Using Prosody to Facilitate Learning

Avi Block, Eric Chen, & Matt Shenton
Tufts University

Abstract

The following paper proposes a project that investigates how understanding prosody could aid a robot in improving its ability to complete a task efficiently. Prosody is the characteristics of speech that are not related to the actual words being said, such as voice pitch, volume, stressed syllables, pauses, and breaths. Prosody carries meaning just as words do, and we are interested in programming a robot who is able to distinguish encouraging prosodic cues from discouraging prosodic cues. After the robot can properly discriminate between the prosodic categories, we will use its understanding of prosody to teach it to play a game. By feeding it encouraging or discouraging audio input , the robot will learn proper strategies to beat the game. We will start with a simple game, and once it has mastered that, move on to a more complex game. We will judge the efficacy of the classifier by measuring how often it correctly classifies encouraging and discouraging tones, with ideal success considered above 80 percent accuracy rate. In order to determine if the machine is in fact using its understanding of prosody to improve its performance at a game, we will compare its scores with our algorithm that includes prosodic understanding to a standard learning algorithm, and establish if there is a significant improvement from the former to the latter.

Introduction

Speaking with other humans is the fundamental part of social interaction. Humans have sophisticated languages that allow us to convey complex thoughts and emotions in a way that no other species can. However, while a staple of communication between humans is words and their meanings, there are several other essential aspects of communication that convey meaning just as words do, such as nonverbal cues like body language and facial expression, as well the interest of this project, prosody. Prosody is described by Yale researchers Elizabeth Kim, Kevin Gold, & Brian Scasselati (2008) as “the music of language”. In other words, it refers to the way we say the words we say, such as tone, volume, duration, and selective pauses. In considering potentially intelligent robots, it is of significant interest, as prosody carries implied meaning that can be key to understanding interactions between humans, which will be discussed below. If a robot can learn to classify speech and commands by characteristics such as tone, pitch, etc., it can use prosodic cues to drastically improve its understanding of spoken language, and in turn the efficiency and efficacy of its learning.

In order to investigate the capability of a robot to understand prosodic cues, and adjust their own behavior accordingly, the present project will have two distinct phases. The phases will be briefly introduced in this section, and explained in further detail later on in the paper. In the first stage of the project, we will attempt to train a robot to be able to binarily classify speech by encouraging prosodic properties, and discouraging prosodic properties. In other words, we will feed data to a robot, and based on the prosody of the speech, the robot will determine if the speech is meant to be encouraging, or if the speech is meant to be discouraging.

After the robot has been trained to accurately classify the two aforementioned prosodic categories, the next phase of the project will be to improve its performance on a task, given

encouraging or discouraging audio input. With encouragement given by us, the researchers, the machine will learn the proper strategies to effectively and efficiently succeed in the games.

Relevant Work

To our knowledge, no significant previous work has been done in attempting to apply prosodic understanding to task efficacy. However, there has been significant research done into the properties of prosody, prosody's importance in human development and interaction, as well as potential applications of its understanding to intelligent, autonomous robots. Prosodic properties in speech convey meaning about the mood or condition of the speaker, as well as implied intentions of the speech that can be completely separate from the words spoken (Scasselati et al., 2006). But how does prosody manifest itself in speech? Prosodic properties differ by language, but in english, people will stress one word in each delivery of speech, otherwise known as a "pitch accent". This accented word, often done naturally and unconsciously, is thereby branded as of greater importance than the other words, a concept that is intuitively understood by the listener (Kim et al., 2008). To illustrate one example of prosody carrying implied meaning, a certain type of higher-voiced pitch accent can indicate to a listener they are being told new information, and they should store the information in their memory. Likewise, a type of lower-voiced pitch can indicate that the speaker expects the listener to already know the information they are being told (Kim et al., 2008).

