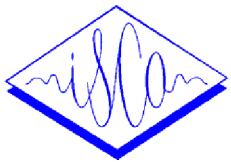

Speech Prosody 2010
Chicago, IL, USA
May 10-14, 2010

Assessing Self-awareness and Transparency when Classifying a Speaker's Level of Certainty

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

This paper is about using prosody to automatically detect one aspect of a speaker's internal state: their level of certainty. While past work on classifying level of certainty used the *perceived* level of certainty as the value to predict, we find that this quantity often differs from a speaker's *actual* level of certainty as gauged by self-reports. In this work we build models to predict a speaker's self-reported level of certainty using prosodic features. Our data is a corpus of single-sentence utterances that are annotated with (1) whether the statement is correct or incorrect, (2) the perceived level of certainty, and (3) the self-reported level of certainty. Knowing the self-reported level of certainty, in conjunction with the perceived level of certainty, allows us to assess what we will refer to as the speaker's *transparency*. Knowing the self-reported level of certainty, in conjunction with the correctness of the answer, allows us to assess what we will refer to as *self-awareness*. Our models, trained on prosodic features, correctly classify the self-reported level of certainty 75% of the time. Intelligent systems can use this information to make inferences about the user's internal state, for example whether the user of a system has a misconception, makes a lucky guess, or needs encouragement.

1. Introduction

In natural conversation, prosody often conveys information about a speaker's internal state [1, 2, 3] or about their social intention [4]. Recent work has investigated how dialogue systems can appropriately respond to a speaker based on their prosody, e.g., by altering the content of system responses [5] or by altering the emotional coloring of system responses [6]. Still, enabling such systems to detect and attend to prosodic cues is a difficult task. This problem is challenging because speakers vary in the degree to which they use prosody to convey their internal state or intentions. As a result, there can be a mismatch between how a speaker feels and how they are perceived to be feeling [4]. In this paper we focus on the domain of *level of certainty*: we examine the prosodic characteristics of speech with respect to whether a speaker's self-reported level of certainty matches how certain he or she is perceived to be.

Previous work on level of certainty classification has focused on classifying an utterance's *perceived* level of certainty, achieving classification accuracies of up to 76% [7, 8]. However, in applications where level of certainty information is useful, such as spoken tutoring systems [5] and second language learning systems [9], we would like to know how certain the speaker actually is, not just how certain they are perceived to be. This knowledge affects the inferences such systems can make about the speaker's internal state, for example whether the speaker has a misconception, makes a lucky guess, or might benefit from some encouragement.

In this work we define two new categories of classification pertaining to level of certainty: *transparency* and *self-awareness*. We consider a speaker to be *transparent* if their self-reported level of certainty matches how certain a set of listeners perceive them to be. We consider a speaker to be *self-aware* if their self-reported level of certainty is in accordance with the correctness of their response. That is, feeling certain when correct and feeling uncertain when incorrect are considered self-aware.

Our approach is to build machine learning models that use prosodic features to distinguish transparent speakers from non-transparent speakers and self-aware speakers from non-self-aware speakers. For this experiment, we consider the situation where the speaker's *perceived level of certainty* and whether or not their statement was *correct* are known quantities. Our training data comes from a corpus containing speech of varying levels of certainty; this corpus is described in Section 3. Using a set of standard prosodic features, our models correctly classify the user's internal state 75% of the time.

2. Self-awareness and Transparency

With knowledge of a speaker's internal level of certainty, we can assess the speaker's self-awareness and transparency for each utterance.

The concept of self-awareness applies to utterances whose correctness can be determined. We consider a speaker to be *self-aware* if they feel certain when correct and feel uncertain when incorrect. The four possible combinations of correctness vs. internal level of certainty are illustrated in Fig 1. For educational applications, systems that can assess self-awareness can assess whether or not the user is at a learning impasse [5]. We claim that the most serious learning impasses correspond to the cases where a speaker is *not* self-aware. If a speaker feels certain and is incorrect, then it is likely that they have some kind of *misconception*. If a speaker feels uncertain and is correct, they either *lack confidence* or made a lucky guess. A follow-up question could be asked by the system to determine whether or not the user made a lucky guess. In our corpus, speakers were self-aware in 72.5% of the utterances.

