Emotion Recognition
Field Session Proposal

April 6, 2018

Research into emotion recognition from audio signals

Tipton Loo
VP of Edge Analytics
425.250.0470
tloo@prokarma.com

Cait Riggs
Data Scientist I
850.572.1518
criggs@prokarma.com

Project Background

ProKarma is interested in using machine learning to recognize emotion and overall sentiment within audio samples of customer call recordings. We are partnering with T-Mobile who is providing the audio data and subject matter experts on customer service calls.

Our data science team has started research into extracting acoustic features that are known to contain emotional signals and using classification techniques to determine the emotion exhibited by a speaker in an audio clip.

Currently, our team is using acoustic features extracted from the raw audio clips over small frames for the duration of the clip. We then use these features as input into our classifier model – best performance so far has been achieved using a bi-directional long short-term memory recurrent neural network (BLSTM RNN).

We’ve tested this proof of concept model on the IEMOCAP open source speech dataset while we work with T-Mobile on transferring and labeling the customer call audio data, but there are several more areas of opportunity for research on emotion recognition discussed in the Potential Research Areas section.

The Emotion Recognition Field Session project will be a good fit for students interested in gaining machine learning, data science, signals processing, NLP, and deep learning experience while working in an Agile team environment. This project will use Python and common data science tools like Jupyter notebooks/lab, numpy, conda, pandas, matplotlib, and tensorflow.

Potential Research Areas

Modeling

Another technique we are interested in trying is using a convolutional neural network (CNN) to discover features about the audio input instead of first extracting pre-defined acoustic features.

Work is also needed to determine if batch normalization on the audio clips would increase performance. Any work on modeling would likely be trained on open source datasets like IEMOCAP at this point while we work with T-Mobile to get a robust, labelled dataset of customer service call recordings.

So far, our research has focused on using just the acoustic features of audio clips, and we have yet to model the transcripts from the customer calls. Our team is also interested in analyzing customer call transcripts for the general sentiment of a caller when interacting with a customer service representative.

Data Processing

Since having quality, labelled data for this project is paramount, work to better automatically split full-length customer calls into shorter clips is also an area to continue working on. This would likely require knowledge of audio signals processing, or a penchant for researching and comparing open source tools.

Our team has tried two methods so far: (1) Using Audacity's Silence Finder tool to semi-automatically split the average 3.5min customer calls into average 3 second clips (std 3.7 sec), and (2) using Google's webrtcvad library to automatically split the customer calls into average 8 second clips (std 9.6 sec). But more work to better segment speakers into separate clips is needed.

We're interested in developing a better automated audio splitting solution since manually splitting would not scale to a dataset of hundreds of thousands of full-length customer calls. Although, we believe that manually **labelling** clips once they're split into smaller segments will be required.

Proposed Deliverables

Since there are several areas of research possible for this proposal, the following are proposed deliverables. The student team can decide which to research and which will be feasible for their team to tackle within the 6-week period. Our team at ProKarma can assist in determining complexity depending on the team's areas of experience.

- Develop a CNN to classify audio clips into emotional states (anger, excitement, neutral, sadness, etc.) or sentiment (positive, negative, neutral) of the customer towards the customer service representative
- Use transcripts extracted from audio data to develop a classifier model to classify the sentiment of a speaker over the duration of a call
    - Transcripts for the IEMOCAP dataset are available, but not for the T-Mobile clips. This deliverable could include a speech-to-text solution for creating transcripts from customer call recordings.
- Develop a tool that better splits full length customer calls into shorter clips which represent only a single speaker in each split clip
    - This could possibly go hand-in-hand with extracting transcripts from the audio calls and splitting audio by sentences instead of silence if synced with transcripts

Project Management

ProKarma's data science team in Seattle will lead weekly stand-up meetings via Skype or Google Hangouts in order to discuss the student team's progress and any blockers which the team may need help overcoming. Our team will be available anytime during the project via email, and project team members should feel free to schedule time on our calendars for longer calls or questions outside of the weekly stand-up conference calls.

ProKarma will also provide any compute resources like AWS instances for model training. We may also provide access to a private project repo on our GitHub for the duration of the project.

A ProKarma representative from the project should be available to attend the final presentations.