Relevant work to developmental robotics and this project has been done to underscore how applicable prosody can be to creating intelligent robots. There is evidence, as proposed by Annett Schirmer (2010), that people implicitly associate word positivity/negativity with the prosodic properties related to when they heard the word. That is to say, humans can intuitively understand intended prosodic cues, and encode them along with the word. This is of great interest to our project, as we would like our game-playing robot to essentially do the same thing: recognize the prosodic properties of an encouraging phrase, and understand that phrase is meant to encourage it in the future. Furthermore, Kim et al., (2008) explain how prosodic understanding could be essential to reducing or eliminating errors in learning. They use the example of a person or robot mishearing a word as "abul", but since the prosodic characteristics with which it was spoken suggest the person or robot should already be familiar with the word, they can correctly identify the meaning to be "apple", not "abul".

Classifying prosody and its subproperties is not completely straightforward, as it is difficult to empirically measure. For the purposes of this project, we will be classifying prosody into only two groups (encouraging/discouraging), but looking at others attempt to define more complex classifications illuminates some of the properties that our robots classification will be based on. One such attempt was the Tones and Break Indices (TOBI) standard. TOBI, in essence, was an attempt to mirror for prosody what the International Phonetic Alphabet (IPA) is to phonology. TOBI consists of three tiers, each with symbols to represent different events within the tiers. The first tier refers to tonal differences in speech, such as pitch and voice loudness, the second tier refers to measuring differences in breaks or pauses between words, and the third tier, called the "miscellaneous tier" refers to other aspects of prosody such as "hesitations, disfluencies, breaths, laughs, false starts or restarts, and other spontaneous speech effects," (Silverman et al., 1992). In moving towards our specific project, understanding the

related work to the importance of prosody, its properties, and previous attempts at classification will be essential to properly training our classifier, and effectively using its knowledge of prosody to effectively play a game.

Problem Formulation and Technical Approach

In our attempt to actually implement a solution to the problem we have posed, the technical formulation and outlining of the step-by-step approach we intend to take has been split into two distinct phases, the first being the affect detection training phase, which will be followed by the affect integration phase.

The first problem we aim to tackle deals with training a robot to receive raw audio data of a human speaker and identify the speaker's affect as one of two categories, either encouraging or discouraging. The formulation of this problem is fairly simple, as it reduces to a problem of basic classification, using the principles of supervised learning in order to gain the ability to correctly identify a speaker's affect through the prosodic elements of their speech. The data for this problem will be comprised of a corpus of raw audio files containing utterances which are intended to convey either encouraging or discouraging affect. The model will then take this audio data and decrease its dimensionality down to the smallest feature space possible which still allows for classification into the two target clusters. This decrease in feature space will rely on research done on which elements of prosody are most directly correlated with emotional affect. For example, the research done on prosodic classification by Kim & Scasselati (Learning, n.d.) reduces audio data down to the measurements of mean pitch over the duration of the utterance and energy range, and calculates a number of statistical values for each of these, which allows her classifier to function in a 15 dimensional feature space. However, if we hope to train our model in a way which allows us to generalize across different speakers, we may be forced to increase our feature space in order to account for the variety of mean pitches of speech encountered.

Once an adequate approach to feature space reduction is determined, the next task will be to select an algorithm for categorizing input data into the categories of encouragement versus discouragement. Because the task at hand requires only binary classification, we will train a linear classifier on the labeled data of the corpus, which will result in the determination of a function which acts as a boundary line between data points representing audio including encouraging utterances and data points representing discouraging audio. This will allow us to then test the model's ability to correctly determine the affective quality of test utterances, which will be represented by the model's probability of correctly predicting the affect conveyed by test audio samples.

