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

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 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 52.57%. 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]**

***Index Terms*—multi-label, speech, emotion, recognition, neural networks**

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 that we are interested in is speech emotion recognition (SER). SER is the problem of recognizing which emotions are present in speech and SER has been approached in two ways:

- Feature engineering
- Machine learning (ML)

A. Literature Review

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). [1] Nassif et al. [2] 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 an artificial neural network that automatically extracts its own parameters for recognition. Examples include support vector machines (SVM), convolutional neural networks (CNN), and recurrent neural networks (RNN). Both approaches are not exclusive and a hybrid approach of feature engineering and machine learning results in better classification accuracy. [2]

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 [3], [4], [5], and [6] have demonstrated that this approach is successful for processing audio and for recognizing emotions in speech.

Current state-of-the-art speech emotion recognition has achieved an accuracy of 52.14% for the case of speaker-independent, single-labeled emotions on the IEMOCAP database. [6] This result was achieved using a 2D CNN long short-term memory (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. [7] is an exception as they consider the multi-label case of emotion recognition but they did not use machine learning and they used visual data on top of audio data. This paper extends upon the state-of-the-art by considering the problem of multi-label speech emotion classification using four databases to train and test the neural network.

The IEMOCAP [8] and CREMA-D [9] 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 by discarding votes.

B. Proposed Approach

We approach the problem of multi-label emotion classification 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:

- Section II details the methodology we use such as the how we combined the four databases and describes the architecture of the neural network.
- Section III describes the results from testing the neural network.
- Section IV discusses the results in the context of the field and how the results can be improved upon.

[1] https://github.com/Brian-Pho/RVST598_Speech-Emotion-Recognition

Databases	Emotions												
	Neutral	**Anger**	**Disgust**	**Fear**	**Happy**	**Sad**	**Surprise**	**Calm**	**Excitement**	**Frustration**	**Amused**	**Sleepy**	**Bored**
IEMOCAP	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✗	✗	✗
CREMA-D	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
TESS	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
RAVDESS	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗
MSP-IMPROV	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
SAVEE	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗
Emo-DB	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✓
EmoV-DB	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗

TABLE I: Comparison of databases with their labeled emotions.

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 [3] and [10]. The choice of using a CNN is based on previous work where Balakrishnan et al. [11] found that CNN-based models have superior performance compared to RNNs. Another justification for using CNNs is due treating the log-Mel spectrograms as images where CNNs have been shown to perform well on. [12]

A high level overview of the data flow is shown in figure 2 where samples flow through the four stages of processing. Both the preprocessing steps and the CNN were implemented in Python using the Librosa [13] and Keras [14] library respectively.

A. Preprocessing

To determine if a ML model can perform multi-label speech emotion recognition, we first need to obtain speech samples with their labeled emotion. We considered eight databases and chose four based on accessibility and the number of overlapping emotions. Table I compares the set of emotions of each database which makes it easier to identify 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

1) IEMOCAP [8]
2) TESS [15]
3) RAVDESS [16]
4) CREMA-D [9]

To control for the set and number of labeled emotions, we consider the following seven emotions for recognition

1) Neutral
2) Anger
3) Disgust
4) Fear
5) Happy
6) Sad
7) Surprise

We chose these seven emotions due to them being considered basic emotions by Ekman [17] 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 samples with that many emotions to be ambiguous and because outliers hinder a neural networks ability to learn. Outliers can cause large gradient updates which prevents a model from converging.

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. The combined database is detailed in figure 1.

Each sample from a database would flow through the same preprocessing steps to maintain consistency across all samples. The detailed steps are described below.