The concept of speaker *transparency* is independent of an utterance's correctness. We consider a speaker to be transparent if they are perceived as certain when they feel certain and are perceived as uncertain when they feel uncertain. The four possible combinations of perceived vs. internal level of certainty are illustrated in Fig 2. If a system uses perceived level of certainty to determine what kind of feedback to give the user, then it will give inappropriate feedback to users who are *not* transparent. In our corpus, speakers were transparent in 63.6% of the utterances. We observed that some speakers acted like ra-

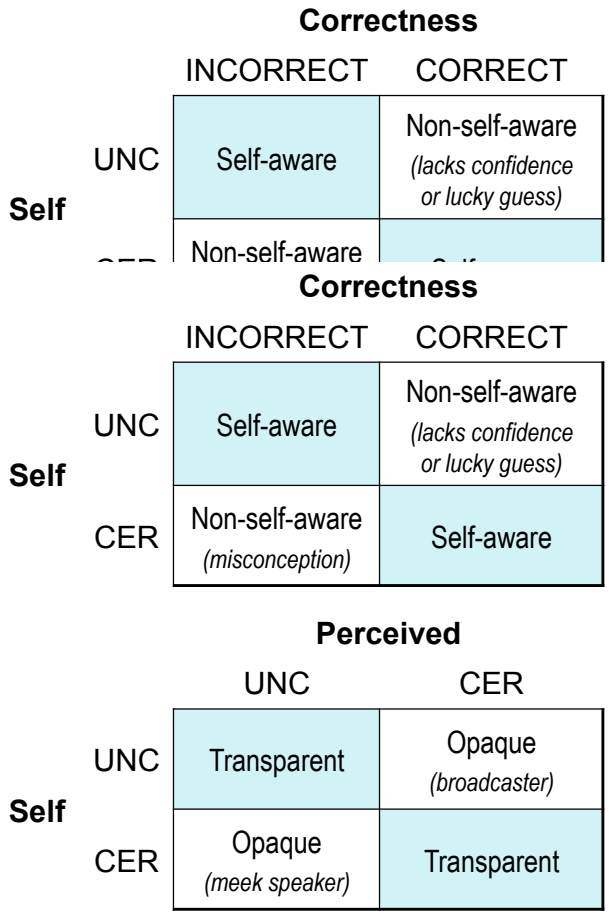

Figure 2: *Transparency: we consider a speaker to be transparent if the speaker's internal level of certainty reflects his or her utterance's perceived level of certainty.*

3. Uncertainty Corpus

In previous work we collected a corpus of utterances spoken under varying levels of certainty [10]. We elicited the utterances from adult native English speakers in the following way. First, we present the speaker with a written sentence containing one or more gaps, then we display multiple options for filling in the gaps and instruct 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 in Table 1.

The corpus contains 600 utterances from 20 speakers. Crucially, speakers rated their own level of certainty on a 5-point scale (1 = very uncertain, 5 = very certain). Each utterance was also annotated for perceived level of certainty on the same 5-point scale by five human judges who listened to the utterances out of context. The average inter-annotator agreement (Kappa statistic) was 0.45, which is on par with past work in emotion detection [2, 7]. We refer to the average of the five annotators' ratings as the 'perceived level of certainty'.

We found that the distribution of self-ratings is heavily concentrated on the uncertain side (with a mean rating of 2.6), whereas the annotators' ratings are more heavily concentrated on the certain side (with a mean rating of 3.5). This is illustrated in Figure 3. This observation motivated our current examination of the dimensions of speaker transparency and speaker self-awareness.

(1) Q: What is the best way to get to North Station from the Harvard T-stop?
A: Take the red line to _____
a. Park Station
b. Downtown Crossing
and transfer to the _____ .
a. green line
b. orange line

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

Table 1: *Examples of a transportation item and a vocabulary item used in eliciting speech of varying levels of certainty.*

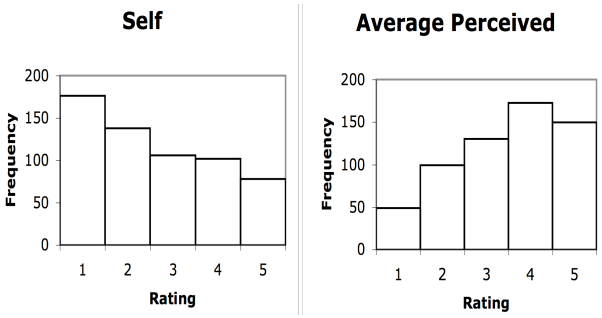

Figure 3: *Histograms illustrating the distribution of self-reported and perceived levels of certainty in our corpus.*

4. Method

4.1. Data Preparation

For our classification experiments, we treat self rating, perceived rating, and correctness as binary features. For both self rating and perceived rating, we map values less than 3 to 'uncertain' and values greater than or equal to 3 to 'certain'. To compute correctness, we code each multiple choice answer or answer-tuple as 'incorrect' or 'correct'. Based on the nature of the corpus, some questions had more than one correct answer and a few had no correct answers.

