Using Ambiguous Handwritten Digits to Induce Uncertainty

Heather Pon-Barry

School of Computing, Informatics, and Decision Systems Engineering
Arizona State University
ponbarry@asu.edu

Abstract

The lack of ground truth labels is a significant challenge in the field of automatic recognition of emotion and affect. 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. In this paper, we describe a method that uses ambiguous handwritten digits for the purpose of inducing natural *uncertainty*. Using a crowdsourcing approach, we quantify the legibility of each handwritten digit. These images are integrated into visual stimuli that are used in a lab experiment for eliciting spontaneous spoken utterances of varying levels of certainty. While we cannot measure a speaker's actual internal level of certainty, our method generates a novel and interesting approximation for internal certainty.

Keywords: Uncertainty, Methodology for Speech Elicitation, Affect Labels and Ground Truth

1. Introduction

Although significant progress has been made in recent years, the problem of automatically recognizing a person's emotional or cognitive state faces many challenges (Schuller et al., 2011). One of the main challenges is in obtaining ground truth labels for a person's emotional or cognitive state. The most common approach to obtaining labels is to measure perceived emotion, as annotated by one or more human judges. This produces labels that are by definition subjective. We treat them as a gold standard, understanding that the subjectivity makes for a challenging classification problem (Devillers et al., 2005).

In this paper, we present a method for inducing natural *uncertainty* in the context of collecting a corpus of affective speech. We use a crowdsourcing approach to identify a set of ambiguous handwritten digits and to calibrate the difficulty of deciphering each digit. The handwritten digit images are integrated into visual stimuli that are used in a question-answering lab experiment for eliciting spontaneous spoken answers of varying levels of certainty. Details on the speech elicitation, the annotation of uncertainty, and the resulting Harvard Uncertainty Speech Corpus are presented in a separate paper (Pon-Barry et al., 2014).

In previous on recognizing uncertainty, there is little control over how uncertain a person is. To obtain labels for level of certainty, researchers have utilized annotators to label perceived certainty (Litman and Forbes-Riley, 2006). In our past work, we compared perceived level of certainty to speaker self-reported level of certainty. We found that self-reported certainty was often lower (rated as less certain) than perceived certainty (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 the perceived certainty or the self-reported certainty was closer to his or her *actual* certainty.

Our interest in improving uncertainty detection is motivated by applications for personalized learning in tutorial dialogue systems, where we are most interested in knowing a student's internal level of certainty. There is evidence indicating 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). Skilled human tutors can gauge a student's level of certainty and tailor the dialogue appropriately. For example, if a student feels certain but gives an incorrect answer, it may be due to a misconception. Studies of learning in human tutorial dialogue suggest a strong connection between impasses (such as misconceptions) and student learning, to the point of proposing that *cognitive disequilibrium* is a necessary precursor to deep learning (VanLehn et al., 2003; Craig et al., 2004).

We describe in this paper a method for approximating internal certainty based upon crowdsourced judgements of handwritten image legibility. We create speech elicitation stimuli around these images that enable the creation of a speech corpus with three kinds of certainty labels: approximate internal certainty, self-reported certainty, and perceived certainty (Pon-Barry et al., 2014).

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

Our process of selecting handwritten digit images and generating legibility scores has three steps.

1. Identify 400 candidate images (out of all 10,000 images) that may have low legibility.
2. Generate legibility scores for these 400 images via crowdsourcing.
3. Narrow down set of 400 images to identify 50 images with varying legibility scores.

The following sections describe these steps in detail.

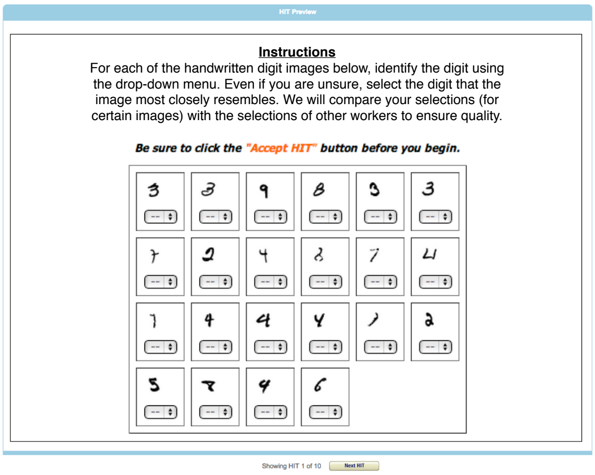
2.1. Identify Candidate Images

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. The 400 images with the lowest confidence measures are used in the crowdsourcing experiment.

2.2. Crowdsourcing Legibility Scores

In the second step, we generate legibility scores for these 400 images by crowdsourcing human labels on Amazon's Mechanical Turk. Mechanical Turk is an online labor market that facilitates the assignment of human workers to quick and discrete *human intelligence tasks*, or HITs (Paolacci et al., 2010; Mason and Suri, 2011). Our crowdsourcing approach enables each image to be labeled by 100 humans in a short amount of time.

