A METHODOLOGY FOR USING CROWDSOURCED DATA TO MEASURE UNCERTAINTY IN NATURAL SPEECH

Lara Martin[*1]*, Matthew Stone*[2]*, Florian Metze*[1]*, Jack Mostow*[1]

[1]Carnegie Mellon University, Language Technologies Institute, Pittsburgh, PA
[2]Rutgers University, Department of Computer Science, New Brunswick, NJ

ABSTRACT

People sometimes express uncertainty unconsciously in order to add layers of meaning on top of their speech, conveying doubts about the accuracy of the information they are trying to communicate. In this paper, we propose a methodology for annotating uncertainty, which is usually a subjective and expensive process, by using crowdsourcing. In our experiment, we used an online database which consists of colors that more than 200,000 users have named. Based on the amount of unique names that users have given each color, an entropy value was calculated to represent the uncertainty level of the color. A model, which performed better than chance, was created to predict whether or not the color that the participant was describing was ambiguous or borderline, given certain prosodic cues of their speech when asked to name the color verbally. Using crowdsourced data can greatly streamline the process of annotating uncertainty, but our methods have yet to be tested in other domains besides color. By using methods such as ours to measure prosodic attributes of uncertainty, it should be possible to increase the accuracy of voice search.

***Index Terms*—** uncertainty, speech analysis, color, annotation, crowdsourcing

1. INTRODUCTION

Human beings use a complicated system of verbal language that encompasses much more than which words we choose (semantics) and what order they belong in (syntax). Conversation is a process of turn-taking. In situations where a person is explaining something, the listener has to show his understanding through back-channel responses, such as "Is this right?" [1]. In turn, the speaker either confirms or rejects the listener's interpretation of what was originally said. Back-channel responses are often marked with an upwards intonation at the end of the word or phrase, called a "try" marker, as in a question, which is used in order to gain acceptance from the speaker [1].

Theory of Mind (ToM), someone's beliefs and desires, can be deduced from their actions [2]. A principle of rational action can be created that demonstrates what an agent wants and believes, and makes the agent choose the best result given the amount of effort that is needed to reach that result. Figuring out how well someone else knows something is known as the Feeling of Another's Knowledge, or FOAK [3]. In speech, people tend to use prosodic cues to reveal how sure they are about what they are saying, and listeners can usually hear these. If a person is answering a question and gives a rising intonation and shows latency, the listener's FOAK will decrease for this person, meaning that the listener does not believe the speaker knows what he or she is talking about. On the other hand, when a speaker uses filled pauses, such as "um" and "uh", the longer the speaker hesitates, the greater the FOAK [3]. It is hypothesized that the FOAK increases for latency in filled pauses since it appears as if the speaker knows the answer but is having trouble accessing it in memory [3].

Human tutors have the ability to know when their student is uncertain. A tutoring system was created using a C4.5 decision tree that was based off of breath groups (segments of speech between two pauses) and the number of turns in the dialogue and was able to achieve 76.42% accuracy [4]. Cues of uncertainty have been used in search techniques where people could not correctly recall the full name of an item, but could produce the letter it begins with. With this input, the computer can find the answer far more accurately than if they were to say the incorrect name [5]. Other, similar experiments have been done to computationally measure uncertainty with reasonable success [6]. Following similar methods to those of Brennan & Williams, Pon-Barry & Shieber used framed prompts, such as giving the user a series of multiple choice questions, to elicit uncertainty. Indeed, in answers where speakers hold low levels of Feeling of Knowing or FOK (they are not confident in their own answers), long delays, filled pauses, and high intonation have been found [7].

In order to use uncertainty in automated systems such as in the work of Paek et al. (2008), we need a way of automating the process of determining whether or not a question is deemed difficult or not. Programs in the past have had some success in incorporating said prosodic cues, but are generally limited in their capabilities. They usually depend on self-reports of uncertainty which are intrusive to obtain, potentially unreliable, and domain-dependent. Unlike the work

[*]The first author performed the work while at Rutgers University

above, which used self-reporting and perceived uncertainty, in our study, uncertainty was derived from the topic's data through measuring the variation in responses. This creates more options for data sources, particularly when it is a topic in which uncertainty has not been researched extensively. We approach this problem through use of a crowdsourced population's uncertainty levels, which we then use as our training labels for inferring an individual's uncertainty.

Does the amount of uncertainty a person expresses correlate with the ambiguity of the color they are trying to describe? Since color is a limited domain (there are only so many colors that we can see) that everyone is familiar with but can allow for a wider variety of uncertainty than a set of discrete items with a well-defined nomenclature, it is a good test for the current initial study. The more borderline a color is (a color found in between two "standard" colors, eg. yellow-orange), the more uncertain the color is considered. People describing more uncertain colors should use these cues of uncertainty to express their doubt, and we can use our automated system of calculated uncertainty in order to simulate FOAK.