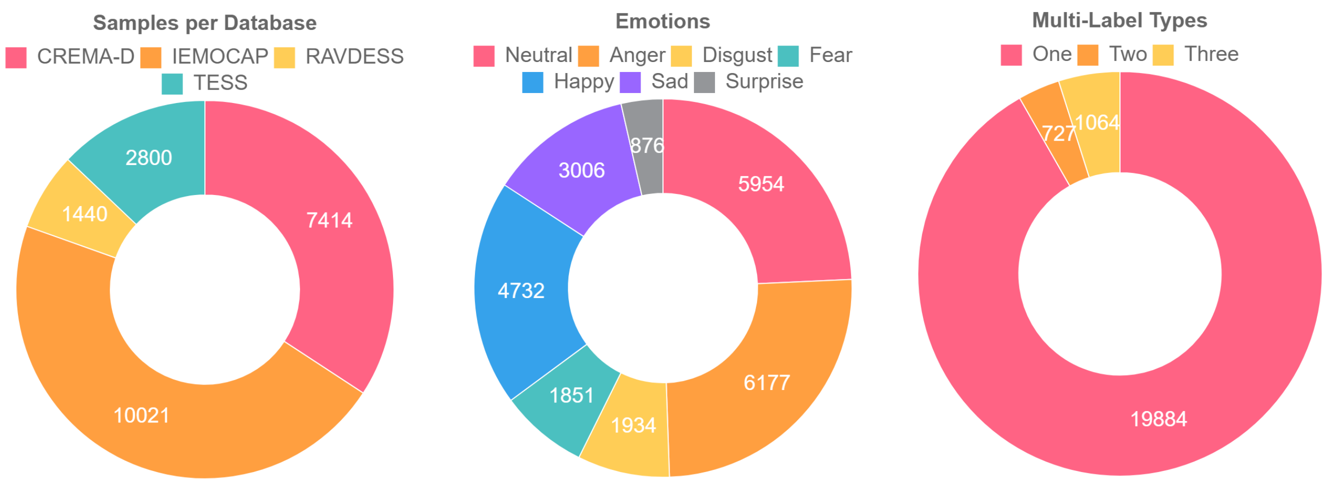

Fig. 1: Proportions of each database, emotion, and label types in the combined database.

1) A sample starts as a raw waveform in the form of time series points specifying the amplitude at a certain 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 cropped and the extra information 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 short-time Fourier transform (STFT) and Mel scale equation. The phase information was discard as it does not seem to hold relevant information. [18]
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.

The discrete-time STFT is defined in (1) where $x[n]$ is the discrete signal and $w[n]$ is the window function.

$$\mathbf{STFT}\{x[n]\}(m,\omega) \equiv X(m,\omega) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-j\omega n} \tag{1}$$

The Mel scale is defined in (2) where f is the frequency in hertz and m is mels. [19]

$$m = 2595log_{10}(1 + \frac{f}{700}) \tag{2}$$

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. [6] where they also use 75% overlap but with a smaller window size (2048) and they achieved excellent results. For the Mel spectrogram, 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. This is due to the Gabor limit.

B. Neural Network

After all of the databases were processed this way, the final combined database was fed into a neural network for training, validation, and testing. We constructed a eight layer neural network consisting of four convolutional layers and three dense layers. The network architecture is shown in figure 3. 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. [20], [21]
- 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 II describes the network hyperparameters and table III describes the training hyperparameters.

The final model is evaluated on the testing set using two measures: categorical cross-entropy and accuracy. Categorical cross-entropy as defined in (3) where N is the number of samples, k is the number of classes, $t_{i,j}$ is 1 if the sample i is in class j and 0 otherwise, and $p_{i,j}$ is the model's predicted probability that sample i is in class j.

$$CCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k} t_{i,j}log(p_{i,j}) \tag{3}$$

