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

Using machine learning to identify multiple emotions in speech

Brian Pho
University of Calgary
Calgary, Alberta
brian.pho@ucalgary.ca

Thomas Truong
University of Calgary
Calgary, Alberta
thomas.truong@ucalgary.ca

Svetlana Yanushkevich
University of Calgary
Calgary, Alberta
syanshk@ucalgary.ca

ABSTRACT

A problem with current speech emotion recognition systems is that they can only recognize one emotion from a set of emotions in speech. This is a problem because it does not match how people recognize emotions in speech. When people hear speech, they can perceive multiple emotions instead of a single emotion. We address this problem by building a 2D convolutional neural network that can also recognize multiple emotions and it achieves an accuracy of 57.64%. This model demonstrates a more realistic approach to speech emotion recognition by more closely matching human data.[1]

CCS CONCEPTS

• **Applied computing → Sound and music computing**;

KEYWORDS

multi-label, speech, emotion, recognition, neural networks

ACM Reference Format:
Brian Pho, Thomas Truong, and Svetlana Yanushkevich. 2020. Multi-Label Speech Emotion Recognition Using 2D Convolutional Neural Networks: Using machine learning to identify multiple emotions in speech. In *Proceedings of ACM SAC Conference (SAC'20).* ACM, New York, NY, USA, Article 4, 7 pages. https://doi.org/xx.xxx/xxx_x

1 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 [18].

1.1 Literature Review

SER has historically been approached in two ways:

- Feature engineering
- 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) [26]. Nassif et al. [23] 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 [23].

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 [13], [7], [2], and [30] 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 [30]. 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. [17] 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 [5] and CREMA-D [6] 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

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

SAC'20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03…$15.00
https://doi.org/xx.xxx/xxx_x

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.

1.2 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:

- 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.

2 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 [13] and [1]. The choice of using a CNN is also based on previous work where Balakrishnan et al. [3] 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 [20].

A high level overview of the data flow is shown in figure 2 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 [22] and Keras [8] libraries respectively.

2.1 Preprocessing

To build a more realistic SER system, we first need to obtain speech samples with their labeled emotion. We considered eight databases and chose four based on accessibility and based on the number of overlapping emotions. Table 1 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 [5]
(2) TESS [11]
(3) RAVDESS [21]
(4) CREMA-D [6]

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 [12] 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 [9].

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 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 [19].
(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.

Table 1: Comparison of databases with their labeled emotions.

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

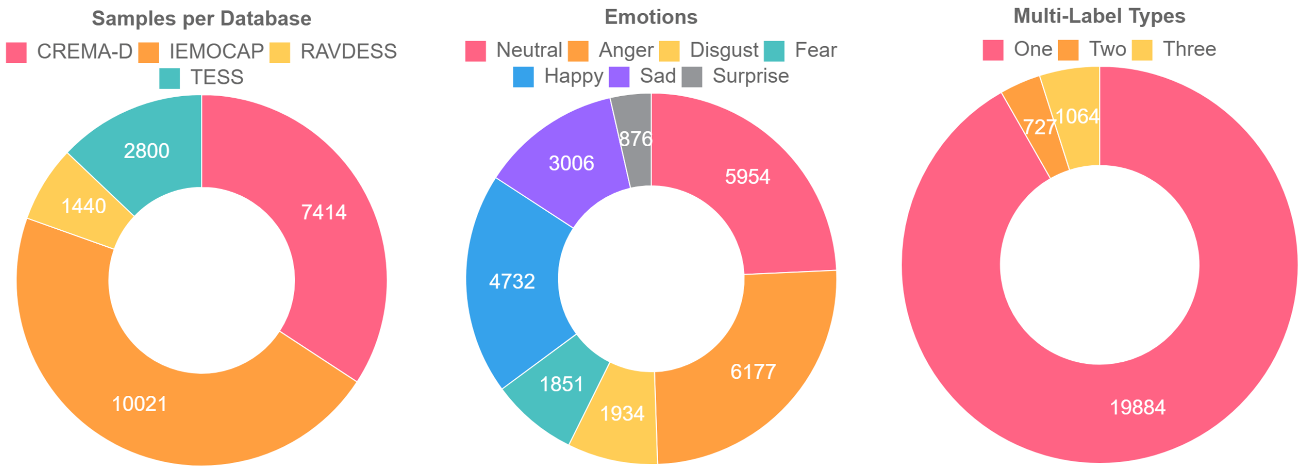

Figure 1: Proportions of each database, emotion, and label types in the combined database.

