Eliciting and Annotating Uncertainty in Spoken Language

Heather Pon-Barry[1], **Stuart M. Shieber**[2], **Nicholas Longenbaugh**[2]

[1]School of Computing, Informatics, and Decision Systems Engineering, Arizona State University
[2]School of Engineering and Applied Sciences, Harvard University
ponbarry@asu.edu, shieber@seas.harvard.edu, nslonge@gmail.com

Abstract

A major challenge in the field of automatic recognition of emotion and affect in speech is the subjective nature of affect labels. The most common approach to acquiring affect labels is to ask a panel of listeners to rate a corpus of spoken utterances along one or more dimensions of interest. For applications ranging from educational technology to voice search to dictation, a speaker's level of certainty is a primary dimension of interest. In such applications, we would like to know the speaker's actual level of certainty, but past research has only revealed listeners' *perception* of the speaker's level of certainty. In this paper, we present a method for eliciting spoken utterances using stimuli that we design such that they have a quantitative, crowdsourced legibility score. While we cannot control a speaker's actual internal level of certainty, the use of these stimuli provides a better estimate of internal certainty compared to existing speech corpora. The Harvard Uncertainty Speech Corpus, containing speech data, certainty annotations, and prosodic features, is made available to the research community.

Keywords: Perception of Certainty, Methodology for Speech Elicitation, Affect Annotation, Prosody

1. Introduction

An exciting goal in human computer interaction is that of adding human-level emotional behavior to intelligent systems, that is, the ability to perceive a user's emotional state and adaptively respond to it (Cowie et al., 2001). In speech systems in particular, there has been a lot of work in recent years on detecting a broad spectrum of affective states in speech, from core emotions (Lee and Narayanan, 2005; Fernandez and Picard, 2005; Schuller et al., 2011) to metacognitive states such as level of certainty (Liscombe et al., 2005; Pon-Barry, 2008; Forbes-Riley and Litman, 2011; Pon-Barry and Shieber, 2011) and engagement (Litman et al., 2012).

A major challenge for the field of affect recognition is the subjective nature of affect labels. The most common approach to obtaining affect labels is to measure perceived affect, as annotated by one or more human listeners. For example, in an existing corpus of Wizard-of-Oz tutorial dialogues, instances of uncertainty are labeled by a single human, the dialogue system "Wizard" (Forbes-Riley et al., 2008). Labels of *perceived certainty* are by definition subjective. We treat them as a gold standard, understanding that the subjectivity makes for challenging classification problems (Devillers et al., 2005). On the other hand, we can consider *self-reported certainty*, when speakers are asked to rate their own level of certainty.

In our prior work, we found that self-reported certainty was often lower (rated as less certain) than perceived certainty (Pon-Barry and Shieber, 2011). In the same vein, related work on interpersonal stance (e.g., friendliness, flirtatiousness) found that in conversation dyads, self-reported affect was not strongly correlated with perceived affect (Ranganath et al., 2013).

For applications in educational technology, we are most interested in knowing a student's *internal level of certainty*. Barring breakthroughs in neuroscience, a person's actual internal level of certainty cannot be determined, but in this paper, we present a data set that provides a novel and interesting proxy for internal certainty by carefully controlling the inherent difficulty of the task leading to the person's level of certainty, which we call *group task certainty*, and we compare it with self-reported and perceived certainty annotations. Our proxy for internal certainty is based upon crowdsourced judgements of handwritten image legibility. We modify an existing affective speech elicitation procedure to create speech elicitation stimuli around these images and we then collect a new corpus of uncertain speech.

2. Method

We present a methodology with two parts. First, we crowdsource human perception judgements to obtain legibility scores for images of handwritten digits (Section 2.1). Second, we create speech elicitation materials incorporating these digit images and collect speech from human subjects in our lab (Section 2.2). The key is that we can assign each image an intrinsic level of difficulty, based on the crowdsourced judgements; we assume that when participants are trying to read the digits, their internal certainty is correlated with the image's level of difficulty. For each utterance, we acquire self-reports of certainty from the subjects and perceived certainty annotations as generated by a panel of human judges.