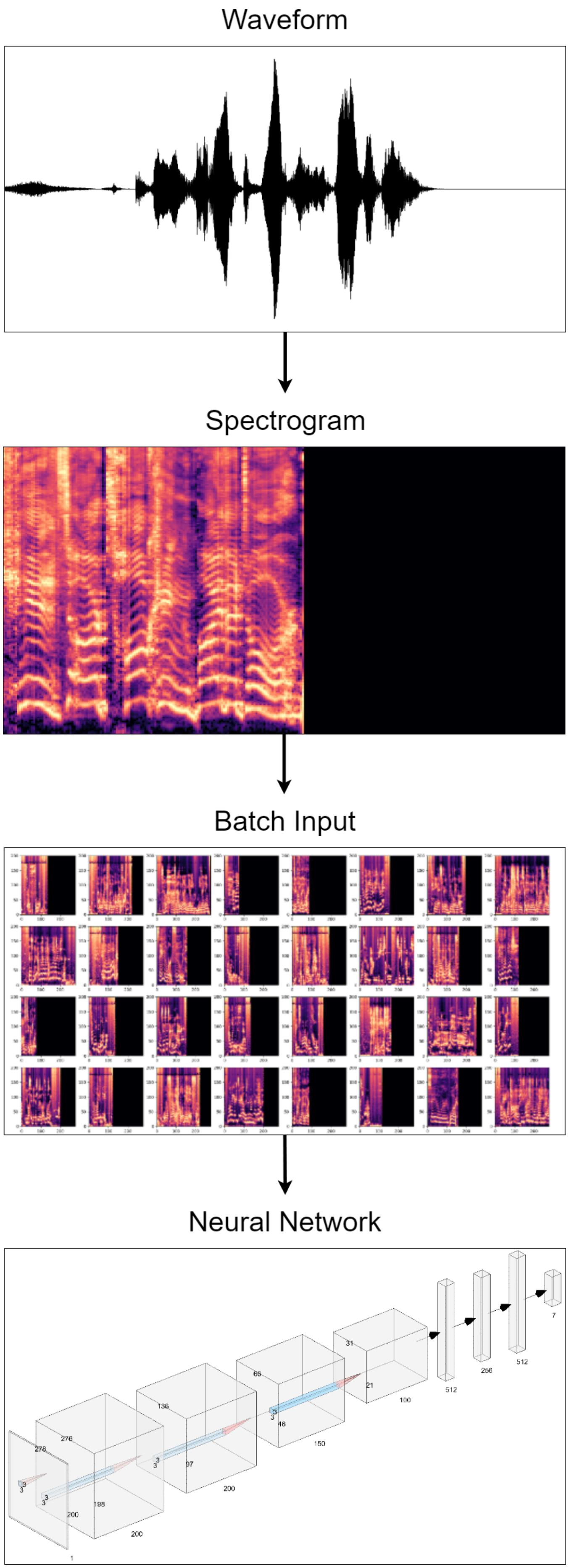

Fig. 2: A high level overview of the processing stages that a speech sample goes through.

TABLE II: The model's hyperparameters.

Hyperparameter	**Value**
Input Dimensions	278w x 200h x 1d
Optimization Algorithm	Adam
Loss Measure	Binary Cross-entropy
Accuracy Metric	Categorical Cross-entropy
Activation Function	Sigmoid

TABLE III: The training hyperparameters.

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

The choice to use categorical cross-entropy is based on its ease of use and how prevalent it is used in mutli-label problems.

The second measure, accuracy, is calculated from the confusion matrices shown in 4. The confusion matrices are calculated by comparing the true labels from the predicted labels for each emotion. There is a confusion matrix for each emotion due to the multi-label problem considering each emotion as independent. We define accuracy as the unweighted average accuracy of all emotions and this is mathematically described in (4) where n is the total number of emotions (7 in this 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 is the total number of predictions.

$$Accuracy = \frac{\sum_{i=1}^{n} \frac{TP_i + TN_i}{Total_i}}{n} \tag{4}$$

We chose this definition of accuracy so that we can compare our results to other papers that also use a similar measure.

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 52.57%. Confusion matrices for each emotion are shown in Figure 4.

We created a neural network model that achieved a 52.57% accuracy on the problem of multi-label speech emotion classification. This is better than the state-of-the-art result of 52.14% achieved in the single-label case. [6] So we conclude that it is possible to achieve a reasonable accuracy on the problem of multi-label speech emotion recognition by applying deep learning to the problem.