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

The discrete-time STFT is defined in (1) where $x[n]$ is the discrete signal and $w[n]$ is the window function. We used a window size of 3072 with a 75% overlap. This choice is based on the work of Zhao et al. [30] where they also use 75% overlap but with a smaller window size (2048) and they achieved excellent results.

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

The Mel scale is defined in (2) where f is the frequency in hertz and m is mels [24]. 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.

$$\begin{aligned} X_{std} &= \frac{X - X_{min}}{X_{max} - X_{min}} \\ X_{scaled} &= X_{std} \cdot (max - min) + min \end{aligned} \tag{3}$$

The min-max scaling is defined in (3) 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 [9].

2.2 Neural Network

After all of the databases were processed as described in section 2.1, 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 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.

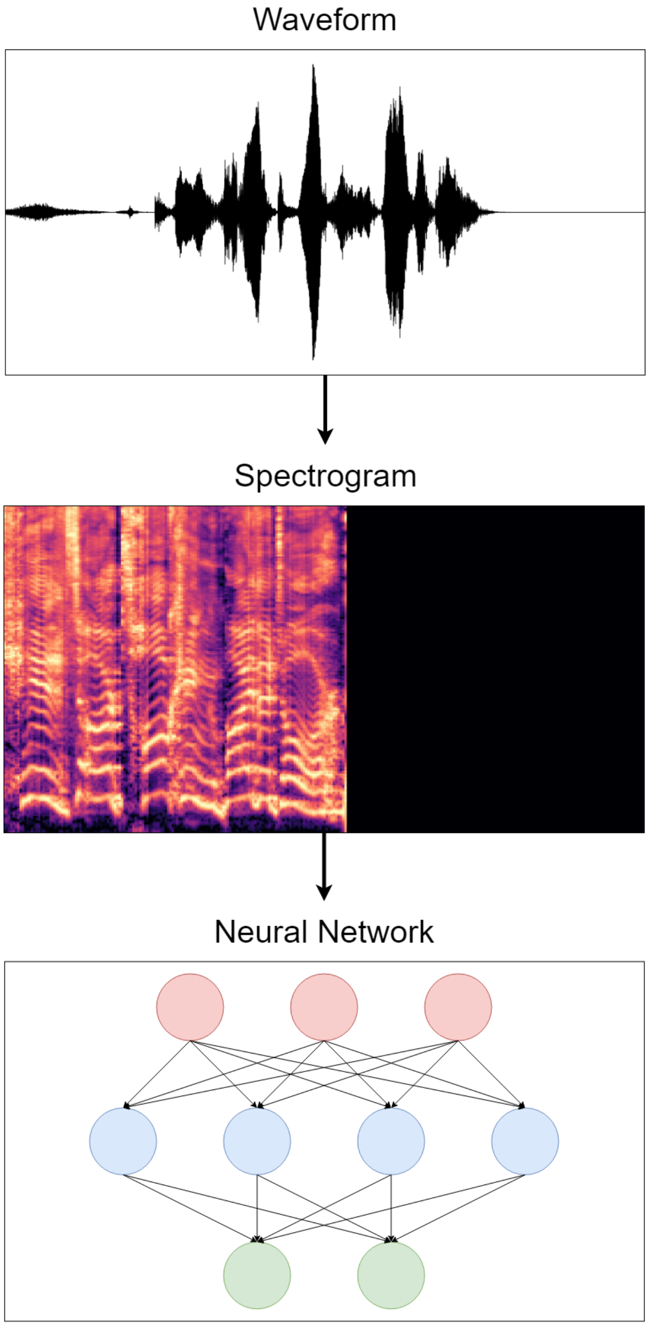

Figure 2: High level overview of the processing stages that a speech sample goes through.

- We applied dropout to the dense layers and batch normalization to the convolutional layers to deal with overfitting [27], [16].
- We updated the model's hyperparameters based on the validation loss and accuracy to improve the model's accuracy and ability to generalize.

Table 2: 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 3: The training hyperparameters.

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

- We used class weights during training to address class imbalance.

Table 2 describes the model's hyperparameters and table 3 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 4. 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 (4) 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.

