SPEECH EMOTION RECOGNITION OF POPULAR DATASETS USING A CONVOLUTION NEURAL NETWORK

Mohammed A, Garnet C, Benedict K, Stephen T
(Data Mining & Machine Learning – Group #7 – Mock Conference Paper)

Electrical and Computer Engineering - University of Calgary, AB, Canada

ABSTRACT

This paper captures research, development, implementation, and evaluation of a 5-layer CNN deep learning speech emotion recognition (SER) model. This work was conducted by Group-7 to satisfy final project requirements and explore team interests in speech emotion recognition. The main scope areas of the project include a preliminary literature search; data collection: exploration and assessment; pre-processing; model development and training; model investigations; and discussion of results. Based on the literature search a CNN architecture was chosen. In this study we use 4 popular datasets (Crema, Tess, Ravdess, Savee) and several data augmentations to balance and expand the data. Training, validation, and testing was carried out on the combined and individual datasets and results are evaluated and discussed. The final model had an accuracy of 48% on the test data. Conclusions are drawn on the effectiveness of different augmentations and data sets, and how they could be more effectively utilized in future models.

Index Terms— Speech emotion recognition (SER), deep learning, convolutional neural network (CNN), data augmentation, human-computer interaction (HCI)

1. INTRODUCTION

In terms of understanding and classifying actual human emotions, several models have been proposed in the literature, some categorical, others dimensional or hybrid. For example, the Ekman model categorizes emotions in six basic classes (happiness, sadness, anger, surprise, disgust, and fear) [4].

"Speech Emotion Recognition (SER) is the task of recognizing the emotional aspects of speech irrespective of the semantic contents" [1]. The goal of a SER model is to classify emotions based on speech.

The development of SER models is an area of ongoing research and is considered an important and complex aspect of Human-Computer Interaction (HCI) [1,2,6]. Accurate SER models will help robots/computers determine human emotional state, thereby enabling a more human-like response, and making it more effective when interacting with people [1,3,4].

More important in the near term, accurate SER models have the potential to improve many aspects of life including mental health assessment, public service call center prioritization, customer service, virtual assistants, and social media analysis [1-3,7]. SER models may also be able to provide accessibility support, aiding people with a difficulty in detecting emotional speech cues (e.g., autism).

As the available speech data and diversity grows, it is also conceivable that advanced language assistance apps can be built to help in understanding the unfamiliar prosody of a new language (i.e., intonation, stress, and patterns of rhythm). Prosodic attributes of speech are considered to "strongly influence emotion recognition" in human listeners [5]. By the same token, it is possible that with context of a speaker specific language and region, SER model accuracy may be improved. In fact, there are studies considering cultural and language influences on SER [7].

With the wide range of potential applications and benefits attributable to an accurate SER model, it is not surprising that speech emotion recognition is an active area of research [6]. SER can be considered a branch of Automated Emotion Recognition (AER). Speech emotion recognition is a challenging and open problem where research has been ongoing for over two decades [6]. Regardless of the approach or technique employed, a central problem for SER is that due to the subjective nature of emotions there is no consensus on measuring or categorizing them; and it is not uncommon for humans to misperceive or misinterpret them [7].

1.2. Motivation

The wide-ranging potential benefits and applications for accurate SER models have captured the interest of our Team. We find the potential for accessibility and language assistance type apps the most interesting, since they could mitigate communication difficulties experienced by those learning a new language, and possibly immersed in a new culture. Our world is becoming more and more connected and so we feel accurate SER tools are needed.

Also, we find the complexity of the speech recognition domain interesting and would like to expand our knowledge of SER data-centric aspects (speech datasets, pre-processing, and augmentation) since the data for SER is a known area of concern.

1.3. Objectives

In line with our motivations, the overall goal of this work is to satisfy project requirements and:

- Expand knowledge of SER and deep learning.
- Explore and assess speech emotion datasets.
- Compare CNN model performance between datasets.
- Assess data augmentation effects on emotion recognition.
- Generate helpful visualizations and summary statistics.
- Evaluate new CNN model architecture if time permits.

1.4. Organization

Section 2 provides supporting conceptual information as a precursor to method discussion. Section 3 outlines our materials and method. Section 4 captures results and discussion. Conclusions along with learnings, challenges, and potential future work are discussed in Section 5. References are included in Section 6.