4.2. Prosodic features

Table 2 lists the 20 prosodic features that we extract from each utterance using WaveSurfer[1] and Praat[2]. These feature-types are comparable to those used in past level-of-certainty prediction experiments [7] [8]. The pitch and intensity features are represented as z-scores normalized by speaker[3]; the temporal

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

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

[3] When computing the absolute slope (semitones) feature, we converted to semitones before normalizing by speaker.

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 2: *Prosodic features extracted from each utterance.*

4.3. Classification Models

We build models for classifying a speaker's self-reported level of certainty in the following way. First, we divide the data into four subsets (see Figure 4) corresponding to the correctness of the answer and the perceived level of certainty, which for this

	Perceived Uncertain	Perceived Certain
Incorrect	Subset A′	Subset A
Correct	Subset B	Subset B′

Figure 4: *We divide the utterances into four subsets and train a separate classifier for each subset.*

In subset A', the distribution of self-reported levels of certainty is skewed: 84% of the utterances in subset A' are self-reported as *uncertain*. This imbalance is intuitive; someone who is incorrect *and* perceived as uncertain most likely feels uncertain too. Likewise, in subset B', the distribution of self-reported levels of certainty is skewed in the other direction: 76% of the utterances in this subset are self-reported as *certain*. This too is intuitive; someone who is correct *and* perceived as certain most likely feels certain as well. Therefore, we hypothesize that for subsets A' and B', classification models trained on prosodic features will do no better than choosing the subset-specific majority class.

Subsets A and B are the more interesting cases; they are the subsets where the perceived level of certainty is not aligned with the correctness. The self-reported levels of certainty for these subsets are less skewed: 65% *uncertain* for subset A and 54% *certain* for subset B. We hypothesize that for subsets A and B, decision tree models trained on prosodic features will be more accurate than selecting the subset-specific majority class. For each subset, we perform a k-fold cross-validation where we leave one speaker out of each fold. Because not all speakers have utterances in every subset, k ranges from 18 to 20. We also train support vector machine models for each subset, as a comparison.

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

5. Results

We first report the results of our decision tree models on subsets A (utterances that are incorrect and perceived as certain) and B (utterances that are correct and perceived as uncertain).[5] The accuracies we report are the averages of the k-fold cross-validation. For subset A, the accuracy in classifying the self-reported level of certainty is 68.99%; for subset B, the accuracy is 69.01%. For both subsets, the decision trees perform better than choosing the majority class for that subset. For subset A, choosing the majority class (self-report = *Uncertain*) would result in an accuracy of 65.19%. For subset B, choosing the majority class (self-report = *Certain*) would result in an accuracy of 53.52%. The decision trees also perform better than choosing the *overall* majority class before dividing the utterances into subsets (self-report = *Uncertain*), which would result in an accuracy of 52.30%. These results are summarized in Figure 5 (The horizontal line in Figure 5 represents the baseline accuracy achieved by choosing the *overall* majority class.)

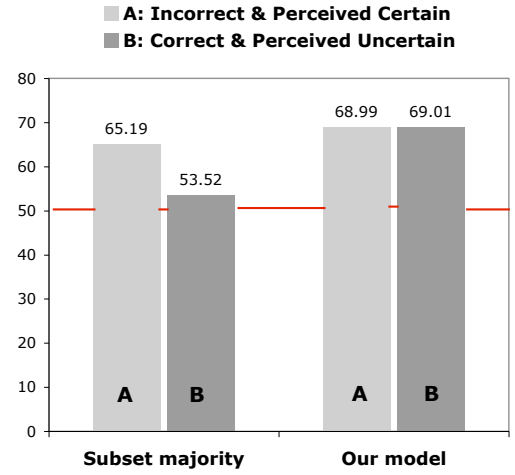

Figure 5: *Accuracy in classifying self-reported level of certainty for two subsets: (A) utterances that are incorrect and perceived as certain, and (B) utterances that are correct and perceived as uncertain. Our decision tree models perform better than choosing the subset-specific majority class (*Uncertain *for subset A;* Certain *for subset B), as well as better than choosing the overall majority class (*Uncertain*) which is represented by the horizontal line.*

For the other two subsets of utterances: A' (utterances that are incorrect and perceived as uncertain) and B' (utterances that are correct and perceived as certain), the decision tree models learned contained a single leaf corresponding to the majority class. Likewise, the SVM models yielded accuracies no better than choosing the most common class. Therefore, our best combined model uses the decision tree models for subsets A and B and chooses the subset-specific majority class for subsets A' and B′. This results in an overall accuracy of 75.30%,

[5]The accuracies for the support vector machine models were 7% lower than the accuracies for the decision tree models.

significantly outperforming the baseline of choosing the overall majority class (52.30%), as well as alternative baselines, e.g., assigning the self-reported level to be the same as the perceived level (63.67%), or training a single decision tree on all 600 utterances (66.33%).

