# 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 address the problem of whether machine learning can be used to detect multiple emotions in speech and how accurate they can identify the emotions present. We created an combined database from four voice emotion-labeled databases and trained 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%. This result shows that it is possible to apply machine learning to the problem of multi-label speech emotion recognition and to achieve a reasonable accuracy.**

***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.

We use "model" and "neural network" interchangeably.

### 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 combined 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.

The paper is outlined as follows:

- Section II details the methodology we use such as the how we combined four databases and the architecture of the neural network.
- Section III details the results from training the neural network.
- Section IV discusses the results in the context of the field and how future work can improve upon the work.

## II. Method

### A. Preprocessing

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 [1]
- TESS [2]
- RAVDESS [3]
- CREMA-D [4]

We specifically picked databases that were spoken in English to keep the language consistent. We combined these four databases for training, validating, and testing the neural network. The combined database is detailed in figures 1, 2, 3.

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

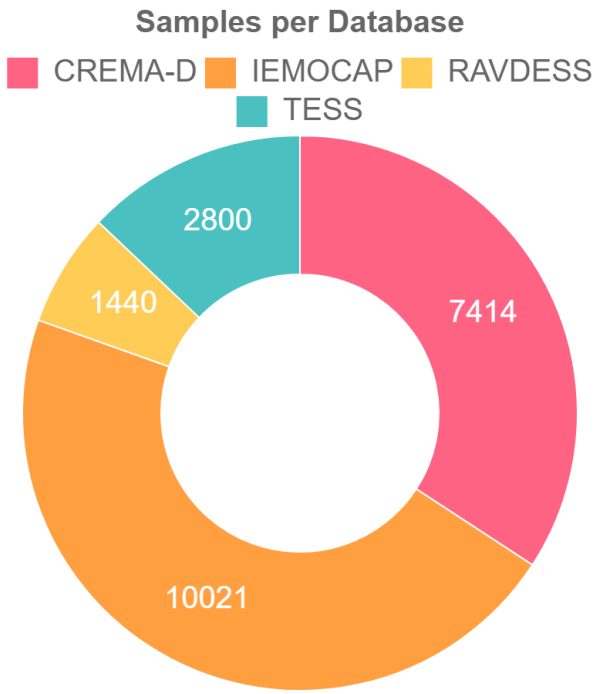


Fig. 1. The proportion of each database in the combined database.

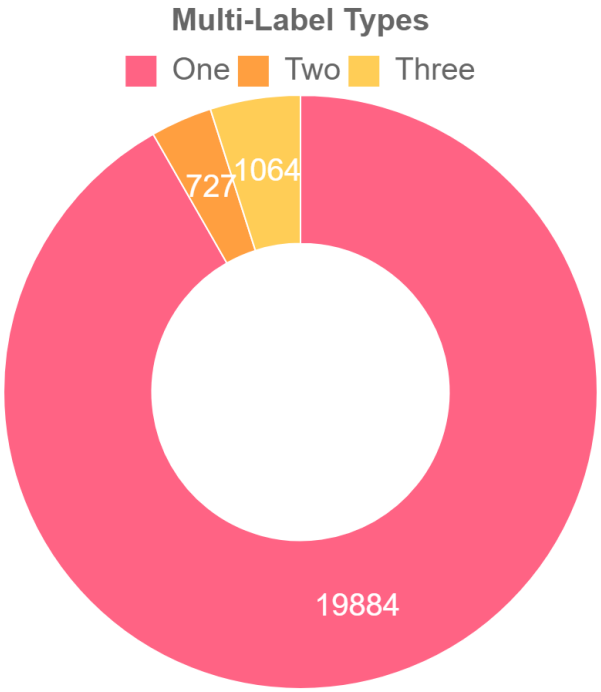


Fig. 3. The proportion of each label type in the combined database.

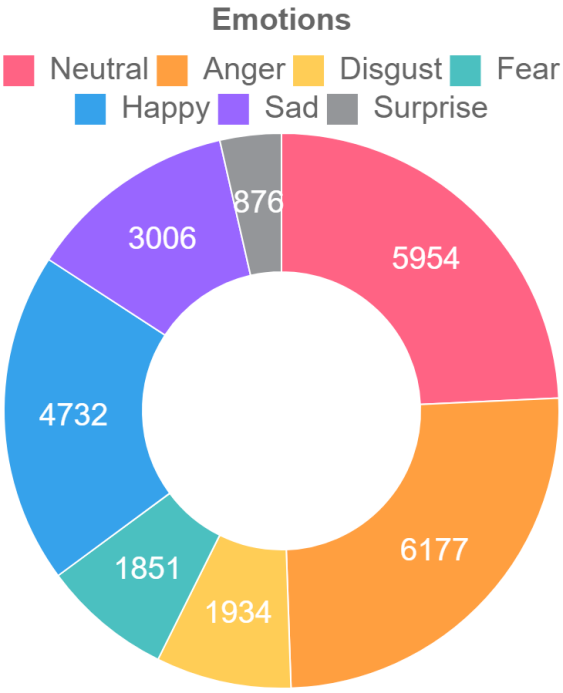


Fig. 2. The proportion of each emotion in the combined database.

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 combined 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 combined database to make each training batch more homogenous and thus prevent large gradient updates.
- We split the combined 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.

