Speech Emotion Recognition with Deep Learning – Project Proposal

Weijia Cao wcao34@wisc.edu
ChuYang Chen cchen466@wisc.edu
Xiu Xie xxie65@wisc.edu

1. Introduction

Speech is one of the fastest and most efficient ways of human communication. People speak to each other every day to exchange ideas, express feelings, etc. Though verbal communication is often considered the most crucial part of speech, non-verbal communication, such as speech emotion, also plays an essential role in human communication. Correctly recognizing the speech emotion makes communication much more efficient. Humans are naturally capable of doing such tasks, but detecting speech emotion is much harder for a machine, which makes human-machine communication still not a well-developed field. The purpose of this project is to build a machine that can successfully recognize basic human speech emotion so that human-machine communication can be improved.

1.1. Related Work

There are plenty of previous works which help study the implementation of this project. In particular, the study conducted by Tal Sobal-shikler [1] demonstrated the feasibility of detecting the emotions and the mental states from non-verbal expressions in speech. Her work showed the possibility of even detecting the subtle expressions and mixtures of emotions.

Also, Huang et al. [2] used Convolutional Neural Networks (CNNs) in deep learning systems to conducted a study for Speech Emotion Recognition (SER). Specifically, they used a semi-CNN to train the data set, and the performance of their models is outstanding. The recognition is stable and robust even under some unfavorable scenes (speaker with background noise).

2. Motivation

Interpreting emotions is critical in interpersonal relationships and therefore also important in human-machine interfaces. Speech is one major channel for emotion recognition, especially when the other channels, such as facial expression, gesture, and posture, are inaccessible. The ability to recognize emotions from speech also enables more efficient interpersonal and human-machine interactions in distance. For instance, implementing SER algorithms might improve voice-controlled devices to better understand the need of their users even when they cannot see or touch the user. Such techniques can also be used in marketing, where emotions expressed in speech can be detected to predict customers' attitudes and actions and to help intelligent agents to provide customers with sensible feedback. In addition, adopting SER algorithms in intelligent in-car systems might help detect the driver's mood state and take preventive actions when the driver is in a mood that is dangerous for driving (e.g., sleepy, angry, etc.) [3].

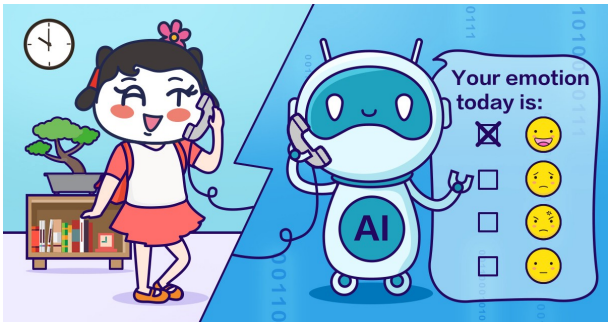

Figure 1. [4] Speech Emotion Recognition enables more efficient human-machine interactions in distance.

Besides such technical and commercial values, implementing deep-learning methods to perform SER also has scientific implications. Scientific researchers are still on their way to find features that are efficient enough in telling apart different emotions.[5] Because deep-learning algorithms can extract useful high-level features from raw speech samples with less manual human intervention, they are promising in helping us find the defining features of different emotions, or at least in performing factor-screening. The results will likely have significance in psychological and linguistic research, which can further benefit mental health services.

3. Evaluation

This project aims to build an SER classifier. Training examples come with seven labels indicating different emotions; namely, the feelings of *angry*, *fear*, *happy*, *neutral*, *sad*, and *surprise*. Correspondingly, prediction output will

take on one of the seven classes. A successful outcome will be the accurate prediction of emotions given the audio data.

An appropriate evaluation metric for the purpose of this study is the Multi-category Cross Entropy (also known as the Negative Log-Likelihood), which takes on the following form:

$$H_a(y) = \sum_{i=1}^{n} \sum_{k=1}^{K} -y_k^{[i]} \log\left(a_k^{[i]}\right),$$

where n is the number of examples, K is the number of classes, y is the predicted label that takes on a value of either 1 or 0, assuming one-hot encoding, and a is the activation function output. A minimal Cross Entropy score is desired.