2.1. Legibility Scores for Handwritten Digits

Here, we discuss our procedure for obtaining the set of handwritten digit images and describe a human computation approach to quantifying each image's *intrinsic ambiguity*, which represents a measure of certainty about the identity of the digit images averaged over the group of participants—a measure of *group task certainty*. We use this measure as a proxy for an individual's level of certainty. (We address the appropriateness of such a move in Section 5.) This proxy allows us to compare a speaker's self-reported certainty to the item's intrinsic level of certainty and verify whether self-reports are a reasonable proxy for internal level of certainty.

We make use of the MNIST database of handwritten digits (LeCun et al., 1998). The MNIST database contains 10,000 handwritten digit images from the United States Postal Service.

The process of selecting handwritten digits to use in the speech elicitation materials has three steps.

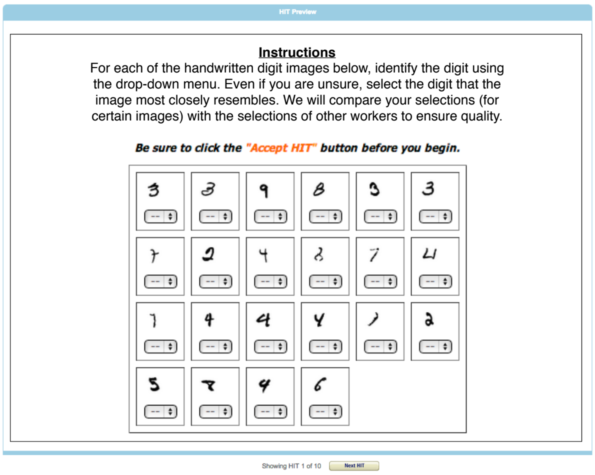
Step 1 Use an SVM classifier to identify 400 images (out of all 10,000 images) that may be difficult to read.

Step 2 Generate legibility scores for these 400 images using Mechanical Turk.

Step 3 Select 50 images (out of the set of 400) of varying legibility to use in the speech elicitation materials.

In the first step, we use an existing support vector machine classifier (Maji and Malik, 2009) to classify all the images in the MNIST database. This classifier outputs a confidence measure along with the most likely label. We select the 400 images with the lowest confidence measures to use in the subsequent step.

In the second step, we use Amazon's Mechanical Turk (Paolacci et al., 2010; Mason and Suri, 2011) to collect human judgements that we use in generating legibility scores for these 400 images. Mechanical Turk is an online labor market that facilitates the assignment of human workers to quick and discrete *human intelligence tasks* (HITs). We divided the digit images into twenty sections so that each HIT consisted of 20 images. We instructed workers to identify each digit using a drop-down menu. Each digit was labeled by 100 human workers.The full instructions and the parameters of the HIT design are available in (Pon-Barry, 2013). Figure 1 shows a screenshot of the Mechanical Turk HIT.



HIT Preview

Instructions

For each of the handwritten digit images below, identify the digit using the drop-down menu. Even if you are unsure, select the digit that the image most closely resembles. We will compare your selections (for certain images) with the selections of other workers to ensure quality.

Be sure to click the "Accept HIT" button before you begin.

Showing HIT 1 of 10

Figure 1: Screenshot of the Mechanical Turk HIT for handwritten digit classification.

Ensuring worker quality and preventing malicious behavior (e.g., bots written to complete all the HITs in a batch) is a challenge for researchers collecting data on Mechanical Turk (Callison-Burch and Dredze, 2010). We took two measures to ensure quality. First, we included a question, such as "What is 4+2?", to verify that the worker was a real person. Second, we randomly included two control images in every HIT. We verified that workers correctly identified those digits before paying them.

We generate a legibility score for each image based on the *entropy* of the human label distribution, a measure of the uncertainty of a random variable X taking on values $x_1, \dots x_N$ defined by,

$$H(X) = -\sum_{i=1}^{N} P(x_i) log P(x_i) \qquad .$$

Using the labels collected on Mechanical Turk, we can compute the maximum likelihood estimate for the probability $P(x_i)$. We take the legibility score to be $1 - H(X)$.