2. RELATED WORK

The field of speech emotion recognition is large and growing. From our preliminary literature search we have found many interesting articles; however further detailed research is required to define the state of the art in SER. This said, our references in section 6 provide a good starting point for further research and future work.

The following paragraphs provide a summary overview of some of the main techniques and advancements made in SER as the research field has progressed [1,2,3,6,7].

Conventional machine learning approaches such as Support Vector Machine (SVMs), Hidden Markov Models (HMMs), Gaussian Mixture Models (GMMs), and shallow artificial neural networks were commonly used in early SER works. These conventional methods relied on using manually generated acoustic features (e.g., pitch, energy, temporal variations) to help recognize emotions in speech.

To improve SER accuracy researchers studied ways to integrate feature extraction and classification. This led to more complex feature extraction techniques such as Mel-frequency cepstral coefficients (MFCCs), which are still in use. Other complex features that have yielded good results in some cases include spectral roll-off, glottal waveforms, Teager Energy Operator (TEO), and spectrograms.

In recent years, the use of deep learning techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM) networks, have become increasingly popular in SER. Such models can automatically learn features from raw speech signals and can achieve state-of-the-art performance for some tasks. Use of DL models can result in significant time savings compared with conventional machine learning where iterative feature engineering is employed to find the best features.

Further SER performance improvements have been sought through the application of concatenated models (e.g., CNN and LSTM); multimodal approaches (e.g., use of speech audio and facial images or dimensional data, etc.); DL attention mechanisms to focus on pertinent audio signal segments; Multitask Learning (MTL) and adversarial networks, and transfer learning.

The different speech datasets/languages, potential features, parameters, pre-processing, network architectures, etc. generally render performance comparisons between SER models in the literature somewhat baseless. However, from our readings CNN-RNN (CRNN) and Domain Adversarial Neural Networks (DANN) models appear to represent the current state of the art in SER.

Example high performing CRNN models include: a 2D CNN LSTM with speaker-dependent accuracies of 89.16% and 95.33% on IEMOCAP and Berlin EmoDB databases, respectively [16]; and a 3-D attention-based convolutional LSTM (ACRNN) achieving 86.99% overall accuracy [2,17].

Work published at Interspeech 2020 shows a novel DANN which achieved 82.68% on the IEMOCAP database [3]. The comparison included in the referenced paper shows the DANN model weighted accuracy out-performing other DANNs (ranging up to 79.20%) by 3.48%. The authors attribute the improved performance to their utilization of contextual and multimodal information during training [3].

Other notable findings in terms of SER model performance include: a merged 1D and 2D CNN yielding speaker-dependent accuracies of 92.71% and 89.77% for EmoDB and IEMOCAP databases, respectively [18]; and a Deep Belief Network (DBN) with an accuracy of 94.6% when used with the Chinese Academy of Sciences (CAS) emotional speech database [19].

A key challenge in developing accurate SER models, is the lack of large quality datasets for training and evaluation [1,7]. Obtaining quality data is difficult for a variety of reasons: actor or induced speech lacks natural characteristics including common background noise; natural speech data collection raises ethical concerns; data collected may be imbalanced; and data labeling once collected is again subject to potential misinterpretation.

Some research has shifted toward semi-supervised learning to reduce the need for labeled data. Employing unsupervised approaches such as using encoders (e.g., Adversarial Autoencoders (AAE), Variational Autoencoders (VAE)) may result in loss of emotional characteristics due to compression of features [3]. Data augmentation also continues to be used to help bridge the speech emotion data gap. In recent years, transfer learning using high performing image classifiers trained on large datasets may also serve to reduce the time and data needed for training.

Multimodal models have the potential to outperform single modal emotion recognition models and mitigate the lack of quality audio data. Potential additional forms of insightful emotional data and information include facial expressions (e.g., using a facial mesh algorithm) [20,21]; bio-signals (e.g., electroencephalogram (EEG), electrocardiogram (ECG), blood pressure, etc.) [22,23]; gaze direction [24]; and body gesture/movement analysis [25].

Multimodal emotion recognition models are not without limitation however – beyond the new modal data, processing, training, and validation required, in some cases (e.g., network architectures, specific emotions) the added information may not significantly improve emotion recognition. Figure-1 demonstrates how additional modality can sometimes greatly improve emotion recognition (bimodal prediction of "happy", 875), while other times it may not (all modalities predict "neutral" well, 875) [8].

