Multi-Label Speech Emotion Recognition Using 2D Convolutional Neural Networks

Anonymous FG2020 submission
Paper ID ****

***Abstract*— Current speech emotion recognition systems often overlook 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. We created a combined database consisting of four speech emotion-labeled databases, and we used it to train a 2D convolutional neural network 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 57.64%. This shows that it is possible to apply machine learning to the problem of multi-label speech emotion recognition and to achieve a reasonable accuracy.**[1]

I. INTRODUCTION

The field of affective computing studies the development of systems that can recognize, interpret, process, and simulate human affects. A subfield of affective computing, the one that we are interested in, is speech emotion recognition (SER). SER is the problem of recognizing which emotions are present in speech and it is important because of its applications to: call center conversations that improve service quality, speech translation systems that are more accurate at conveying emotions, and robotic pets that are more compassionate [1].

A. Literature Review

SER has historically been approached in two ways: feature engineering and machine learning (ML).

Feature engineering is a method of manually extracting desired parameters from the data for use in recognition. For audio, examples of parameters include pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCC) [2]. Nassif et al. [3] found that most researchers used the MFCC parameter for deep learning models but also recommended other parameters such as Linear Predictive Coding.

Machine learning is a method of creating models that automatically extract their own parameters for recognition without being manually programmed. Examples include convolutional neural networks (CNN), recurrent neural networks (RNN), and long short-term memory (LSTM). Both approaches are not exclusive and a hybrid approach of feature engineering and machine learning results in better classification accuracy [3].

A common example of combining both approaches is to convert an audio waveform into a spectrogram using the Short-Time Fourier Transform (STFT) and to feed the spectrogram into a neural network. Papers such as [4], [5], [6], and [7] have demonstrated that this approach is successful for processing audio and for recognizing emotions in speech.

Current state-of-the-art SER has achieved an accuracy of 52.14% for the case of speaker-independent, single-labeled emotions on the IEMOCAP database [7]. This result was achieved using a 2D CNN LSTM model where the audio waveform was converted into a log-Mel spectrogram and then fed into the model.

However, this model and most other models only consider a single emotion per speech sample and are often only trained on one or two databases. Kim et al. [8] is an exception as they consider the multi-label case of SER but they did not use machine learning and they used visual data on top of audio data. This paper extends upon current literature by considering the problem of multi-label SER using four databases to train and test a CNN.

The IEMOCAP [9] and CREMA-D [10] databases both include multiple labels for each speech sample but the data is discarded by considering the emotion with the majority of votes. We argue that discarding the non-majority votes results in a less realistic model of emotion classification due to not matching human performance and due to the loss of information. For example, in the case of speech translation systems, translating ”surprise” speech lacks the nuance of whether the surprise is ”surprise-happy” such as the case of receiving a gift or if it is ”surprise-sad” such as the case of receiving news of the death of a loved one. A single label is not able to capture this nuance compared to multiple labels and could result in inaccurate or offensive translations.

B. Proposed Approach

The main idea of this paper is to build a more realistic speech emotion recognition system that can recognize multiple emotions in speech. We approach this problem by using more data and by using a 2D CNN to classify speech samples into multiple emotions. We collected four databases: two with multi-label and single-label samples, and two with only single-label samples. We then combined the four databases into an a single database by processing all of the speech samples into log-Mel spectrograms. These log-Mel spectrograms were then fed into an eight-layer neural network consisting of four convolutional layers and four dense layers.

This paper is outlined as follows:

[1]Code for this paper is available on Github at: `https://github.com/Brian-Pho/RVST598_Speech-Emotion-Recognition`

- Section 2 details the methodology we use such as the how we combined the four databases and describes the architecture of the neural network.
- Section 3 describes and discusses the results from testing the neural network.
- Section 4 concludes this paper and provides suggestions for future directions.

II. METHODOLOGY

We approach the problem of SER by preprocessing samples from four databases and then feeding those samples into a CNN. We preprocess the speech samples into log-Mel spectrogram to help the neural network extract features relevant to emotion recognition. This choice is based on previous work such as [4] and [11]. The choice of using a CNN is also based on previous work where Balakrishnan et al. [12] found that CNN-based models have superior performance compared to RNNs. Another justification for using CNNs is due to treating the log-Mel spectrograms as images and CNNs have been shown to perform well on images [13].

A high level overview of the data flow is shown in figure **??** where samples flow through three stages of processing. The stages are detailed further into the methodology. Both the preprocessing steps and the CNN were implemented in Python using the Librosa [14] and Keras [15] libraries respectively.