Accordingly, legibility scores fall in the range [0,1]. A legibility score of 1 (entropy of 0) indicates high legibility (all 100 people select the same label). We find that 36% of images were unambiguous (legibility score = 1) and the most ambiguous image has a legibility score of 0.19. Table 1 shows several digits of varying legibility, the frequencies of the human labels, and the associated entropy values and legibility scores.

In the final step, we select 50 images to use in the speech elicitation stimuli based on the entropies of the human-label distributions. We draw as uniformly as possible from the binned range of entropies.

Table 1: Handwritten digits: label frequencies and entropy.

	Crowdsourced Label Frequencies				
Label					
'0'	-	-	-	-	2
'1'	-	-	-	5	34
'2'	-	22	-	-	9
'3'	-	-	-	-	20
'4'	-	-	69	-	4
'5'	100	-	-	-	15
'6'	-	1	31	-	3
'7'	-	77	-	58	5
'8'	-	-	-	-	8
'9'	-	-	-	37	-
Entropy	0.00	0.25	0.27	0.36	0.81
Legibility Score	1.00	0.75	0.73	0.64	0.19

2.1.1. Image Ambiguity

When generating legibility scores, we assume that ambiguous images will appear ambiguous to nearly all people. To test this, we conducted a second experiment on Mechanical Turk that asked 100 people whether they found an image to be ambiguous or unambiguous. Figure 2 shows the fraction of people who rated an image as unambiguous versus the image's legibility score. The distribution confirms our hypothesis. Images with lower legibility scores (less than 0.75) were deemed ambiguous for the majority of people.

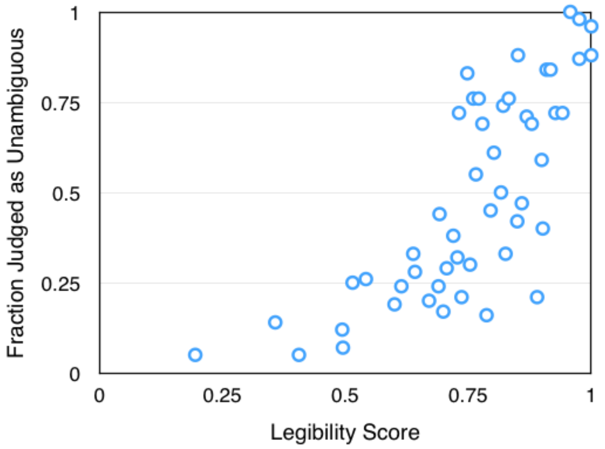


Figure 2: For each image, the fraction of people who judged it to be unambiguous vs. its legibility score.

2.2. Eliciting Speech with Varying Levels of Certainty

We collect speech data with emphasis on controlling a speaker's likely internal level of certainty as predicted by the group task certainty. The method for eliciting uncertain speech is a modification of the method used in a prior round of affective speech collection (Pon-Barry and Shieber, 2011). In that work, we did not attempt to control the speaker's internal level of certainty. As a result, there was no way to verify whether a speaker's self-reported certainty was aligned with his or her actual, internal certainty.

In the present work, the salient difference is that the speech elicitation materials are designed in a way that controls the level of certainty of the stimulus. This is achieved by asking participants to engage in a task that necessitates speaking a spontaneous utterance that incorporates reading the handwritten digits, which we controlled using the method above to exhibit widely varying degrees of legibility.

The materials for eliciting speech are designed so that participants would speak the selected MNIST digit aloud, in the context of answering a question. The handwritten digit images are embedded in an illustration of a train route connecting two U.S. cities. An example train route illustration is shown in Figure 3.

At the start of the data collection experiment, participants read a task scenario explaining why they are deciphering handwritten train conductor notes and answering questions about them. (See (Pon-Barry, 2013) for details of the task scenario.) For each train route illustration, participants are asked a single question. The participants responds aloud, speaking spontaneously. However, their word choice is influenced by a warm-up task where they are given answers to read aloud. This lets us have some influence over the length and lexical content of the utterances without the participant explicitly reading a sentence aloud. Two example questions and answers are shown below.

(1) Q: Which train leaves Los Angeles and at what time does it leave?