Once the model has been sufficiently trained and reports reliably correct predictions, we will move on to the second phase of the problem, which involves the integration of detected prosodic feedback into an approach to a secondary task. We will first attempt to integrate the effect of the speaker into the model's approach to the Mountain Car problem (Mountain, 2017), which has been used as a testing environment for a significant body of reinforcement learning research. In this problem, a car must learn that in order to reach the goal state, reaching the peak of one tall hill, it must first accelerate a certain distance up an opposite hill in order to coast back down it and combine this momentum with an acceleration up the target hill until it reaches its goal. In this scenario, the reward function which guides the model's learning is based only on

minimizing the time taken to realize the goal state. However, we intend to integrate auditory feedback into this reward function, allowing the model to recognize when it is approaching the problem in a slower or less efficient fashion earlier than it otherwise would have. This will allow it to reach its goal state more quickly and learn the correct approach to the problem more efficiently.

Once we have shown the efficacy of integration of auditory feedback into the reinforcement learning algorithm as applied to the mountain car problem, we intend to expand the application of this feedback into models which use RL to learn to play video games, specifically Pacman. The model we plan to use in this sub-phase includes a reward function, as is necessary in a reinforcement learning strategy, but the reward function is based exclusively on changes in score while playing the game. This means that the model does not ever learn to avoid ghosts, or to avoid its own death. Our integration of prosodic analysis of auditory input into the model's reward function will allow human spectators to make statements with encouraging or discouraging affect in a far more diverse and useful range of scenarios over the course of gameplay which will hopefully allow the model to become skilled at the game after significantly fewer trials.

Our approach to the integration of Phase 1 into the reward functions of these two tasks will also be split into a preliminary phase followed by a full integration phase. We will first create proxy inputs which will represent encouraging and discouraging auditory feedback, such as the pressing of a certain key to represent encouragement and another to represent discouragement. Once we have demonstrated that this change to the reward function is effective, we will then substitute the results of our actual, real-time auditory analysis model in the place of these proxy inputs, which will result in the final product of a model which can include auditory feedback from humans in the reinforcement learning process when solving a task in order to increase the efficiency of the learning process.

Testing our Project

Part 1 - Recognizing Positive and Negative tones of speech

Testing this part of our project is relatively straight forward compared to the second part of the project described below. Ideally, we would want the classifier to be 100% accurate in predicting whether or not an input is positive or negative. Of course, audio files and speech patterns in general have a large amount of variation, so 100% accuracy is not a feasible goal in such a short amount of time. Instead, we will define success on a scaling basis as follows:

- PERFECT - 95% - 100%
- Very good - 80% - 95%
- Fair - 70% - 80%
- Poor - 60% - 70%
- Fail - < 60%

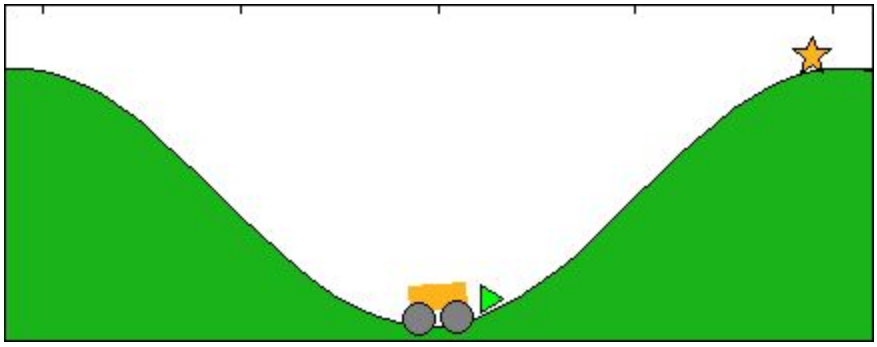
A score of < 50% is worse than the predicted outcome if the classifier just guessed between positive and negative each time. Anywhere around 50% (i.e. < 60%) should be deemed a failure because it's too close to the expected outcome of just random guesses. We will be

aiming for the “Very good” range, but being in the upper “Fair” range should be enough for us to implement the second part of the project.