In our decision tree models we find that the *percent silence* and *speaking rate* features are consistently selected[6] as attributes to split on, in other words, they lead to the highest information gain. Lower values of percent silence correspond to speakers feeling certain and higher values correspond to speakers feeling uncertain. This is not surprising; in our prior work on classifying *perceived* level of certainty, percent silence was strongly correlated with perceived level of certainty [8]. For speaking rate, very slow and very fast speaking rates correspond to speakers feeling certain. Values in the middle correspond to a mix of speakers feeling certain and uncertain. This observation is surprising; in our prior work, speaking rate was not strongly correlated with perceived level of certainty. This suggests that perhaps speaking rate is important in distinguishing internal levels of certainty from perceived levels of certainty.

Lastly, our data shows a wide range of variation among individual speakers. Average transparency values for individuals range from 0.27 to 0.80, with a median of 0.67. Similarly, average self-awareness values for individuals range from 0.43 to 0.87, with a median of 0.72. When we examine average transparency, per individual, for correct versus incorrect utterances, we see that some speakers are more transparent when correct than when incorrect. This is illustrated in Figure 6. The presence of outliers and clusters of individuals in Figure 6 suggests that there may be different 'speaking personalities' related to transparency (similar in spirit to the personality types described in [11]). If such speaking personality clusters exist, identifying them may enhance our ability to classify a speaker's internal level of certainty.

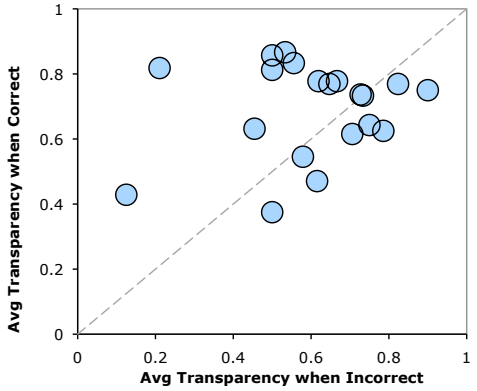

Figure 6: *Average transparency when correct vs. incorrect, per speaker. Each dot corresponds to an individual speaker. If speakers were equally transparent regardless of their correctness, we would expect them to fall along the dashed line.*

6. Conclusion

In this paper we explain why the concepts of self-awareness and transparency are important properties of speakers. When a speaker's internal state indicates that they are *not* self-aware or are *not* transparent, intelligent systems can use this knowledge to remediate misconceptions, offer encouragement, or determine the right feedback to give. We find that for two of the four subsets of utterances in our corpus, decision tree models trained on prosodic features are better at classifying a speaker's self-reported level of certainty than models that assign the subset majority class. For the other two subsets of utterances, the decision tree models learn to choose the subset majority class (i.e., gain no information from the prosodic features). Combining these models results in an overall accuracy of 75.3%, whereas choosing the overall majority class would be accurate only 52.3% of the time.

In addition, our results suggest that *speaking rate* may be useful in distinguishing self-reported ratings from perceived ratings. Also, associating utterances with particular 'speaking personalities' may be helpful in classifying a speaker's internal level of certainty.

[6]Speaking rate was one of the first two attributes split on in 100% of the cross-validation models; percent silence was one of the first two attributes in 95% of the cross-validation models.

7. Acknowledgements

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

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] 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.

[4] R. Ranganath, D. Jurafsky, and D. McFarland, "It's not you it's me: Detecting flirting and its misperception in speed-dates," in *Proceedings of EMNLP*, Singapore, 2009.

[5] 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.

[6] J. Acosta and N. Ward, "Responding to user emotional state by adding emotional coloring to utterances," in *Proceedings of Interspeech 2009*, Brighton, UK, 2009, pp. 1587–1590.

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

[8] 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.

[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] F. Mairesse, M. Walker, M. Mehl, and R. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of Artificial Intelligence Research*, vol. 30, pp. 457–500, 2007.