A: Train number 7 leaves Los Angeles at 1:27.

(2) Q: Which train arrives in Dallas and at what time does it arrive?

A: Train 2 arrives in Dallas at 9:12.

2.2.1. Procedure for Speech Elicitation

The procedure for eliciting speech and certainty self-reports is summarized below.

1. The participant sees a train route illustration.
2. The participant hears a question about the train route (while viewing illustration).
3. A beep is played, prompting the participant to answer.
4. The participant answers (while viewing illustration).
5. The participant rates his or her level of certainty on a 1 to 5 scale.

The procedure is an adaptation of the procedure described in previous work (Pon-Barry and Shieber, 2011), with two differences: (1) the questions are pre-recorded and integrated into the experimental interface, and (2) the participants answer the questions spontaneously. Twenty-two participants completed the experiment, 11 male and 11 female.

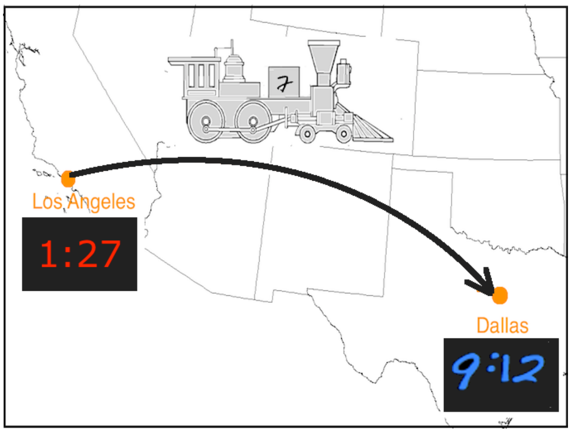


Figure 3: Speech elicitation stimulus. The handwritten digit on train was perceived as a '7' some of the time and perceived as a '2' some of the time.

3. Annotating Level of Certainty

We collect level of certainty annotations from the speaker's perspective and the hearer's perspective. This is a key distinction between our corpus and other corpora that focus on annotations of the perception of certainty. The distribution of legibility scores, certainty labels from the speaker's

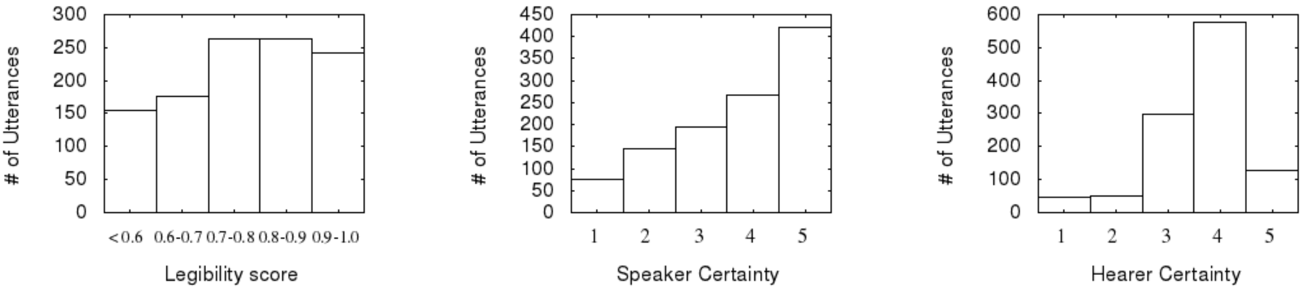


Figure 4: Three histograms: (left) distribution of difficulty scores for the stimuli that prompted the utterances in the corpus, (middle) distribution of certainty labels from the speaker's perspective, (right) distribution of certainty labels from the hearer's perspective.

perspective, and from the hearer's perspective are shown in Figure 4.

We obtain certainty labels from the *speaker's perspective*, that is, self-reported certainty, during the speech elicitation experiment. After answering a question, the speaker is asked, "How certain were you about the answer you just gave?" They indicate their certainty on a 1 to 5 Likert scale (1=very uncertain, 5=very certain).

