44%	67.44%
Target Word, Non-prosodic	40	90.70%	79.07%
Target Word	20	86.05%	79.07%
Target Word, Context, Utterance	60	79.07%	81.40%
Target Word, Context, Utterance, Non-prosodic	80	76.74%	81.40%
Target Word, Utterance	40	69.77%	72.09%
Combination Set (Target Word)	4	72.09%	67.44%
Combination Set (Target Word, Context, Utterance)	20	72.09%	53.49%
Context	20	48.84%	27.91%

text feature set in Table 3 suggest that they are detrimental in determining which word a speaker is uncertain about, using our proposed method. The task we examine in this paper, distinguishing the actual 'target word' from a control word, is different than the task the models are trained on (predicting a real-valued level of certainty), therefore we do not expect the models with the highest classification accuracy to necessarily perform well on the task of identifying the word causing uncertainty.

It is not clear which model type, linear regression or support vector machine regression, is better suited for the general task of identifying uncertain words within an utterance. Among all the models we trained, the two with the highest accuracies were both linear regression models (one with only prosodic features and one with a mixture of prosodic and non-prosodic features). However, the SVM models yielded higher accuracies than the linear regression models for three of the eight sets of features examined.

6. Conclusion

We built level-of-certainty prediction models that take utterance and sub-utterance prosodic features and non-prosodic features as input. Using these models, we compared the predicted level of certainty using the correct 'target word' segmentation with the predicted level using an alternate segmentation with a control word as the proposed target word. On the task of identifying the correct segmentation, our best linear regression and SVM models achieve error reductions of 71% and 43%, respectively, over the baseline models trained on only non-prosodic features. These results imply that prosodic information is crucial in identifying uncertain words within an utterance.

The experiment described in this paper is an initial step towards understanding whether prosodic information can be used more generally to determine which word or phrase within an utterance, among all candidate words and phrases, is the cause of a speaker's uncertainty. Since our prediction model was able to choose the correct target word over the control word 91% of the time we have reason to believe that this method of using prediction models trained on sub-utterance prosodic features will be successful in the more general setting.

7. Acknowledgements

This work was supported in part by a National Defense Science and Engineering Graduate Fellowship. We thank the reviewers for their helpful comments.

8. References

[1] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2002, pp. 2037–2040.

[2] C. M. Lee and S. Narayanan, "Towards detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

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

[4] H. Pon-Barry and S. Shieber, "The importance of sub-utterance prosody in predicting level of certainty," in *Proceedings of NAACL-HLT*, Boulder, CO, June 2009.

[5] E. Krahmer and M. Swerts, "How children and adults produce and perceive uncertainty in audiovisual speech," *Language and Speech*, vol. 48, no. 1, pp. 29–53, 2005.

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

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

[8] T. Paek and Y.-C. Ju, "Accommodating explicit user expressions of uncertainty in voice search or something like that," in *Proceedings of Interspeech*, Brisbane, Australia, September 2008.

[9] A. Alwan, Y. Bai, M. Black, L. Casey, M. Gerosa, M. Heritage, M. Iseli, B. Jones, A. Kazemzadeh, S. Lee, S. Narayanan, P. Price, J. Tepperman, and S. Wang, "A system for technology based assessment of language and literacy in young children: the role of multiple information sources," in *Proceedings of IEEE International Workshop on Multimedia Signal Processing*, Chania, Greece, 2007, pp. 26–30.

[10] H. Pon-Barry, "Prosodic manifestations of confidence and uncertainty in spoken language," in *Proceedings of Interspeech*, Brisbane, Australia, September 2008, pp. 74–77.

[11] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in Kernel Methods: Support Vector Machines*.

[12] F. Keller and M. Lapata, "Using the web to obtain frequencies for unseen bigrams," *Computational Linguistics*, vol. 29, no. 3, pp. 459–484, 2003.