A. Preprocessing

To build a more realistic SER system, we first collect speech samples with their labeled emotion. We considered eight databases and chose four based on accessibility and based on the number of overlapping emotions. We chose four databases because it becomes more difficult to maintain consistency due to database variability. Databases can vary in

- The set and number of labeled emotions
- The number of labels per sample
- The audio quality such as sampling rate and noise
- The spoken language

Given this variability, we chose the following four databases: IEMOCAP [9], TESS [16], RAVDESS [17], and CREMA-D [10].

To control for the set and number of labeled emotions, we consider the following seven emotions for recognition: neutral, anger, disgust, fear, happy, sad, and surprise. We chose these seven emotions due to them being considered basic emotions by Ekman [18] and due to these seven being the most common among all databases.

To control for the number of labels per sample, we mixed the multi-labeled data from the IEMOCAP and CREMA-D databases with the single-labeled data from the TESS and RAVDESS databases. We chose to mix of single- and multi- labeled data to increase the number of samples that the model can learn from and because the single-labels can be considered as special cases of multi-labeled data. However, we removed samples that were labeled with four or five emotions because we consider these samples to be ambiguous and because outliers can hinder a neural network's ability to learn. Outliers can hinder a neural network's ability to learn by causing large gradient updates that prevent the model from converging [19].

To control for the audio quality, we resampled all samples to 48 kHz, applied noise reduction to samples from the IEMOCAP and CREMA-D databases, and cropped all samples to 4.5 seconds. We used a kaiser filter for resampling and used spectral gating for noise reduction. We only applied noise reduction to the IEMOCAP and CREMA-D databases after listening to the audio. Lastly, we chose databases that are spoken in English to maintain language consistency.

We combined these four databases into a combined database to train, validate, and test the neural network. Each sample from a database went through the same preprocessing steps to maintain consistency across all samples. The detailed steps are described below.

1) A sample starts as a raw waveform in the form of time series points specifying the amplitude at a point in time.
2) The sample is then padded or cropped to the desired length of 4.5s. Shorter samples were zero-padded on the right tail. Longer samples were right-tail cropped and the extra information was discarded.
3) If the sample came from a database that we considered noisy, then a noise reduction filter was applied to the sample. We consider the IEMOCAP and CREMA-D databases to be noisy.
4) The sample is then converted into a log-Mel spectrogram using the STFT and Mel scale. The phase information was discarded as it does not appear to hold relevant information [20].
5) The final step is to normalize the spectrograms to have values between negative one and one. This was done by using a min-max scaling function.

For the STFT, we used a window size of 3072 with a 75% overlap. This choice is based on the work of Zhao et al. [7] where they also use 75% overlap but with a smaller window size (2048) and they achieved excellent results.

For the Mel scale, we set the minimum frequency to 20 Hz and maximum frequency to 12 kHz with 200 Mel bins. The frequency range was chosen after experimenting with various ranges and selecting the range that resulted in visually clean spectrograms. The number of Mel bins was also chosen after experimentation as too few bins resulted in poor temporal resolution while too many bins resulted in poor frequency resolution.

$$X_{std} = \frac{X - X_{min}}{X_{max} - X_{min}} \qquad (1)$$
$$X_{scaled} = X_{std} \cdot (max - min) + min$$

The min-max scaling is defined in (1) where X_{min} is the smallest value in tensor X, X_{max} is the largest value in tensor X, min is the lower bound set to -1, and max is

the upper bound set to 1. We scaled all spectrograms to only have values between -1 and 1 as this reduces the chances of triggering large gradient updates that prevent the model from converging [19].

B. Neural Network

After all of the databases were processed as described in section II-A, the final combined database was fed into a neural network for training, validation, and testing. We constructed an eight layer neural network consisting of four convolutional layers and four dense layers. The training process is described below.

- We shuffled the combined database to make each input batch more uniform thus mitigating large gradient updates.
- We split the combined database into 80% training, 10% validation, and 10% testing.
- We applied dropout to the dense layers and batch normalization to the convolutional layers to deal with overfitting [21], [22].
- We updated the model's hyperparameters based on the validation loss and accuracy to improve the model's accuracy and ability to generalize.
- We used class weights during training to address class imbalance.

Table I describes the model's hyperparameters and table II describes the training hyperparameters. The decision to use *binary cross-entropy* as the loss measure is due to multi-label problems considering each output as independent and thus each output acts like a binary classifier. This lets the model only penalize the outputs that are incorrect without penalizing the other outputs. However, we use *categorical cross-entropy* as the accuracy metric because binary cross-entropy would provide the incorrect accuracy as it defined for a binary output such as 0 or 1, not vectors containing binary outputs. We could have picked a different accuracy metric but chose categorical cross-entropy due to its ease of use.

Likewise, we use *sigmoid* as the output layer's activation function instead of the more commonly used *softmax* function in classification. The justification is the same as in our choice of loss measure; sigmoid treats each output as independent which means multiple outputs can be true while softmax maximizes a single output while minimize all other outputs.