We obtain certainty labels from the *hearer's perspective*, that is, perceived certainty, by selecting the majority among the judgements of a panel of six annotators, all native English speakers. Every annotator listened to and rated the entire set of 1100 utterances. The annotators rated level of certainty for each utterance on the same 5-point scale used for the self-reports (1 = very uncertain, 5 = very certain). They did not see any contextual information such as the handwritten images.

The agreement among the six annotators highlights the subjective nature of the hearer-centric affect labeling paradigm. Across all pairs of annotators, we find an average pairwise agreement of 54.3%, average Cohen's kappa of 0.235, and average Spearman correlation coefficient of 0.494. If we look only at the pair of annotators with the highest agreement, we see much higher values: pairwise agreement of 74.1%, Cohen's kappa of 0.407, and Spearman correlation of 0.62.

4. Harvard Uncertainty Speech Corpus

The materials described here form part of the Harvard Uncertainty Speech Corpus, which contains speech recordings, level of certainty annotations, and acoustic feature vector data. The speech elicitation materials include items from three domains: vocabulary and public transportation (described in previously published work (Pon-Barry and Shieber, 2011)), and the handwritten digits domain described here. In total, the Harvard Uncertainty Speech Corpus has 1700 utterances and 148.79 minutes of speech. The speech recordings are available upon request for research purposes. The level of certainty annotations, acoustic feature vector data, and speech elicitation materials are available for download through the Dataverse Network (`http://dvn.iq.harvard.edu/dvn/dv/ponbarry`).

There are three main benefits of the corpus. First, it contains certainty annotations from the speaker's point of view (self-reports) as well as annotations from the hearer's point of view (listener judgements). Second, the difficulty of the questions can be controlled. Third, the corpus contains several instances of specific words and phrases such as "train one" or "train two". These phrases are spoken multiple times by each speaker, with differing levels of certainty. Figure 5 shows the spectrograms of three utterances from the same speaker saying "train two" while feeling uncertain, neutral, and certain. This allows for the analysis of subtle differences in prosodic expressivity (for example, (Pon-Barry and Nelakurthi, 2014)).

5. Discussion

Our initial analysis suggests that self-reported certainty and group task certainty are more strongly correlated than perceived certainty and group task certainty. The correlation coefficient for the former is $r = 0.818$, while the latter is $r = 0.687$. Given that self-reported certainty is "closer" to a speaker's internal level of certainty than perceived certainty, this finding goes some way toward validating the assumption that the speaker's internal level of certainty is closely associated with the group task certainty.

Of course, group task certainty is not the same as internal level of certainty. First, internal level of certainty may depend on various aspects of the task, not only how uncertain the particular manipulated stimulus component of the task is for the individual. However, the particular task used here was designed so that all aspects other than interpretation of the ambiguous digit were straightforward to accomplish and did not vary in difficulty across stimuli. Second, how uncertain the stimulus component of the task is *for the individual* may differ from how uncertain the stimulus component of the task is *for the group*. For instance, in the digit identification task, it could be the case that a digit possessed high group task uncertainty (that is, a low legibility score) not because each individual participating in the crowdsourced annotation was uncertain about the digit identification, but because each individual was quite certain about identifying the digit differently. Though introspection may indicate this possibility as implausible, it cannot be ruled out; eliminating the possibility would re-