Models utilized in this project include the Convolutional Neural Network, Recurrent Neural Network, and Transformer, with logistic regression being the baseline model. Given the baseline model, it is expected that after the implementation of the three algorithms mentioned above, a lower Negative Log-Likelihood score will be achieved.

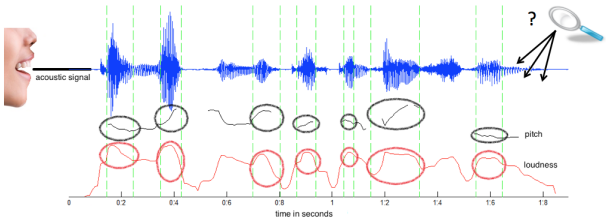

Figure 2. [6]

4. Resources

Since there are limited data in the area of emotional speech, currently, there is no single database large enough to serve the purpose of training in deep learning. Hence, we will need to put several databases together to form our total dataset. Databases that are frequently used in the field of SER include but are not limited to: the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [7], the Toronto emotional speech set (TESS) [8], the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) [9], the Surrey Audio-Visual Expressed Emotion (SAVEE) Database [10], the Emotional Voices Database (EmoV-DB) [11]. The databases mentioned above together contain around 25,000 emotional speech samples. Our tentative plan is to use the parts that share the same labels of emotions and are produced by English speakers (regardless of accents). If we are able to find more data in the future to form a more homogeneous dataset, we will likely do so. Another possible plan to expand the size of our dataset is to use emotional speech databases that are produced by speakers of different languages and to mix all the samples regardless of languages when training and testing the algorithm.

We will be doing the computation using Python through Jupyter notebook for algorithm development. The packages we will mainly be employing are PyTorch, NumPy, Pandas, Keras, and librosa. The computer hardware applied is the laptops of each group member. Also, we will be using Google Cloud GPU for model computation.

5. Contributions

As a group, it is decided that each member is taking responsibility to evenly divide the computational and writing tasks. For this proposal, Cao worked on the planning, motivation, resources portion, Chen worked on the introduction, motivation, and references of the proposal, and Xie worked on the evaluation and contribution parts. Regarding the final report, Chen will be focusing on the introduction, related work, and the organization; Xie will be responsible for the Proposed Method and Experimentation sections. Cao will be handling the Results, Discussions and Conclusion. All three members are responsible for vetting through all reports and proposals before submission.

For the computational tasks, each group member is responsible for finding one or more SER datasets, and the team will be working cooperatively to reconstruct and combine the datasets into an operable one. Chen will be taking care of data cleaning and feature engineering. Cao will be responsible for conducting experiments and building deep learning models. Xie will be comparing and evaluating the models.

The group will be setting up deadlines for each phase in the project and will hold group meetings on Zoom to code and work together to successfully find the model that most accurately classifies and predicts the emotions.

References

[1] Tal Sobol-Shikler. Analysis of affective expression in speech. Technical report, University of Cambridge, Computer Laboratory, 2009.

[2] Zhengwei Huang, Ming Dong, Qirong Mao, and Yongzhao Zhan. Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, 2014.

[3] Shivamburnwal. Speech emotion recognition, May 2020.

[4] Alibaba Tech. An ensemble framework of voice-based emotion recognition, Apr 2018.

[5] S. Suganya and E. Y. A. Charles. Speech emotion recognition using deep learning on audio recordings. In *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)*, volume 250, pages 1–6, 2019.

[6] Mailchimp. Speech emotion analyzer, 2018.

[7] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), apr 2018.

[8] M. Kathleen Pichora-Fuller and Kate Dupuis. Toronto emotional speech set (TESS), 2020.

[9] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

[10] Philip Jackson and Sana ul haq. Surrey audio-visual expressed emotion (savee) database, 04 2011.

[11] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 2018.