1) Big area/problem

2) Main issue

3) Your fix

4) Your method

5) The main advantage

6) Contribution to 1

1. Using voice/audio data for social signal recognition is an integral part of affective computing and natural language processing (NLP).
2. However, classifying the given data into particular social stands (emotions) is a challenging task. Variety of approaches to cluster the audio data exist, but only few of them are effective.
3. In this paper, we propose an approach to classify the audio data into positive, negative, and neutral using feature extraction technique.
4. We first collect and annotate a dataset of ***217*** audio files with 3 phrases, normalize them, and finally apply different machine learning (ML) algorithms to classify each audio.
5. The most effective model identifies the intonation of the voice with ***74%*** accuracy compared to the other 2 models.
6. Our work provides an annotated dataset on 3 phrases and an effective model, which extends and justifies existing methods in social signal recognition using voice/audio data.

 This a longer version of sentences #1-3 of the abstract, and includes citations to existing work.

 Describe the big picture motivation and importance of the goal you chose, followed by the main issue(s)/problem(s) preventing current approaches from reaching the goal, ending with what you're contributing.

 The introduction should situate your paper amongst related works and describe its contribution/importance -- think of your Assignment 1. Are there any datasets on your topic that already exist, and why is yours novel? See one of my paper [introductions (Links to an external site.)](http://www.angelicalim.com/papers/humanoids2017_paper.pdf) for an example (although, this one was for a much longer paper.)

 The first sentence to your paper is the most important. Don't let it be generic, such as "the field of affective computing is becoming more and more popular", i.e. if another group could use the same sentence, re-write the sentence.

Can you ask the person to shut up, but mean absolutely the opposite of what you said? The answer is yes! The nature of the human emotions is a complicated phenomenon questioned by the greatest minds throughout the history. Although understanding the meaning of ‘shut up’ (or any other phrase) is comprehensible given the context, facial expression, and intonation, what are the criteria that differentiate the direct meaning of ‘shut up’ from its positive (excited) meaning?

All of these are the questions that are asked when we talk about Human-Robot interaction (HRI). An interactive robot Pepper, some voice interfaces such as Amazon Alexa, OK Google, Siri, and other interactive systems for customer service are the successful examples of the artificial intelligence (AI). Although some of these systems demonstrate impressive results, most of them fail to detect the social signals, especially when using voice data. Thus, scientists propose different methods to handle this issue. Some researchers use various modalities combining different types of data, while others only focus on audio data.

Various studies are conducted to identify an emotion given a dataset of audios and using different modalities. Some of the research focus on both facial expressions (video) and emotional speech (audio), where authors combine two modalities to classify 6 basic emotions (anger, dislike, fear, happiness, sadness, and surprise) based on 2 human subjects [1]. Similarly, the other research focuses on detecting the emotion of the song based on its lyrical and audio features [2]. In this paper [2], the authors combine 2 modalities, where the lyrical features are generated by segmentation of lyrics during the process of data extraction and features like energy, tempo, and danceability are extracted, to compute Valence and Arousal values.

At the same time, other studies for social signal or ‘emotion’ recognition are only based on one modality: the audio itself. For example, the research on the evaluation of musical features for emotion classification uses a ground truth data set of 2904 songs that have been tagged with one of the four words “happy”, “sad”, “angry” and “relaxed”, on the Last.FM web site [3]. They then extract 55 features from the audios using MIR toolbox and use k-nearest neighbor (KNN) and support vector machines (SVM) to classify the songs. Another research on audio signal processing for speech emotion recognition uses Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) datasets to classify the audios using a SoftMax classifier [4].

To train the model that would be able to differentiate emotions for this paper, we will focus on audio data only. To narrow down and systemize the study, we chose 3 phrases: ‘Shut Up’, ‘Are you serious’, and ‘Are you kidding me’. All these phrases can have different ‘moods. For example, the first phrase can be used in its actual meaning angrily (negative), it can be used when the person is surprised and excited (positive way), or it can be said without involving any emotions (neutral). The audio samples are collected using the Internet sources and annotated into positive, negative, or neutral manually. Given the prior work done and the secondary research we have conducted, the features of the audios such as loudness, energy, jitter, etc. helps clustering different types of emotion. As a result of extracting some of these features, several models could be trained to detect whether the emotion of the phrase as positive, negative or neutral.

Approach

The general approach to build such system had 3 steps:

1. *Collect and manually annotate data.* Going through Youtube and Younglish websites, recording the audios of different people pronouncing 3 phrases, and manually annotating/naming each audio file into positive, negative, neutral. After converting all audios into .wav format using online converters[5], they were placed under 3 folders for each phrase.
2. *Preprocessing and normalizing the dataset.* Removing noise and unnecessary silence from the audios, making sure the length of each audio does not exceed 3 seconds. Some of the work was done manually by cutting the files, while other part was done using ***Python LIBRARY.***
3. *Feature extraction and training model.* Extracting necessary features (numeric/graphic) from the audios and training Decision Tree and SVM classifiers on the features. Choosing the best model based on their performance.

However, there are different methods that were used to extract features and train the models. The first method is based on *numeric* values of the audio, while the second one uses *graphical* features for training.

For the numeric method, we extracted 11 features (rmse, chroma\_stft, spec\_cent, spec\_bw, rolloff, zcr, mfcc, intensity, sample\_width, frame\_rate, frame\_width) from each audio returning them first as an array, and then, as a mean. The features were extracted using librosa, numpy, and and pydub (AudioSegment) libraries and the dataset was created using Pandas library. After creating a labeled and processed dataset of 217 audios as a Pandas dataframe, the dataset was divided into testing and training set (ratio 8:2). After training on Decision Tree Classifier (trained on 11 features) and SVM (trained on the first 10 features), the SVM outperformed its competitor with ***41.6%*** of accuracy after applying 5-fold cross validation. Sklearn is the main library that was used for training, and performance evaluation.

1. <https://ieeexplore.ieee.org/abstract/document/840655>
2. <https://arxiv.org/abs/1506.05012>
3. <https://www.researchgate.net/profile/Yading-Song/publication/277715954_Evaluation_of_Musical_Features_for_Emotion_Classification/links/557187c408ae7467f72ca201/Evaluation-of-Musical-Features-for-Emotion-Classification.pdf>
4. <https://www.mdpi.com/1424-8220/20/1/183>
5. <https://cloudconvert.com>