2. METHODS

Participants were shown colors, one by one, with their speech being recorded while the color was displayed. The participants were told that they were going to be shown a series of colors and were asked to verbally state the name of the color they saw within 6 seconds. The participants were also informed of the sequence of button presses they had to perform (outlined in the next paragraph). After the instructions were given, there was a practice session in which participants were shown 5 randomly-chosen colors in a setup identical to the actual experiment. During this time and immediately afterward, participants were allowed to ask questions about how to do the experiment. The recordings from the practice were not used.

A color was presented as a 2x2-inch square in the middle of the computer screen, surrounded by a pure white background. Next to the display was a button that was labeled "Stop". If the participant did not click the "Stop" button after 6 seconds, it was programmatically pressed. After the recording stopped, the button changed its text to say "Next". This was to prevent the experiment from continuing on without the participant being ready. This button had to be manually clicked in order to proceed with the experiment. The recording started as soon as the color was displayed on the screen. The investigator was outside of the room throughout the duration of the experiment, with the exception of the practice session.

Participants were shown 60 predetermined RGB-value colors from two classes (Tables 1 and 2) presented one at a time in a random order. Colors were selected from an online color survey by the webcomic XKCD where 222,500 users were shown various colors and asked to type the name [8]. Based on this database, Shannon entropy values were calculated to measure the uncertainty of the color based on the probability of a name occurring for a given RGB value. The colors were chosen so that 30 colors were of high entropy and 30 colors were of low entropy, creating a binary classification problem. A color is considered low entropy if there is very little variation in the names that people have given it, such as a pure red or blue. A high entropy color has more variation in its names, and the color itself is more borderline. For example, one of these high entropy colors in the experiment was called "uh tannish beige", "gray green", and "sandy green moss" by three different participants.

The experiment was run on a Samsung Ultrabook in a quiet, standard classroom setting. Seventeen participants (9 female and 8 male) volunteered from Rutgers psychology and linguistics subject pools. They were American native English speakers, claiming not to be colorblind. One additional participant was discarded since the person was not a native English speaker.

Audio was recorded using a CAD U2 USB Stereo Headphones with Microphone, but the subjects were told not to put the headphones on their ears, so that they could hear themselves talk. The monophonic audio was saved with a 44,000Hz sample rate and 16-bit sample size.

3. RESULTS

All results were calculated using Praat. Pitch was hand-annotated by the first author using Tones and Break Indices (ToBI) conventions for English. Results from ToBI were evaluated using only the final tones with 0 as shorthand for L-L%, 1 for H-L%, 2 for L-H%, and 3 for H-H%. To gather information on the hesitation of the speaker, a Praat script called "Praat Script Syllable Nuclei" was used to find the nuclei of each syllable (voiced peaks) [9]. The script found: the number of syllables, the number of pauses (where the sound levels were lower than a specified threshold for at least 0.3 seconds), the total duration of the utterance (in seconds), the length of phonation (how long the person was actually speaking for, in seconds), the speech rate (number of syllables / duration), the articulation rate (number of syllables / phonation time), and ASD (phonation length / number of syllables). The length of the filled pauses was not taken into account such as in the work by Brennan et al. [3]. In addition to these features, the script was modified to calculate initial pause duration (the number of seconds from the beginning of the recording until the first voiced location was found). The number of filled pauses (counted by hand) and the gender of the participant were also taken into consideration. Practice answers were not analyzed. Some utterances were discarded because the speaker did not finish talking in time or did not say anything within the allotted 6 seconds. As a result, we collected a total of 488 utterances while a low-entropy color was displayed and 468 when a high-entropy color was displayed.

Table 1. List of low-entropy RGB colors used in experiment and their entropy values.

Color	R	G	B	Entropy
	0	0	0	0.27539
	230	0	10	0.32522
	240	240	40	0.39506
	240	120	30	0.40559
	210	0	0	0.40889
	130	130	130	0.45110
	90	50	10	0.46779
	170	170	170	0.47696
	110	110	110	0.51228
	230	240	0	0.58083
	10	30	230	0.59096
	200	10	10	0.61332
	240	110	20	0.62190
	120	80	30	0.63549
	0	0	10	0.64179
	10	150	40	0.68038
	120	20	150	0.73436
	120	70	10	0.73441
	110	0	120	0.81534
	20	90	210	0.81936
	10	30	190	0.83575
	70	170	40	0.83812
	240	240	70	0.84079
	150	30	160	0.84271
	0	180	0	0.86676
	240	110	180	0.87548
	240	140	0	0.99733
	10	200	70	1.00116
	140	10	210	1.00308
	240	90	150	1.01758

Table 2. List of high-entropy RGB colors used in experiment and their entropy values.