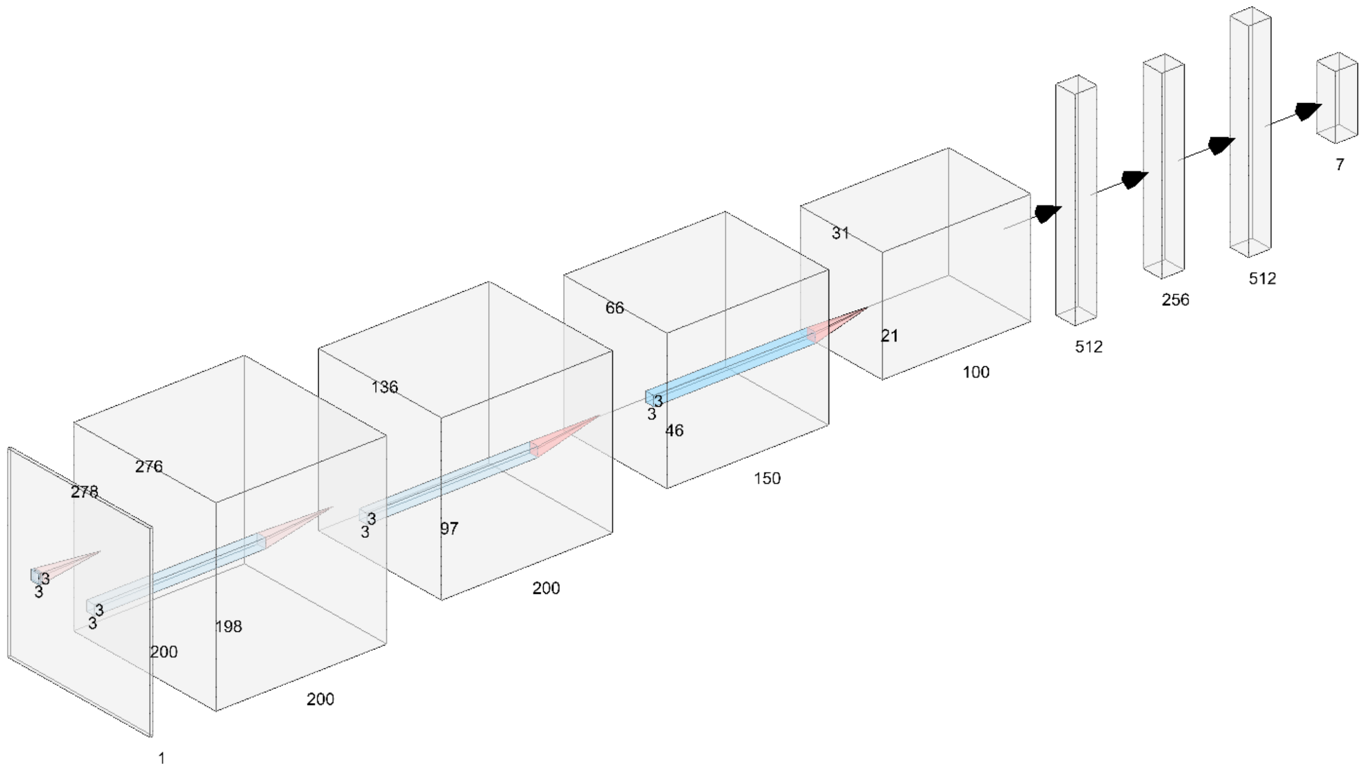

Fig. 3: Architecture for the speech emotion recognition network including the input layer and eight processing layers.