We divide the digit images into twenty sections so that each HIT consists of 20 images. We instruct workers to identify each digit using a drop-down menu. Figure 1 shows a screenshot of the Mechanical Turk HIT. Pon-Barry (2013) includes the full instructions and experiment settings.



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 legibility scores.

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 $x_1, \ldots x_N$ 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 $1 - H(X)$.

Thus, legibility scores fall in the range [0,1]. A legibility score of 1 (entropy of 0) indicates high legibility (all 100 people choose the same label).

Table 1: Handwritten digits of varying legibility. The individual label frequencies and legibility scores are shown in the columns below each image.

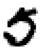
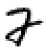
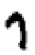
	Crowdsourced Label Frequencies				
Label					
'0'	-	-	-	-	2
'1'	-	-	-	5	34
'2'	-	22	-	-	9
'3'	-	-	-	-	20
'4'	-	-	69	-	4
'5'	100	-	-	-	15
'6'	-	1	31	-	3
'7'	-	77	-	58	5
'8'	-	-	-	-	8
'9'	-	-	-	37	-
Entropy	0.00	0.25	0.27	0.36	0.81
Legibility Score	1.00	0.75	0.73	0.64	0.19

Table 2: The distribution of legibility scores for the 400 images that were classified by human workers on Mechanical Turk.

Legibility Score s	Number of Images
$0.1 < s < 0.2$	1
$0.3 < s < 0.4$	1
$0.4 < s < 0.5$	3
$0.5 < s < 0.6$	2
$0.6 < s < 0.7$	8
$0.7 < s < 0.8$	26
$0.8 < s < 0.9$	33
$0.9 < s < 1$	181
$s = 1$	146

Table 1 shows five digits of varying legibility, the frequencies of the human labels, and the associated entropy values and legibility scores. Table 2 shows the frequency of legibility scores for the 400 images that were classified by workers on Mechanical Turk.

Ensuring Quality. Preventing malicious behavior (e.g., artificial bots designed to complete all the HITs in a batch) is a challenge for researchers collecting data on Mechanical Turk (Ipeirotis et al., 2010; Callison-Burch and Dredze, 2010). We take two measures to ensure worker quality. First, we include a question, such as "What is 4+2?", to verify that the worker is a real person. Second, we include two control images in every HIT. Before paying workers, we verify that they correctly identify the control images.

Experiment Running Time. Our Mechanical Turk experiment was staged in two rounds, with 10 unique HITs per round. Round 1 took 126 hours (about five days) to complete with an average time/HIT of 72 seconds. Round 2 took 33 hours (about one and a half days) to complete, with an average time/HIT of 61 seconds.[1]

2.3. Narrow Down Set of Images

In the final step, we identify 50 images to use in the speech elicitation stimuli based on the entropies of the human-label distributions. We drew uniformly (as uniformly as possible) from the binned range of legibility scores. The resulting set of 50 images is shown in Figure 2. The images are displayed from easiest to hardest (low entropy to high entropy) starting from the top-left and moving left-to-right across the rows.

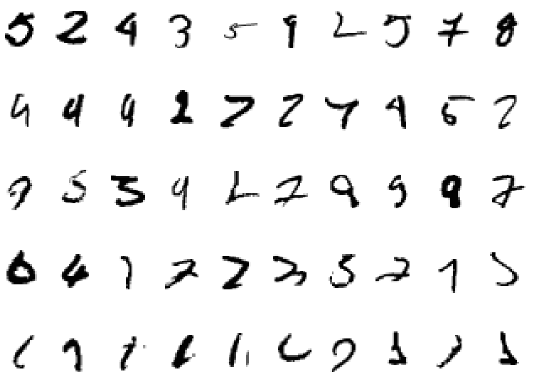


Figure 2: Handwritten digit images of varying legibility, ordered from easiest to hardest.

2.4. 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 3 shows the fraction of people who rated an image as unambiguous versus the image's legibility score. The distribution confirms our hypothesis. Images found unambiguous by a majority of people all have legibility scores in the upper range (greater than 0.75).

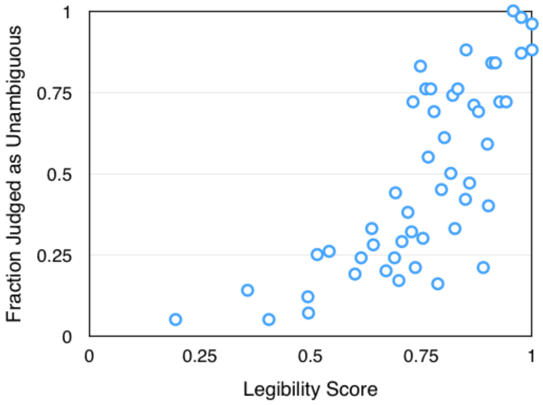


Figure 3: For each image, the fraction of people who judged it to be unambiguous vs. its legibility score.