The final model was evaluated on the testing set by using average accuracy that was calculated from the confusion matrices shown in figure 1. The confusion matrices are calculated by comparing the true labels to the predicted labels for each independent emotion.

We define accuracy as the unweighted average accuracy of all emotions and this is mathematically described in (2) where n is the total number of emotions (7 in our case), TP_i is the number of true positives for emotion i, TN_i is the number of true negatives for emotion i, and $Total_i$ is the total number of predictions for emotion i.

TABLE I
THE MODEL'S HYPERPARAMETERS.

Hyperparameter	Value
Input Dimensions	278w x 200h x 1d
Optimization Algorithm	Adam
Learning Rate	0.001
Beta 1	0.9
Beta 2	0.999
Epsilon	10^{-8}
Loss Measure	Binary Cross-entropy
Accuracy Metric	Categorical Cross-entropy
Hidden Layer Activation Function	ReLU
Output Layer Activation Function	Sigmoid

TABLE II
THE TRAINING HYPERPARAMETERS.

Hyperparameter	Value
Epochs	20
Batch Size	32
Training Set	17,341
Validation Set	2,167
Testing Set	2,167

$$Accuracy = \frac{\sum_{i=1}^{n} \frac{TP_i + TN_i}{Total_i}}{n} \quad (2)$$

In summary, we calculate the accuracy per emotion and then average over these accuracies to obtain the final accuracy. We chose this definition of accuracy so that we can compare our results to other papers that also use a similar measure of accuracy that is calculated from the confusion matrices.

III. RESULTS AND DISCUSSION

After training the model, it was evaluated on the testing set to get the final accuracy. The final accuracy achieved was 57.64%. Comparisons to the current literature is shown in table III. We compare our results to the single-label case of SER as we could not find papers with results for the multi-label case.

The results achieved in this paper do not reach the state-of-the-art accuracy achieved for the single-label case. But considering how this is the first attempt on the more realistic problem of multi-label SER, we perform reasonably well. Compared to the literature shown in table III, no other paper considers as many emotions as we do while also using four databases and also tackling the multi-label problem. So we conclude that we are successful in building a more realistic SER system by using ML and more data to tackle the problem of recognizing multiple emotions in speech.

In analyzing the confusion matrices, we see that the model predicts an emotion is absent in most of the samples with "surprise" being the exception. We suspect that "surprise" being an exception is due to the class imbalance that is shown in figure **??**. The "surprise" class was the least represented class out of the seven classes and this makes the

TABLE III
COMPARISON OF SER ACCURACY IN LITERATURE.

Research Work	Method	Testing Method	Number of Emotions	Databases Used	Label Type	Accuracy (%)
Zhao et al. [7]	CNN + LSTM	Testing set	6	IEMOCAP	Single	52.1
Our Work	CNN	Testing set	7	IEMOCAP, TESS, RAVDESS, CREMA-D	Multi	57.6
Etienne et al. [23]	CNN + LSTM	Testing set	4	IEMOCAP	Single	61.7
Zhang et al. [24]	CNN	Testing set	4	IEMOCAP	Single	63.9
Fayek et al. [25]	CNN	Testing set	4	IEMOCAP	Single	64.8
Yenigalla et al. [26]	CNN	Testing set	4	IEMOCAP	Single	73.9
Badshah et al. [6]	CNN	Testing set	7	Emo-DB, Korean dataset	Single	80.8

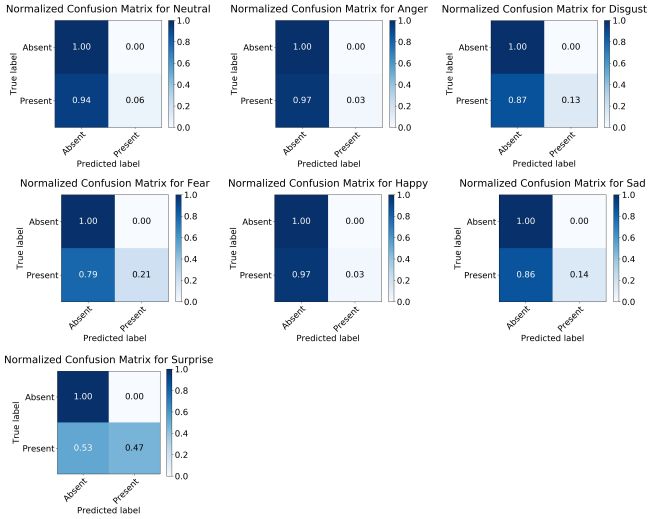

Fig. 1. Confusion matrices for each emotion.

model predict it more due to the use of class weights. Class weights bias the model towards underrepresented classes in an attempt to balance classes but this seems to have affected the model's ability to learn.

