INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

Emojive! Collecting Emotion Data from Speech and Facial Expression using Mobile Game App

Ji Ho Park[†], *Nayeon Lee*[†], *Dario Bertero*[†], *Anik Dey*[†*], *Pascale Fung*[†*]

[†]Human Language Technology Center
Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology
[*]EMOS Technologies Inc.

[jhpark, nyleeaa, dbertero, adey]@connect.ust.hk, pascale@ece.ust.hk

Abstract

We developed Emojive!, a mobile game app to make emotion recognition from audio and image interactive and fun, motivating the users to play with the app. The game is to act out a specific emotion, among six emotion labels (happy, sad, anger, anxiety, loneliness, criticism), given by the system. Double player mode lets two people to compete their acting skills. The more users play the game, the more emotion-labelled data will be acquired. We are using deep Convolutional Neural Network (CNN) models to recognize emotion from audio and facial image in real-time with a mobile front-end client including intuitive user interface and simple data visualization.

Index Terms: speech emotion recognition, gamification, data collection

1. Introduction

Most challenging problem of training modules for tasks like emotion recognition and sentiment analysis is that each requires huge amounts of data, since it became a common practice to use statistical and machine learning methods. Moreover, supervised learning, which is a frequently used method, requires the data to be labelled, making it even harder to collect data for training.

To tackle this problem of data collection, we built a mobile app to collect more data for the emotion recognition task by incorporating gamification to motivate more users to play with the existing emotion recognition module. Gamification has been one of the approaches taken to collect data effectively in different fields such as market research [2] and pollution monitoring [3].

By using this app, we can acquire speech and facial data that are labelled with six emotions - angry, anxiety, criticism, happy, loneliness, and sad. Our mobile app includes a frontend user interface and the backend server that uses Convolutional Neural Network (CNN) to extract emotion scores real-time from speech audio [1] and from facial image [4].

2. System Description

2.1. System architecture

The system consists of mainly two parts: front-end user interface (UI) and back-end server, which includes the emotion detection modules. The front-end UI takes the user inputs from the camera and audio recorder of the smartphone, sends them to the emotion detection back-end server, and displays the detection results back to the user. The architecture diagram is shown in Figure 1.

The front-end UI is implemented using React Native framework. User input modules are responsible for getting input, speech and image, from the users. Collected user inputs are passed to backend server to obtain emotion scores.

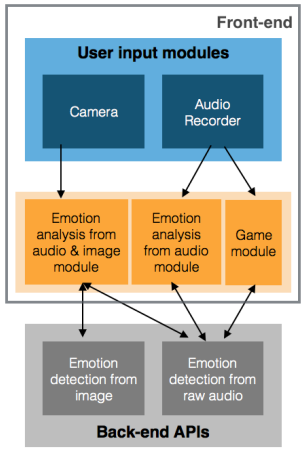

Figure 1: *System Architecture Diagram*

2.2. System UI Design

There are two detection demonstrations in this app: emotion detection from only speech and from both speech and image.

In the speech-only demo, a user can input his or her voice (the length of the speech should be within 30 seconds). Once the users finish recording, real-time detection will be processed and displayed in forms of arc charts as shown in Figure 2.

In the speech & facial image demo, the user's facial image from the camera is additionally detected. The speech emotion scores and facial emotion scores have six and seven emotions respectively, but to provide more intuitive and user-friendly visualization, only top five emotions are shown in descending order.

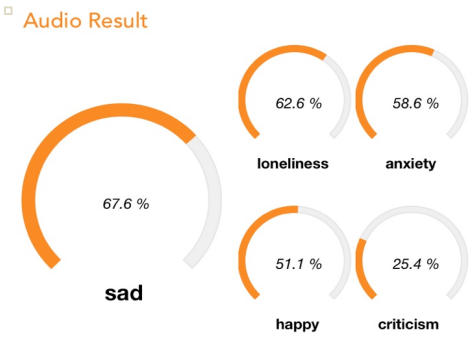

Figure 2: Emotion Scores

2.3. Gamification

We also implemented a simple mobile game that asks users to express a certain emotion. In the game, memes related to one emotion are randomly generated to direct the user about which emotion to act out. When the user acts out an emotion the app will give an acting score from the speech detection module. Two people can play the "Double Player mode", which decides the winner by comparing the acting scores, as demonstrated in Figure 3.