3. Integrating Images into Stimuli

The materials for eliciting speech are designed so that participants utter a specific digit aloud in the context of answering a question. The handwritten digit images are embedded in an illustration of a train route

[1] The two experiment rounds were identical in all ways except for the images themselves. We speculate that Round 2 took less time than Round 1 due to the time of posting, i.e., weekday vs. weekend.

connecting two U.S. cities. The handwritten digit indicates the train number. An example train route illustration is shown in Figure 4. The handwritten digit on the train was identified as a ‘7’ by 76 people, as a ‘2’ by 22 people, and as a ‘6’ by 2 people.

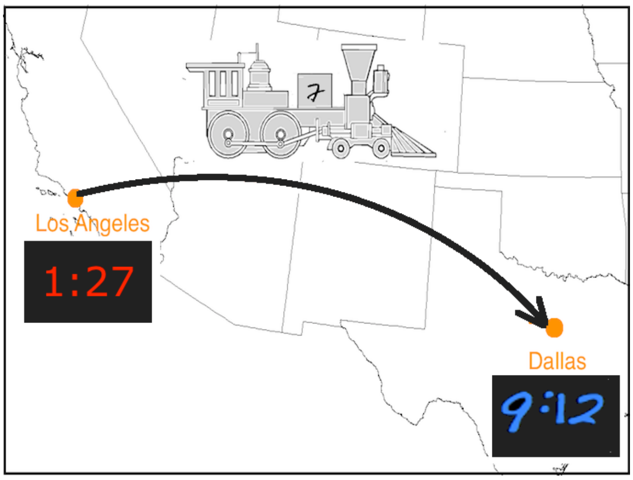


Figure 4: Speech elicitation stimulus integrating an ambiguous handwritten digit indicating the train number.

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. A question that requires reading the train number is asked and participants respond spontaneously. For example:

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

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

Although the question responses are spontaneous, word choice is influenced by a warm-up task where participants are given answers to read aloud. This lets us have indirect influence over the length and lexical content of the utterances, which aids future analysis of utterance-level and word-level prosody.

The key point is that we can assign each image a legibility score, based on the crowdsourced judgements. We assume that when participants are trying to read the digits, their internal certainty is proportional to the image’s legibility score. We compare two kinds of certainty labels to these legibility scores: labels from the speaker’s perspective and labels from the hearer’s perspective. The former, labels from the speaker’s perspective, are more strongly correlated with the legibility scores (Pon-Barry et al., 2014).

4. Harvard Uncertainty Speech Corpus

The results of our Mechanical Turk experiment and speech elicitation stimuli are available to the research community through the Dataverse Network.[2] At this site, researchers can also access the level of certainty annotations, acoustic feature vector data, and request access to the audio data. Details on the Harvard Uncertainty Speech Corpus can be found in previous and concurrent published works (Pon-Barry and Shieber, 2011; Pon-Barry et al., 2014).

5. Discussion and Conclusion

In this paper, we introduced a novel method for approximating internal certainty based upon crowdsourced judgements of handwritten image legibility. We collected affective speech in a controlled experiment in a laboratory setting that utilized these images. This allowed us to analyze subtle differences in prosodic expressiveness to better understand individual speaking styles (Pon-Barry and Nelakurthi, 2014). However, there are limitations associated with speech collected in a lab. Integrating these images into new experiments to collect spontaneous affective speech in real-world learning and tutorial environments is an exciting avenue for future research.

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? We have presented a method for identifying ambiguous handwritten digits for the purpose of inducing natural uncertainty and we used crowdsourcing to generate a legibility score for each handwritten digit. While crowdsourcing has been used as a way of obtaining labels for a given audio or video segment, we claim that it also has utility in designing stimuli for inducing natural affect. Our work is done in the context of examining uncertainty, though the method is 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 the Mechanical Turk experiment, and Nicholas Longenbaugh and Stuart Shieber for their contributions to the development of the Harvard Uncertainty Speech Corpus.

[2] http://dvn.iq.harvard.edu/dvn/dv/ponbarry

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.

Scotty D. Craig, Arthur C. Graesser, Jeremiah Sullins, and Barry Gholson. 2004. Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media*, 29(3):241–250.

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.

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.

Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 64–67. ACM.

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.

Diane Litman and Kate Forbes-Riley. 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech Communication*, 48(5):559–590.

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. *Behavior 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, Stuart M. Shieber, and Nicholas Longenbaugh. 2014. Eliciting and annotating uncertainty in spoken language. In *Proceedings of the 9th International Language Resources and Evaluation Conference (LREC)*.

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

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.

Kurt VanLehn, Stephanie Siler, Charles Murray, Takashi Yamauchi, and William B. Baggett. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):209–249.