IV. CONCLUSION AND FUTURE WORK

Overall, this paper presented a 2D CNN model that achieved an accuracy of 57.64% on the problem of multi-label speech emotion recognition using four combined databases. We obtained this result by transforming raw speech samples into log-Mel spectrograms using the STFT and the Mel scale. The log-Mel spectrograms are then fed into an eight layer neural network for classification. While this result is a promising start, we suggest improvements that future work can build upon to improve the accuracy of the model and to expand the scope of emotions considered.

One limitation of this work is that we only accounted for seven emotions but recent research has suggested that there are more emotions such as boredom, shame, and triumph [27]. However, one issue with expanding the set of considered emotions is the lack of databases with the labeled emotion.

Another limitation of this work is that all samples are spoken in English so the model is biased towards Anglophones. In theory, the basic emotions are universal across languages and cultures so incorporating databases spoken in different languages, such as the Emo-DB database, would help the model generalize across languages [28].

The following list is a suggestion that future work could pursue:

- Using more sophisticated neural network architectures such as LSTMs or using more databases.
- Incorporating the phase data from the STFT.
- Testing a binary relevance approach to this multi-label problem.
- Replacing the use of STFT with a wavelet transform.

Rana et al. [29] has also shown that SER systems can be more robust by introducing noise into the samples which is another promising future direction.

V. ACKNOWLEDGMENTS

This work was funded by the Program for Undergraduate Research Experience (PURE) award granted by the University of Calgary.

REFERENCES

[1] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," pp. 99–117, jun 2012.
[2] J. Rybka and A. Janicki, "Comparison of speaker dependent and speaker independent emotion recognition," *International Journal of Applied Mathematics and Computer Science*, vol. 23, no. 4, pp. 797–808, dec 2013. [Online]. Available: http://content.sciendo.com/view/journals/amcs/23/4/article-p797.xml
[3] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech Recognition Using Deep Neural Networks: A Systematic Review," *IEEE Access*, vol. 7, pp. 19 143–19 165, 2019.
[4] J. Engel, K. K. Agrawal, S. Chen, I. Gulrajani, C. Donahue, and A. Roberts, "GANSynth: Adversarial Neural Audio Synthesis," feb 2019. [Online]. Available: http://arxiv.org/abs/1902.08710

[5] M. Chen, X. He, J. Yang, and H. Zhang, "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1440–1444, oct 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8421023/

[6] A. M. Badshah, N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5571–5589, mar 2019. [Online]. Available: http://link.springer.com/10.1007/s11042-017-5292-7

[7] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, jan 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809418302337{#}bib0265

[8] Y. Kim and J. Kim, "Human-Like Emotion Recognition: Multi-Label Learning from Noisy Labeled Audio-Visual Expressive Speech," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2018-April. IEEE, apr 2018, pp. 5104–5108. [Online]. Available: https://ieeexplore.ieee.org/document/8462011/

[9] 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.

[10] 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.

[11] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in *2017 International Conference on Platform Technology and Service, PlatCon 2017 - Proceedings*. IEEE, feb 2017, pp. 1–5. [Online]. Available: http://ieeexplore.ieee.org/document/7883728/

[12] A. Balakrishnan and A. Rege, "Reading Emotions from Speech using Deep Neural Networks," Tech. Rep., 2017.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[14] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 2015.

[15] F. Chollet, "Keras," 2015. [Online]. Available: https://github.com/keras-team/keras

[16] 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

[17] 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.

[18] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[19] F. Chollet, *Deep Learning with Python*, 1st ed., 2017.

[20] P. Kozakowski and B. Michalak, "Dcgan and spectrograms," 2017. [Online]. Available: http://deepsound.io/dcgan_spectrograms.html

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 448–456. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045167

[23] C. Etienne, G. Fidanza, A. Petrovskii, L. Devillers, and B. Schmauch, "CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation," pp. 21–25, feb 2018. [Online]. Available: http://arxiv.org/abs/1802.05630http://dx.doi.org/10.21437/SMM.2018-5

[24] Y. Zhang, J. Du, Z. Wang, J. Zhang, and Y. Tu, "Attention Based Fully Convolutional Network for Speech Emotion Recognition," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2018 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., mar 2019, pp. 1771–1775.

[25] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, aug 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089360801730059X?via{%}3Dihub

[26] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2018-September. International Speech Communication Association, 2018, pp. 3688–3692.

[27] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures." *Emotion*, vol. 18, no. 1, pp. 75–93, 2018.

[28] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech." in *INTERSPEECH*. ISCA, 2005, pp. 1517–1520. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/interspeech2005.html#BurkhardtPRSW05

[29] R. Rana, "Emotion Classification from Noisy Speech - A Deep Learning Approach," mar 2016. [Online]. Available: http://arxiv.org/abs/1603.05901