Color	R	G	B	Entropy
	100	100	10	2.99391
	150	180	0	3.00584
	150	150	190	3.01965
	170	60	30	3.03180
	120	230	170	3.10680
	230	240	190	3.13145
	170	130	140	3.13320
	150	50	70	3.13367
	80	140	120	3.13368
	80	90	130	3.15574
	180	210	130	3.16946
	110	110	150	3.18806
	130	170	150	3.22612
	190	150	40	3.22841
	150	150	90	3.24022
	200	190	90	3.24460
	200	170	170	3.25080
	50	80	80	3.28384
	90	130	130	3.29242
	180	160	30	3.29425
	200	90	60	3.29685
	130	70	80	3.35728
	130	130	70	3.35742
	190	200	110	3.35804
	170	160	0	3.38156
	190	90	80	3.41413
	160	150	50	3.42388
	160	200	180	3.43635
	180	110	110	3.48140
	180	190	80	3.58763

We measured these features since they were shown to express uncertainty. Our goal is to see if our entropy values are an accurate representation of uncertainty in the color-naming domain; therefore other features which can be used to illustrate uncertainty in speech could be used instead.

A six-second time window was specified in order to limit the participants to one-word responses, avoiding "descriptions" of the color instead of just stating its name, as participants have done during the pilot. By giving this constraint, we are possibly eliminating non-answers which would give the opposite correlation and would have confused results since we were not noting the words being said [7].

After initial tests with various models, we implemented a naive Bayes tree (a tree with different Naive Bayes classifiers at the leaves) created by Weka [10], training and testing it using 10-fold cross validation. The results of the Naive Bayes tree model showed that articulation rate, the gender of the participant, and the number of filled pauses were the most significantly related to entropy (see Table 4). Based off of the results from the Naive Bayes classifiers, having a short duration is highly indicative of low entropy (88% of the instances that reach this Naive Bayes classifier are of low entropy), as is the same with long duration and high entropy (72%). Also, a longer initial pause ($>$ 1.8 seconds) makes it likely that it is high entropy (71%) and a ToBI value of L-L% will most likely result in lower entropy (75%). The rest of the Naive Bayes classifiers were more evenly split between the two classes, resulting in a lower overall accuracy rate.

The Naive Bayes tree appears to have been fairly successful in predicting entropy, correctly classifying 65.27% of the instances. This means that the model is performing above the baseline mark of 50%, where the model would be choosing randomly whether the color was of high or low entropy. Fisher's exact test was performed on the data, producing a two-tailed P value $<$0.0001 (see Table 3). Without incorporating gender, the naive Bayes tree performed at 62.34% accuracy with P $<$0.0001, relying on articulation rate, phonation time, and the number of filled pauses.

Table 3. Confusion matrix of high and low entropy classifications produced from the alternating decision tree, where "hyp" is the hypothesized value, and "ref" is the reference value.

	Hyp-Low	Hyp-High	Total
Ref-Low	**346**	142	488
Ref-High	190	**278**	468
Total	536	420	

Table 4. Naive Bayes tree classifier using 10-fold cross validation for finding entropy, where "NB" stands for each leaf's individual Naive Bayes classifier and "IPL" stands for initial pause length. After analyzing these series of features, the tree will point to which Naive Bayes classifier to use in order to decide the entropy value.

$articulationrate <= 0.9697$
| $gender = M$
| | $numberoffilledpauses <= 0.5$
| | | $numberofsyllables <= 1.5$
| | | | $IPL <= 1.8$
| | | | | $ToBI = 0 : NB$
| | | | | $ToBI = 1 : NB$
| | | | | $ToBI = 2 : NB$
| | | | | $ToBI = 3 : NB$
| | | | $IPL > 1.8 : NB$
| | | $numberofsyllables > 1.5$
| | | | $duration(s) <= 2.44 : NB$
| | | | $duration(s) > 2.44 : NB$
| | $numberoffilledpauses > 0.5 : NB$
| $gender = F$
| | $numberoffilledpauses <= 0.5 : NB$
| | $numberoffilledpauses > 0.5 : NB$
$articulationrate > 0.9697$
| $phonationtime(s) <= 0.915$
| | $gender = M$
| | | | $duration(s) <= 3.565 : NB$
| | | | $duration(s) > 3.565 : NB$
| | | $numberofsyllables > 2.5 : NB$
| | $gender = F : NB$
| $phonationtime(s) > 0.915$
| | $numberoffilledpauses <= 0.5 : NB$
| | $numberoffilledpauses > 0.5 : NB$

4. DISCUSSION

We were able to create a classifier which operated above chance using prosodic features that have been shown to express uncertainty in previous research [3, 6], demonstrating that using our methods for extracting classes from crowdsourced data can work for binary classification of uncertainty.