By asking the user to act out a specific emotion, we can use the data from the game as a labelled instance to re-train the detection module. We expect this game can be one of the methods to gather labelled emotion data.

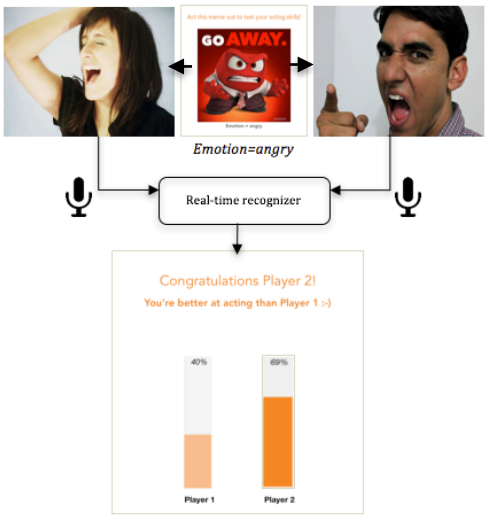

Figure 3: Illustration of Game Mode

2.4. Emotion Recognition

The audio input from the user is analyzed by a real-time speech emotion recognition module, which is a convolutional neural network (CNN) to detect emotions without the need of any feature engineering [1]. The CNN is designed with one filter, convolution window size of 200, which is 25ms of audio, and overlapping step size of 50, around 6ms. The convolution layer extracts the features, and the following max-pooling adds the contributions of all frames, and gives a segment-based vector as an output. This output vectors are inputted into a fully connect layer and then finally the softmax layer, which gives probabilities over the six emotion labels as shown in Figure 4.

The model is trained on a dataset we built, which includes 207 hours of speech extracted from 1495 TED talks. We annotated the data with existing commercial API followed by manual correction by humans. We divided the data into segments of average length of around 13 seconds.

The image input of the user's facial expression is analyzed by a real-time facial expression recognition module, which is also a convolutional neural network (CNN) trained on facial action units [4].

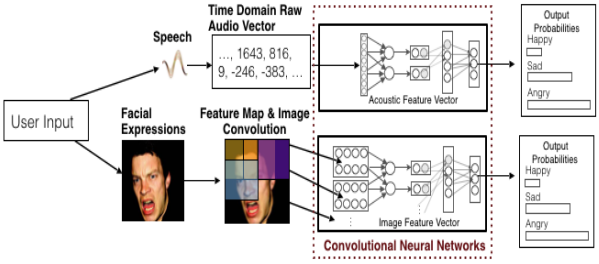

Figure 4: Architecture of the CNN

3. Data Collection

Since our mobile game guides the user to act out a certain emotion, we can assume that the collected input contains that guided emotion. To filter out invalid instances, we do not collect inputs that have scores lower than probability 0.5 from our pre-trained audio emotion recognizer. Manual correction may be needed to ensure the quality of the data from the app. The collected instances will be used for re-training the classifier to improve performance.

We are planning the publish the app through the Apple App Store. We expect more users will be able to interact with our app so we can more data from different kinds of users from various places.

4. Conclusion

We have described Emojive!, a mobile game app that effectively recognizes and collects emotion data from users. We made a mobile app user interface with effective visualization of the emotion scores and a simple game element with a real-time emotion recognizer. Our approach shows that gamification can make research outcomes more fun and interactive and, thus, can help us collect labelled data effectively.

5. References

[1] D. Bertero et al, "Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems", 2016

[2] J. Cechanowicz et al, "Effects of gamification on participation and data quality in a real-world market research domain," in Proceedings of the First International Conference on Gameful Design, Research, and Applications, 2013, pp. 58-65.

[3] I. G. Martí et al, "Mobile application for noise pollution monitoring through gamification techniques," in International Conference on Entertainment Computing, 2012, pp. 562-571.

[4] Z. Yuqian and E. Bertram, "Action Unit Selective Feature Maps in Deep Networks for Facial Expression Recognition", In Proceedings of IEEE International Joint Conference on Neural Networks, 2017.