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 1. The ”surprise” class was the least represented class out of the seven classes and this makes the 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 52.57% 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. [22] 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. [23]

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

Acknowledgment

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

References

[1] 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

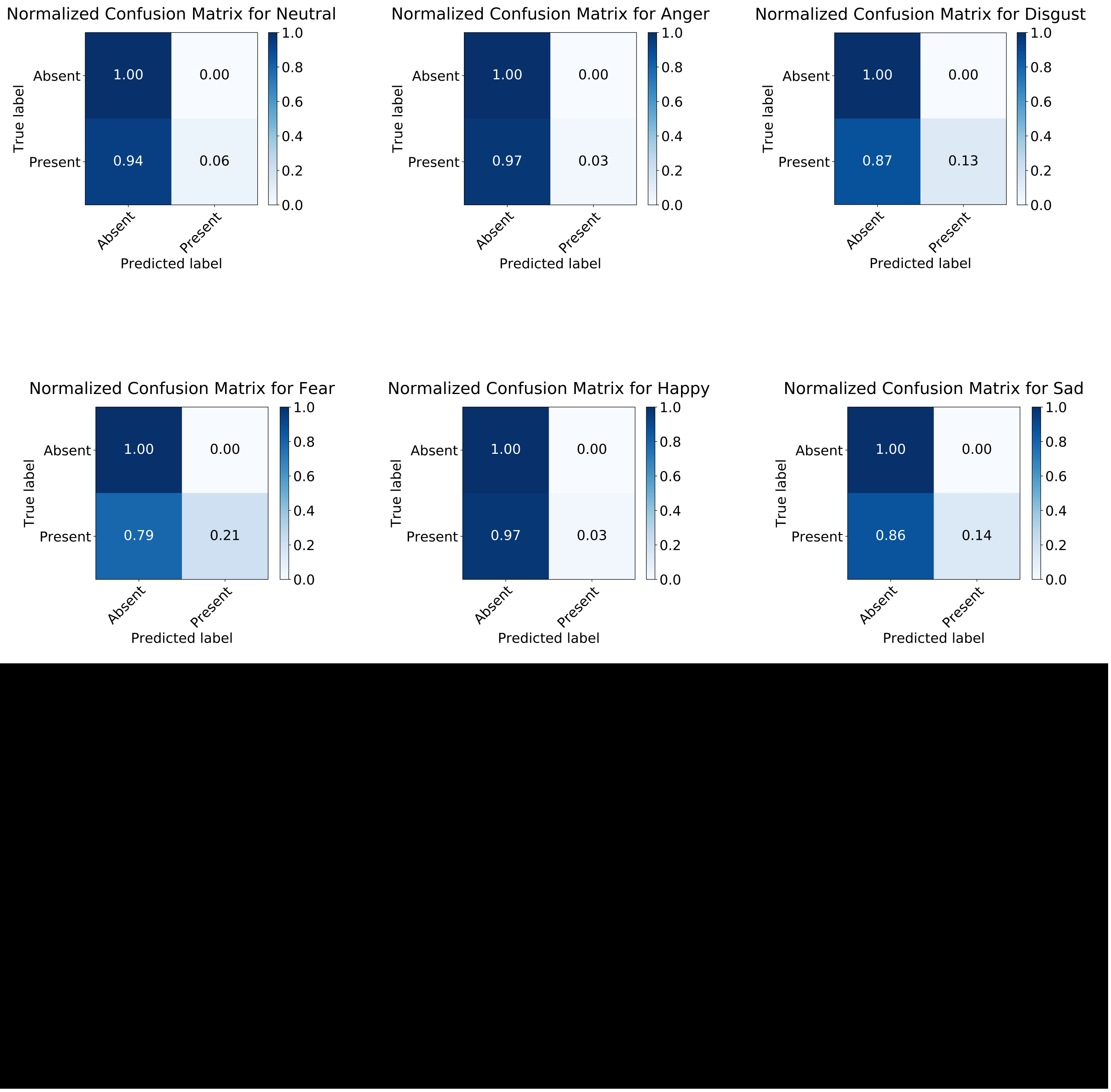

Fig. 4: Confusion matrices for each emotion.

[2] 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.

[3] 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

[4] 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/

[5] 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

[6] 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

[7] 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/

[8] 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.

[9] 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.

[10] 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/

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

[12] 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

[13] 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.

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

[15] 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

[16] 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.

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

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

[19] D. O'Shaughnessy, *Speech communication*. Addison-Wesley Pub. Co., 1990.

[20] 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

[21] 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

[22] 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.

[23] 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

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