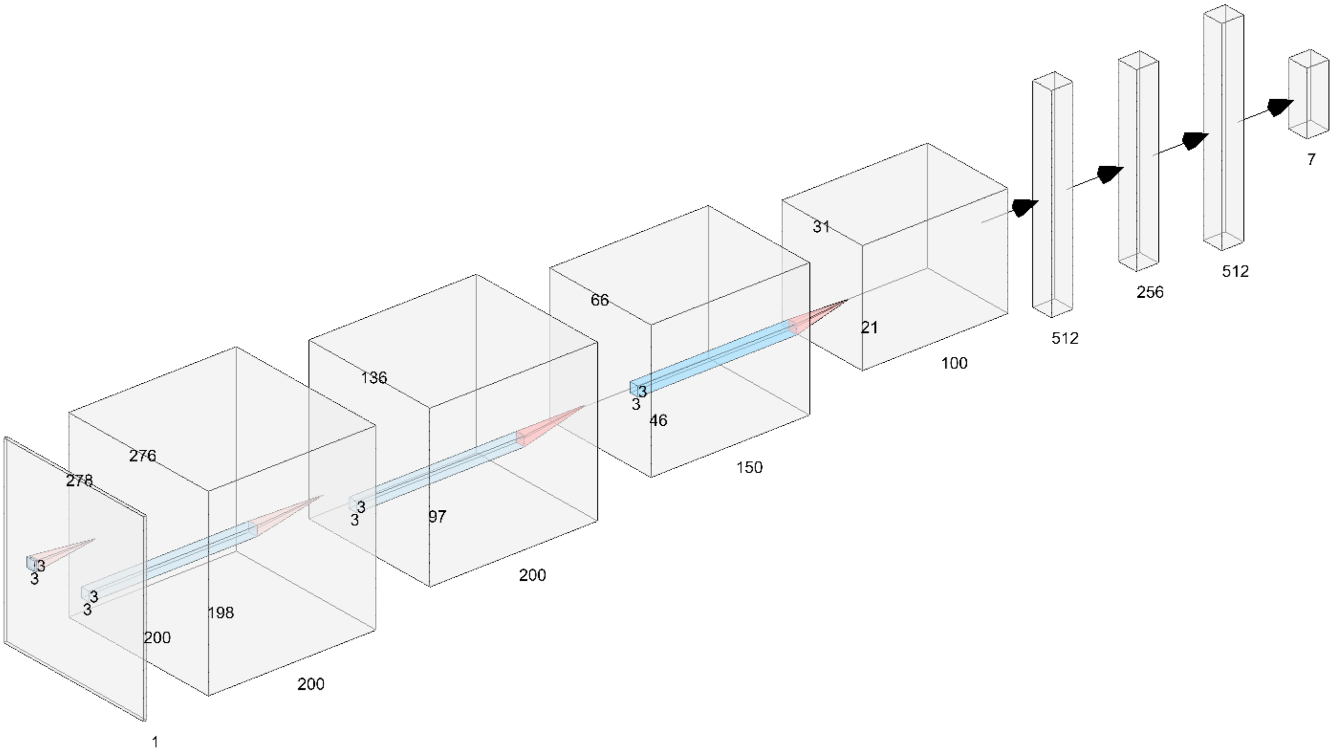

Figure 3: Architecture for the SER network including the input layer and eight processing layers.

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

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.

3 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%. Confusion matrices for each emotion are shown in Figure 4 and comparisons to the current literature is shown in table 4. 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 4, 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 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.

4 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 [10]. 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

