Multi-Label Speech Emotion Recognition Using 2D CNNs

Brian Pho, Thomas Truong, Svetlana Yanushkevich
Biometric Technologies Laboratory, Department of Electrical and Computer Engineering
University of Calgary, Canada
{brian.pho, thomas.truong, syanshk}@ucalgary.ca

***Abstract*—Current emotion classification systems using machine learning often ignore the multi-label data that comes with databases. In this paper we question whether machine learning can be used to detect multiple emotions in speech. To answer this question, we created an aggregate database from four voice emotion-labeled databases and train a 2D convolutional neural network on it to determine if the model could recognize multiple emotions in a speech sample. The model was able to classify the samples with an accuracy of 50%. While this result is better than the baseline, we suggest future improvements.**

***Index Terms*—speech, emotion, classification, neural networks**

I. Introduction

A. Literature Review

The field of affective computing studies the development of systems that can recognize, interpret, process, and simulate human affects. One subfield of affective computing that we are interested in is speech emotion classification. Classifying speech based on its emotional content is applicable to the development of more user friendly kiosks, more interactive digital assistants such as Siri or Google Assistant, and allows for a better understanding of what parts of speech contribute to its emotional content. One method of classification that has demonstrated great success is deep learning. Specifically, the use of convolution neural networks have proven to be successful at classifying emotions in speech as shown by []. Current research has successfully applied deep learning to automate this process and achieves an accuracy of 90% [**?**]

However, these models only consider a single emotion per speech sample and are often only trained on one to two databases. This paper extends upon this work by considering the problem of multi-label emotion classification of speech using four databases. The IEMOCAP and CREMA-D databases both include multiple labels for each speech sample but the data is usually discarded by only considering the emotion with the majority votes. We argue that discarding the other emotions results in a less realistic model of emotion classification due to not matching human performance and because it sidesteps the problem of ambiguous samples.

B. Summary

We approach the problem of multi-label emotion classification by using 2D convolution neural networks to classify speech samples into multiple emotions. We collected four databases, two with mutli-labeled and single-labeled samples, and two with only single-labeled samples. We then combined these four databases into an aggregate database by processing all of the speech samples into log-Mel spectrograms. These log-Mel spectrograms were then fed into a seven layer neural network consisting of four convolutional layers and three dense layers. After training the neural network, the model achieved a 50% test accuracy.

II. Results

A. Method Summary

To determine if a machine learning model could perform multi-label emotion classification on speech, we first need to get speech samples with their labeled emotion. We considered eight databases and chose four based on accessibility and the number of common emotions. Given the variability of emotions in each database and the constraint adding more databases increases the difficulty of controlling the consistency of the samples, we chose the following databases

- IEMOCAP
- TESS
- RAVDESS
- CREMA-D

We specifically picked databases that were spoken in English to keep the language consistent.

Each sample from a database would flow through the same preprocessing steps to maintain consistency. The preprocessing steps are

1) The sample starts as a raw waveform in the form of time series points specifying the amplitude at a certain time.
2) we then pad or crop the sample to a desired length. In our case, we set a maximum length of 4.5 seconds. Samples shorter than this were zero-padded on the right tail. Samples longer than this were cropped to this length and the extra information was discarded.
3) If the sample came from a database that we considered noisy, then a noise reduction function was run on the sample to reduce the noise. We consider the IEMOCAP and CREMA-D databases to be noisy.
4) The sample is then converted into a log-Mel spectrogram using a short-time Fourier transform (STFT) and Mel scale equation. The phase information was discard as it does not seem to hold relevant information. []

5) The final step is to normalize the spectrograms to have values between negative one and one. This was done by using a minmax scaling function.

After all of the databases were processed this way, the final aggregate database was fed into a neural network for training. We constructed a seven layer neural network of four convolutional layers and three dense layers. The network architecture is shown in Figure **??**. The training details follow.

- We applied dropout to the dense layers and batch normalization to the convolutional layers to deal with overfitting.
- We shuffled the aggregate database to make each training batch more homogenous and thus prevent large gradient updates.
- We split the aggregate database into 80% training, 10% validation, and 10% testing.
- We updated the model's hyperparameters based on the validation loss and accuracy to improve the model's accuracy and generalizability.

After training, the model was run on the testing set to get the final accuracy of the model. The final accuracy achieved was 50%.

B. Rationale

III. Discussion

A. Results Summary

We created a neural network model that achieved a classification accuracy of 50% on the problem of multi-label emotion classification. This model can successfully classify multiple emotions in a sample and while this is a good result, we can do better.

B. Limitations and Future Work

C. Conclusion

Acknowledgment

Acknowledge PURE award. Test citation: [1]

References

[1] N. Fahlgren, M. Gehan, and I. Baxter, "Lights, camera, action: high-throughput plant phenotyping is ready for a close-up,," *Current Opinion in Plant Biology*, vol. 24, pp. 93–99, apr 2015.