Duration, initial pause, phonation time, and speech rate might be noteworthy characteristics of hesitancy, which can account for uncertainty (as seen in the tree in Table 4). Accordingly, these factors made significant contributions to the calculation of entropy in the current study. Other factors of speech are not as directly related to uncertainty and could either be artifacts or be conveying other meanings besides uncertainty.

These hesitancy cues could also indicate the difficulty, rather than the uncertainty, of the task. Perhaps participants

take longer because of the difficulty of coming up with a name for the color. Intonation might be a better indicator of uncertainty.

Subjects demonstrated the try tone on certain colors, yet there were subjects that did not use it at all and maintained either a monotone (H-L%) or a listing (L-H%) intonation. Either the cues for uncertainty that were examined might not hold universally for all American English speakers or colors are not difficult enough to elicit uncertainty.

As in the work of Swerts & Krahmer, upwards intonation–or in their case, high intonation–did not play a large role [7]. One reason why gender might be a factor is that men and women have been known to perceive color differently [**?**] and some women can perceive more colors than men can [**?**]. Another reason could be, as Munroe noticed in the data he collected, "women were slightly more liberal with the modifiers" when naming the colors [8]. Perhaps the number of syllables or the length of the utterance did not matter as much in the distinction between borderline and definite colors since women were more likely to use more adjectives.

5. FUTURE RESEARCH

Future participants should be tested for colorblindness in case this has an effect on the results. The effect of colorblindness in studies of uncertainty expressed in prosody is still unknown. It would be interesting to test people who are certain they are colorblind and see if the correlation still exists. Such a finding would be particularly important if this work is to be developed in a real-life application where people might or might not be colorblind. Regardless, color is a medium in this experiment and not the main focus. The focus is uncertainty. Future research will see if the model is flexible enough to work with other media, such as animals, which is a much larger category.

For future experiments, the entire process can be automated and used in conjunction with a voice search program such as in the work of [5]. The model could be inserted so that it could be computed automatically whether the current object being described is high or low entropy. It is expected that measuring these prosodic attributes of natural speech should be useful in increasing the accuracy of voice search by changing the range in which the computer searches for answers in its database.

In the same vein as the experiments performed by Swerts and Krahmer, a second experiment should be done to verify whether or not the uncertainty classes created from our entropy values can correspond to the prosodic cues of uncertainty to an unbiased listener. New participants would listen to the recordings and rate their FOAK, seeing if they can properly correspond to the entropy rating given to the color, or perhaps the participants could even be able to guess which color the speaker was originally naming.

It would also be beneficial to investigate other languages besides American English. The use of pitch is expected to change, perhaps making it more significant, but hesitancy would most likely remain fairly consistent. Regardless, different prosodic cues should be experimented with, such as the 76 various spoken features extracted for dialog systems in work by Ward & Vega [11]. The accuracy of our experiment is significantly better than chance, and with more or better features, our methodology could give an alternative when annotating for uncertainty research in speech.

6. REFERENCES

[1] Herbert H. Clark and Deanna Wilkes-Gibbs, "Referring as a collaborative process," *Cognition*, pp. 1–39, 1986.

[2] Chris L. Baker, Rebecca R. Saxe, and Joshua B. Tenenbaum, "Bayesian theory of mind: Modeling joint belief-desire attribution," 2011.

[3] Susan E. Brennan and Maurice Williams, "The feeling of another's knowing," *Journal of Memory and Language*, pp. 383–398, 1995.

[4] Jackson Liscombe, Julia Hirschberg, and Jennifer J. Venditti, "Detecting certainness in spoken tutorial dialogues," 2005.

[5] Tim Paek and Yun-Cheng Ju, "Accommodating explicit user expressions of uncertainty in voice search or something like that," *International Speech Communication Association*, 2008.

[6] Heather Pon-Barry and Stuart M. Shieber, "Recognizing uncertainty in speech," *EURASIP Journal on Advances in Signal Processing*, November 2010.

[7] Marc Swerts and Emiel Krahmer, "Audiovisual prosody and feeling of knowing," *Journal of Memory and Language*, June 2005.

[8] Randall Munroe, "Color survey result," XKCD: http://blog.xkcd.com/2010/05/03/color-survey-results/, May 2010.

[9] Ton Wempe and Nivja de Jong, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, no. 2, pp. 385–390, 2009.

[10] University of Waikato, "Weka (waikato environment for knowledge analysis) [computer program]," July 2014.

[11] Nigel G. Ward and Alejandro Vega, "A bottom-up exploration of the dimensions of dialog state in spoken interaction," *13th Annual SIGdial Meeting on Discourse and Dialogue*, 2012.