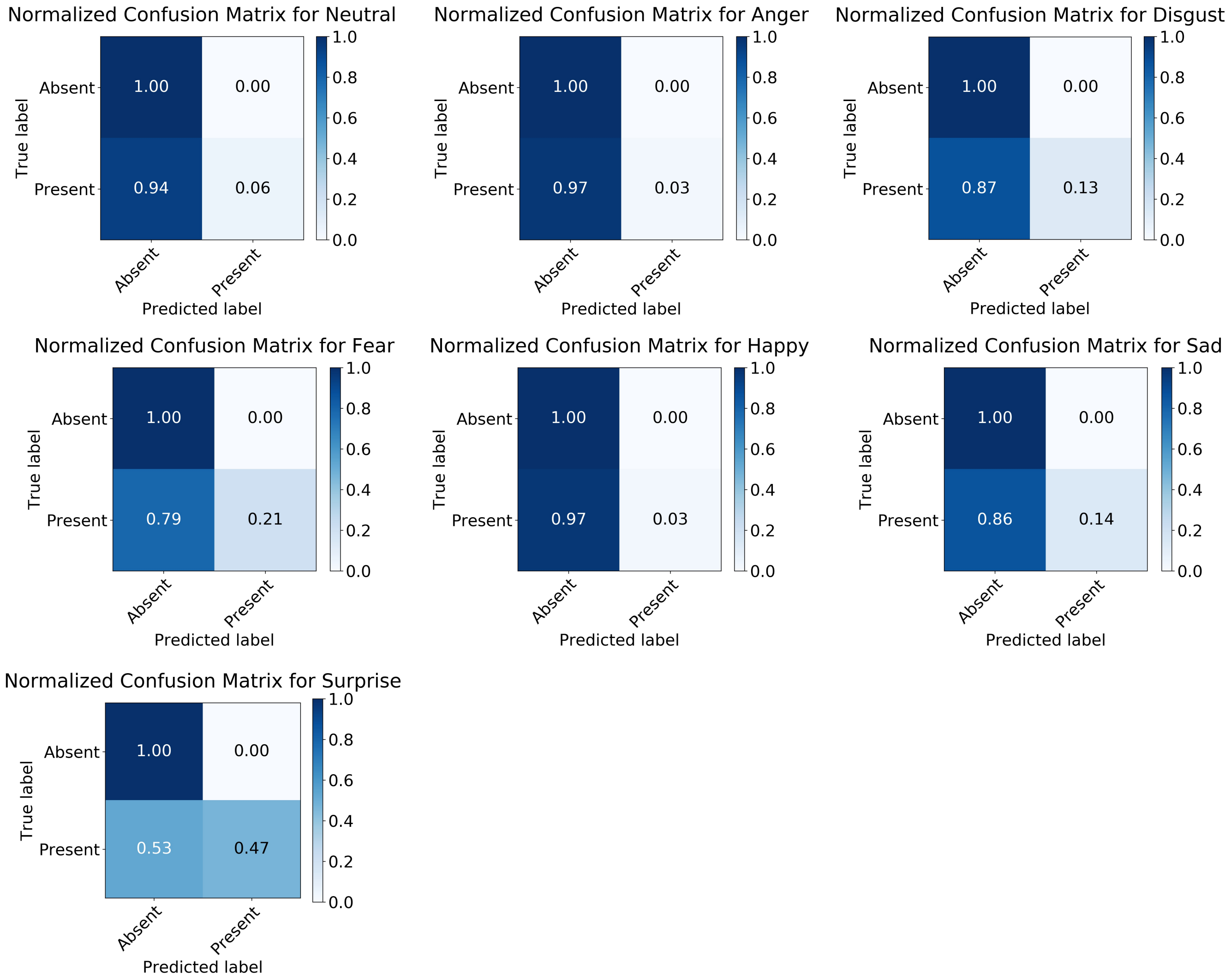

Figure 4: Confusion matrices for each emotion.

Table 4: Comparison of SER accuracy in literature.

Research Work	Method	Testing Method	Number of Emotions	Databases Used	Label Type	Accuracy (%)
Zhao et al. [30]	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. [14]	CNN + LSTM	Testing set	4	IEMOCAP	Single	61.7
Zhang et al. [29]	CNN	Testing set	4	IEMOCAP	Single	63.9
Fayek et al. [15]	CNN	Testing set	4	IEMOCAP	Single	64.8
Yenigalla et al. [28]	CNN	Testing set	4	IEMOCAP	Single	73.9
Badshah et al. [2]	CNN	Testing set	7	Emo-DB, Korean dataset	Single	80.8

as the Emo-DB database, would help the model generalize across languages [4].

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

ACKNOWLEDGMENTS

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

REFERENCES

[1] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. 2017. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In *2017 International Conference on Platform Technology and Service, PlatCon 2017 - Proceedings*. IEEE, 1–5. https://doi.org/10.1109/PlatCon.2017.7883728

[2] Abdul Malik Badshah, Nasir Rahim, Noor Ullah, Jamil Ahmad, Khan Muhammad, Mi Young Lee, Soonil Kwon, and Sung Wook Baik. 2019. Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications* 78, 5 (mar 2019), 5571–5589. https://doi.org/10.1007/s11042-017-5292-7

[3] Anusha Balakrishnan and Alisha Rege. 2017. *Reading Emotions from Speech using Deep Neural Networks*. Technical Report.

