Using Prosodic Cues from Uncertainty in Natural Speech to Help Voice Recognition

Lara Martin[1] and Matthew Stone[2]
Language Technologies Institute Student Research Symposium, August 21st 2013
[1]Carnegie Mellon University, [2]Rutgers University

Introduction

Prosody, the melodic structure of speech, can be used to convey meaning besides what is being said lexically. Uncertainty is realized in a number of prosodic features, such as pauses in speech, speech rate, rising intonation at the end of a phrase (try tone), and filled pauses (words like “um” and “uh”).

Questions:
Based on this, can there be a computational model that can use the prosodic cues to predict what someone is talking about?
In particular, **does the amount of uncertainty a person expresses correlate with the ambiguity of the color they are trying to describe?**

Method

XKCD has produced a large color survey that has collected from readers the names of certain colors (Munroe 2010). For this reason, color was chosen as the medium.
Participants included 18 American native English speakers (10 female, 8 male) who stated that they were not colorblind.
Participants were shown 60 predetermined RGB colors presented in a random order and given 6 seconds to state the name of the color. The sets of colors were either of high or low entropy.

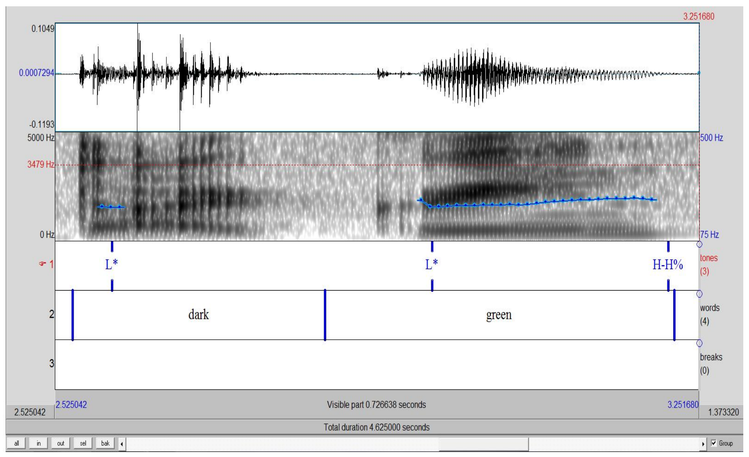

Figure 1: Participant saying “dark green” with try tone.

Data was examined using Praat. A script was used to automate pause and speech rate calculations (Wempe & de Jong 2008). Intonation was measured by hand using ToBI.

Results

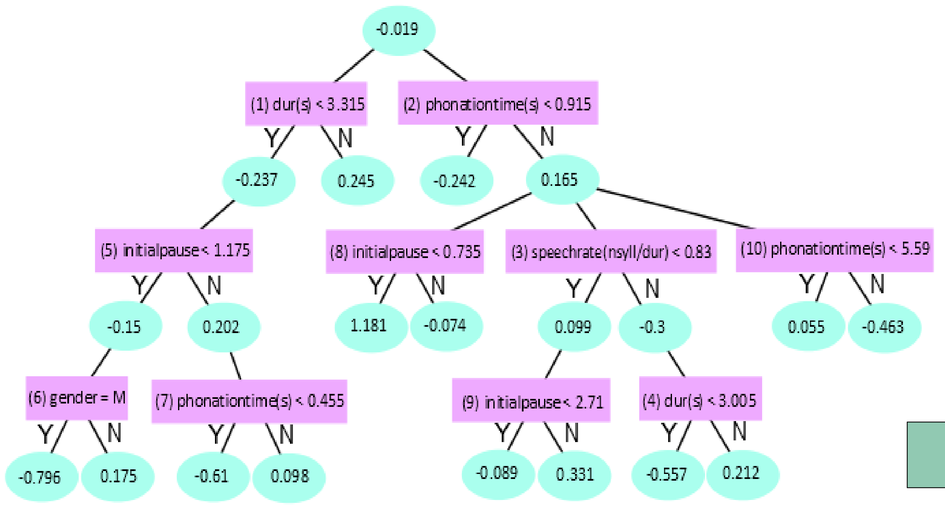

Figure 2: Alternating decision tree model solving for entropy with 10-fold cross-validation.

The model was able to perform above baseline, correctly guessing the entropy category 62.5% of the time. The features involved in analysis were: the duration of the file, the initial pause before the subject started talking, the amount of time the participant spent talking, the speech rate, the number of filled pauses, the final tones, and the gender of the participant. All of these parameters were included in the model except filled pauses and final tones, which were used by participants, but were deemed insignificant due to machine learning techniques.

References

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian Theory of Mind: Modeling Joint Belief-Desire Attribution.
Brennan, S. E., & Williams, M. (1995). The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to Listeners about the metacognitive States of Speakers. Journal of Memory and Language, 383-398.
Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a Collaborative Process. Cognition, 1-39.
Liscombe, J., Hirschberg, J., & Venditti, J. J. (2005). Detecting Certainnes in Spoken Tutorial Dialogues.
Munroe, R. (2010, May 3). Color Survey Results. Retrieved from XKCD: http://blog.xkcd.com/2010/05/03/color-survey-results/.
Paek, T., & Ju, Y.-C. (2008). Accommodating Explicit User Expressions of Uncertainty in Voice Search or Something Like That. International Speech Communication Association.
Wempe, T., & de Jong, N. H. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods, 41(2), 385-390.

Discussion

Duration, initial pause, phonation time, and speech rate might be noteworthy characteristics of uncertainty. Accordingly, these factors made significant contributions to the determination of entropy in the current study.

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.

It was found that some speakers are more likely to insert filled pauses than others, regardless of how certain they are. An explanation of why filled pauses were not significant could be that the length of the filled pauses themselves was not taken into account such as in the research by Brennan & Williams (1995).

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.

Future Research

Future participants should be tested for colorblindness in case this has an effect on the results. Regardless, color is a medium in this experiment and not the main focus. Future research will see if the model is flexible enough to work with other media, such as names of animals, which is a much larger category.

For future experiments, the Praat analysis will be automated and used in conjunction with voice search. The model will be inserted so that the computer can decide 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. When the system is integrated into search, it would be beneficial to investigate other languages besides American English.