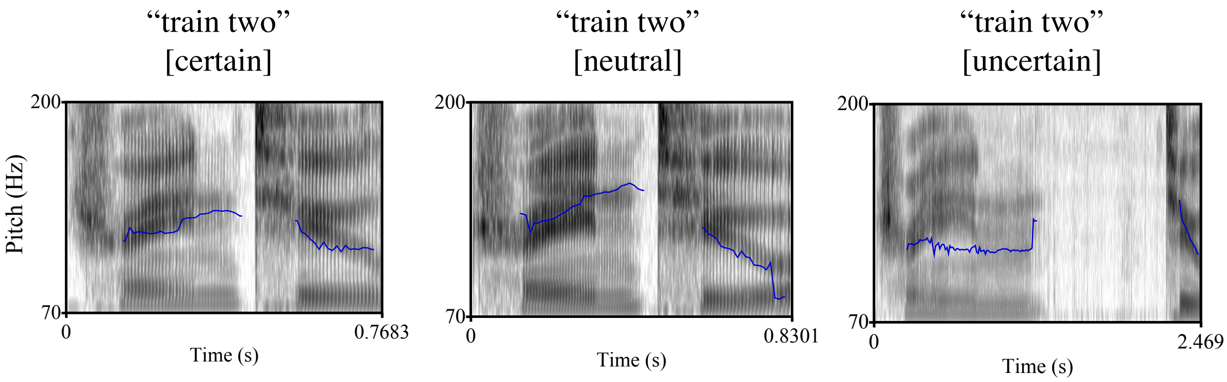


Figure 5: Three instances of a single speaker saying "train two" with varying certainty: certain (left), neutral (middle) and uncertain (right). The pitch estimate (blue line) is overlaid atop the spectrogram.

quire knowing the internal level of certainty of the annotators, thereby begging the question. Barring breakthroughs in neuroscience or parapsychology, we are unlikely to see approaches to determining true internal level of certainty. In the meantime, this new measure of certainty based on intrinsic task ambiguity may prove useful as a proxy.

This work addresses an issue central to human language technologies and affect recognition: what are the best practices with respect to measuring speaker affect and speaker state? For speaker uncertainty, there is evidence that adapting to uncertainty can improve learning, but also that accurately detecting uncertainty is a bottleneck for fully-automated adaptive systems (Forbes-Riley and Litman, 2011). We believe that a speaker's perception of certainty is the measure we ought to care about, despite the challenges associated with measuring it.

We have presented a method for acquiring, along with such self-reports and annotator labelings, information about the actual source for the level of certainty, allowing us to investigate the relationship between these external and internal types of annotation. We do so in the context of examining uncertainty, though the method may be applicable to other forms of affect as well, ones where the source of the affectual state is manipulable.

6. Acknowledgements

This work was supported in part by a National Science Foundation Graduate Research Fellowship and by a Siebel Scholarship. We thank Vanessa Tan and Spencer de Mars for their contributions to this project. We thank Luke Hunsberger for initiating our thinking about the use of digit manipulation for controlling internal level of certainty.

7. References

Chris Callison-Burch and Mark Dredze. 2010. Creating speech and language data with Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 1–12.

Roddy Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor. 2001. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. 2005. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422.

Raul Fernandez and Rosalind Picard. 2005. Classical and novel discriminant features for affect recognition from speech. In *Proceedings of Interspeech*, pages 473–476, Lisbon, Portugal.

Kate Forbes-Riley and Diane Litman. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53:1115–1136.

Kate Forbes-Riley, Diane Litman, Scott Silliman, and Amruta Purandare. 2008. Uncertainty corpus: Resource to study user affect in complex spoken dialogue systems. In *Proceedings of the 6th Language Resources and Evaluation Conference*.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

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. In *Proceedings of Interspeech*, pages 1837–1840, Lisbon, Portugal.

Diane Litman, Heather Friedberg, and Kate Forbes-Riley. 2012. Prosodic cues to disengagement and uncertainty in physics tutorial dialogues. In *Proceedings of Interspeech*.

Subhransu Maji and Jitendra Malik. 2009. Fast and accurate digit classification. Technical Report UCB/EECS-2009-159, EECS Department, University of California, Berkeley.

Winter Mason and Siddharth Suri. 2011. Conducting behavioral research on Amazon's Mechanical Turk. *Be-*

havior Research Methods, 44:1–23.

Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419.

Heather Pon-Barry and Arun Reddy Nelakurthi. 2014. Challenges for robust prosody-based affect recognition. In *Proceedings of Speech Prosody*.

Heather Pon-Barry and Stuart M. Shieber. 2011. Recognizing uncertainty in speech. *EURASIP Journal on Advances in Signal Processing*, 2011(251753).

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

Heather Pon-Barry. 2013. *Inferring Speaker Affect in Spoken Natural Language Communication*. Ph.D. thesis, Harvard University.

Rajesh Ranganath, Dan Jurafsky, and Daniel A. McFarland. 2013. Detecting friendly, flirtatious, awkward, and assertive speech in speed-dates. *Computer Speech and Language*, 27(1):89–115.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53:1062–1087.