[4] Felix Burkhardt, Astrid Paeschke, M. Rolfes, Walter F. Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech.. In *INTERSPEECH*. ISCA, 1517–1520. http://dblp.uni-trier.de/db/conf/interspeech/interspeech2005.html#BurkhardtPRSW05

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation* 42, 4 (2008), 335–359. https://doi.org/10.1007/s10579-008-9076-6

[6] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2014. CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset. *IEEE Transactions on Affective Computing* 5, 4 (2014), 377–390. https://doi.org/10.1109/taffc.2014.2336244

[7] Mingyi Chen, Xuanji He, Jing Yang, and Han Zhang. 2018. 3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition. *IEEE Signal Processing Letters* 25, 10 (oct 2018), 1440–1444. https://doi.org/10.1109/LSP.2018.2860246

[8] Franï£¡ois Chollet. 2015. Keras. https://github.com/keras-team/keras

[9] Francï£¡ois Chollet. 2017. *Deep Learning with Python* (1 ed.). 101 pages.

[10] Daniel T. Cordaro, Rui Sun, Dacher Keltner, Shanmukh Kamble, Niranjan Huddar, and Galen McNeil. 2018. Universals and cultural variations in 22 emotional expressions across five cultures. *Emotion* 18, 1 (2018), 75–93. https://doi.org/10.1037/emo0000302

[11] Kate Dupuis and M. Kathleen Pichora-Fuller. 2011. Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set. *Canadian Acoustics* 39, 3 (Sep. 2011), 182–183. https://jcaa.caa-aca.ca/index.php/jcaa/article/view/2471

[12] Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3-4 (1992), 169–200. https://doi.org/10.1080/02699939208411068

[13] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. 2019. GANSynth: Adversarial Neural Audio Synthesis. (feb 2019). arXiv:1902.08710 http://arxiv.org/abs/1902.08710

[14] Caroline Etienne, Guillaume Fidanza, Andrei Petrovskii, Laurence Devillers, and Benoit Schmauch. 2018. CNN+LSTM Architecture for Speech Emotion Recognition with Data Augmentation. (feb 2018), 21–25. https://doi.org/10.21437/smm.2018-5 arXiv:1802.05630

[15] Haytham M. Fayek, Margaret Lech, and Lawrence Cavedon. 2017. Evaluating deep learning architectures for Speech Emotion Recognition. *Neural Networks* 92 (aug 2017), 60–68. https://doi.org/10.1016/j.neunet.2017.02.013

[16] Sergey Ioffe and Christian Szegedy. 2015. 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 (ICML'15)*. JMLR.org, 448–456. http://dl.acm.org/citation.cfm?id=3045118.3045167

[17] Yelin Kim and Jeesun Kim. 2018. 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, 5104–5108. https://doi.org/10.1109/ICASSP.2018.8462011

[18] Shashidhar G. Koolagudi and K. Sreenivasa Rao. 2012. Emotion recognition from speech: A review. , 99–117 pages. https://doi.org/10.1007/s10772-011-9125-1

[19] Piotr Kozakowski and Bartosz Michalak. 2017. DCGAN and spectrograms. http://deepsound.io/dcgan_spectrograms.html

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. 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., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[21] Steven R. Livingstone and Frank A. Russo. 2018. 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* 13, 5 (2018), e0196391. https://doi.org/10.1371/journal.pone.0196391

[22] Brian McFee, Colin Raffel, Dawen Liang, Daniel Patrick Whittlesey Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python.

[23] Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. 2019. Speech Recognition Using Deep Neural Networks: A Systematic Review. *IEEE Access* 7 (2019), 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

[24] Douglas O'Shaughnessy. 1990. *Speech communication*. Addison-Wesley Pub. Co.

[25] Rajib Rana. 2016. Emotion Classification from Noisy Speech - A Deep Learning Approach. (mar 2016). arXiv:1603.05901 http://arxiv.org/abs/1603.05901

[26] Jan Rybka and Artur Janicki. 2013. Comparison of speaker dependent and speaker independent emotion recognition. *International Journal of Applied Mathematics and Computer Science* 23, 4 (dec 2013), 797–808. https://doi.org/10.2478/amcs-2013-0060

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* 15, 1 (Jan. 2014), 1929–1958. http://dl.acm.org/citation.cfm?id=2627435.2670313

[28] Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. 2018. 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, 3688–3692. https://doi.org/10.21437/Interspeech.2018-1811

[29] Yuanyuan Zhang, Jun Du, Zirui Wang, Jianshu Zhang, and Yanhui Tu. 2019. 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., 1771–1775. https://doi.org/10.23919/APSIPA.2018.8659587

[30] Jianfeng Zhao, Xia Mao, and Lijiang Chen. 2019. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control* 47 (jan 2019), 312–323. https://doi.org/10.1016/j.bspc.2018.08.035