### B. Neural Network

## III. Results

After training, the model was evaluated on the testing set to get the final accuracy of the model. The final accuracy achieved was 50%. Confusion matrices for each emotion are shown in Figure **??**.

## IV. 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

One limitation of this work was that we only accounted for seven emotions but we know that there are more emotions than this. [] Another limitations was that Here we present future improvements of our work

- Improving the recognition accuracy by using more sophisticated neural network architectures such as LSTMs or using more databases.
- Incorporating the phase data from the STFT and testing to see if that improves the accuracy.
- Testing a binary relevant approach to this multi-label problem and comparing it to this model.

### C. Conclusion

Overall, this paper presented a convolutional neural network model that achieves a 50% accuracy on multi-class, multi-label speech emotion recognition using four combined databases. We obtained this result by transforming raw speech samples into log-Mel spectrograms using STFT and the Mel scale. These log-Mel spectrograms are then fed into a eight layer neural network for classification for training. When we have completed training the neural network, the model was evaluated and achieved as 50% accuracy. While this result is a promising start, we suggest improvements that future work can build upon to improve the accuracy.

## Acknowledgment

Acknowledge PURE award.

## References

[1] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[2] K. Dupuis and M. Kathleen Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set," *Canadian Acoustics*, vol. 39, no. 3, pp. 182–183, Sep. 2011. [Online]. Available: https://jcaa.caa-aca.ca/index.php/jcaa/article/view/2471

[3] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, p. e0196391, 2018.

[4] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

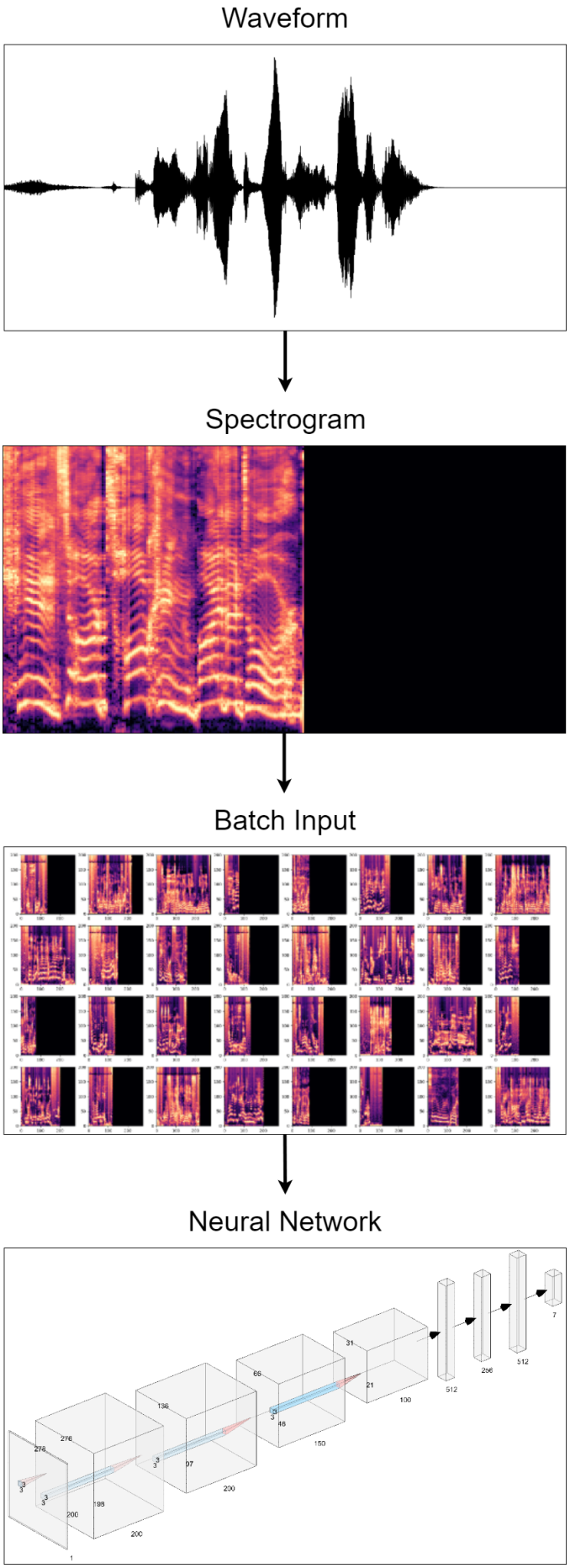


Fig. 4. A high level overview of the processing stages that a speech sample goes through.

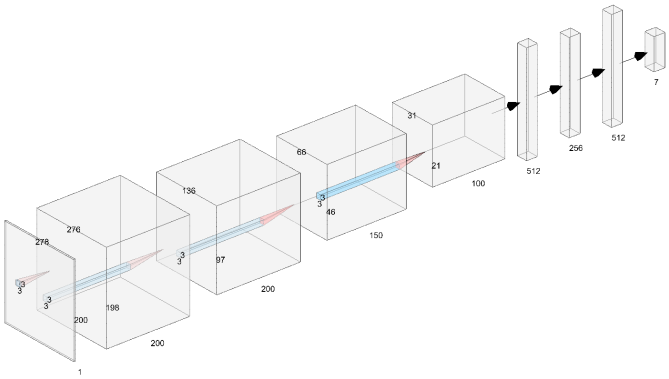

Fig. 5. The detailed architecture of the neural network including its eight layers and convolution filter size.