To calculate these percentages, we will have a large dataset of positive affect and negative affect audio files that we will feed into the final processor. From there, we will simply divide the total number of input files with the total number of successful predictions. Currently, we are still looking for databases with large amounts of audio recordings separated by emotions. If we can’t find a dataset that we can use, we will probably create our own dataset.

Part 2 - Teaching AI software how to accomplish a task with the help of prosody

The second part of our project is a lot harder to pinpoint what is a success and what is a failure. We plan on using the classifier we created for part 1 to train progressively harder tasks. For the simpler tasks, we will declare the test a success if we can teach a program to accomplish the task with prosody audio files alone. For the more complicated tasks, such as playing Pacman, we will compare the score from a standard learning algorithm and our algorithm which includes prosody.



Mountain Car Problem: the car doesn’t have enough acceleration to overcome gravity, so the program must learn how to use gravity to reach the star at the top of the hill.
(image source: https://en.wikipedia.org/wiki/File:Mcar.png)
(example solution: https://www.youtube.com/watch?v=XEKlSx1hLYg)

The first learning problem we will attempt to train is one of the simpler reinforcement located in the burlap reinforcement training library. For the first part of the project, we will probably stick with the simpler tasks such as the Mountain Car Problem (Mountain, 2017), which is shown above . Since the task is relatively simple, we will not compare our final solution with one that uses a standard reinforcement learning algorithm. Instead, we will define it as a success if we can solve the Mountain Car Problem using Prosody alone. To do this, we will modify an already existing implementation of the Mountain Car Problem taking in user input, and add an audio input to the program.

The second learning problem we will attempt to train will most likely be a version of pacman. Plenty of reinforcement learning solutions for pacman already exist. We will modify the reward system of the pacman game to handle audio input for a specific case. For example, instead of the penalty for being killed by a ghost being hardcoded into the reward system, we can have someone spectate the game and give a disapproval tone when the player dies. After

sufficient training, we will determine if we succeeded based on the comparison of scores the normal pacman training implementation achieves on average, and our implementation.

Timeline/Schedule

We have decided to allocate the most time to creating a proper classifier for various tones of speech. We think that this will be one of the harder parts of our project, since part 2 of our project will just be us modifying the reward functions of already implemented solutions.

Oct 26th - Nov 1st
- Research Prosody features, try and decide what features of speech we want to use for the classifier

Nov 1st - Nov 7th
- Determine what classifier we will use, start working on classifier

Nov 7th - Nov 14th
- Finish classifier, begin testing input

Nov 14th - Nov 21st
- Did the classifier work? Tweak classifier as needed

Nov 21st - Nov 31st
- Mountain Car Problem/Begin working on Pacman problem

Nov 31st - Dec 10th
- Pacman problem

Dec 10th - Dec 15th
- Final project writeup

References

Kim, E. S., Gold, K., & Scassellati, B. (2008). What Prosody Tells Infants To Believe. *Development and Learning*. doi:10.1109/DEVLRN.2008.4640842

Kim, E., & Scasselati, B. (n.d.). *Learning to Refine Behavior Using Prosodic Feedback*[Scholarly project]. Retrieved October 24, 2017. Research done at Yale University

Mountain car problem. (2017). Retrieved October 24, 2017, from https://en.wikipedia.org/wiki/Mountain_car_problem

Scassellati, B., Crick, C., Gold, K., Kim, E., Shic, F., & Sun, G. (2006). Social Development. *IEEE Computational Intelligence Magazine, 1*(3), 41-47. doi:10.1109/MCI.2006.1672987

Schirmer, A. (2010). Mark My Words: Tone of Voice Changes Affective Word Representations in Memory. *PLoS One, 5*(2), e9080. doi:10.1371/journal.pone.0009080

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., & Hirschberg, J. (1992). TOBI: A Standard for Labeling English Prosody, 2nd International Conference on Spoken Language Processing, Banff, Alberta, Canada, October 12-16, 1992.