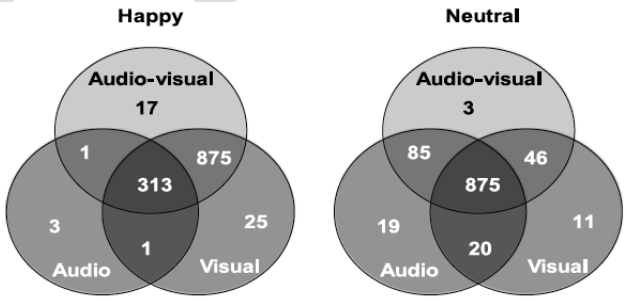

Fig. 1. CREMA-D emotion recognition Venn diagrams [8]

3. MATERIALS AND METHODS

The focus of our project is on further development of a single modal SER CNN, its performance, and data-related investigations. This section outlines our software and hardware tools, as well as the main approach and workflow areas: Data collection; Data exploration and assessment; Data pre-processing; Classification model and training; SER model investigations.

In general, with respect to materials our team collected publicly available speech emotion data; and for equipment, we availed of a team member's home-based GPU (Nvidia RTX 3080) to facilitate data processing and model work. In terms of software, we used a single Jupyter Notebook file and built our workflow around PyTorch and Librosa (a popular data science library for

audio data), as well as some utilities from other libraries such as scikit-learn, Pandas, NumPy etc. For reference, our project notebook and data files can be found in the team repo located on GitHub [14]. Project notebook references are included in Section 6 [26,27,28].

Based on the high accuracies observed for CNNs in the image domain [15], we decided on a CNN for our SER single modal use case. From our preliminary literature search, we found that CNNs can also yield high accuracies in the SER domain, supporting our model choice. Pipeline-wise we chose to train our model with pre-extracted features. Computing features in advance broke-up the work and compute time and allowed experimenting with features.

3.1 Data collection

For our project we used several datasets in a systematic manner to better understand the data and their influence on model performance. It is hoped that comparing different dataset results may reveal some insights on how to improve our CNN SER model, and on the applicability of our model.

Table 1. Project datasets summary [8-11]
(A-Audio, V-Visual, A-V-Audio-Visual, na-not available)

Dataset	CREMA-D	RAVDESS	SAVEE	TESS
No. of people - male / female	48 / 43	12 / 12	4 / 0	0 / 2
Natural vs. actor	actor	actor	actor	actor
Age range	20 - 74	21 - 33	27 - 31	26 - 64
No. of speech samples	7,441	1,440	480	2800
No. ratings per sample	9.8 (avg.)	10	10	na
No. of raters	2,443	319	10	na
Language	English	English	English	English
Modalities	A, V, A-V	A, V, A-V	A, V, A-V	A
Audio sampling rate (kHz)	44.1	48	44.1	24.4
Audio bit depth	16	16	16	16
Audio file type	WAV	WAV	WAV	WAV

The four pre-labeled datasets selected for the project are summarized in Table-1 above along with some key attributes. As can be observed, there appears to be a significant number of quality audio samples to work with – particularly considering the CREMA-D and RAVDESS datasets. Figure-2 illustrates the proportions of each of the 7 emotions in the combined dataset. We chose the four noted datasets because they are well known, easily accessible, and have a similar length and format.

The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D) is a large dataset generated by using a group of 91 English-speaking actors diverse in age and ethnicity. A large and diverse group of raters, with 2443 participants, was used to obtain ratings [8].

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was also generated using a diverse group of English-speaking actors. The 24 actors and 319 raters were selected from the Toronto area in Ontario, Canada [9]. As noted in the dataset title, the audio files include speech and song clips; however, we chose not to use the song clips for our project.

The Surrey Audio-Visual Expressed Emotion (SAVEE) is an imbalanced database comprised of 480 British English-spoken recordings of 7 emotions from 4 male actors. Of note, the SAVEE database was used to build an AER classifier using 3 modalities. Emotion recognition rates of 61%, 65% and 84% were realized for audio, visual, and audio-visual modalities, respectively [10].

The TESS dataset is comprised of 2800 audio recordings from two female actors portraying seven emotions. Both actors were from the Toronto area of Ontario, Canada [11].

Other datasets found in our search which have been identified for potential future work include: the English language Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database, in English; the Berlin Database of Emotional Speech (EmoDB), in German; and the Airplane Behavior Corpus (ABC), in German.

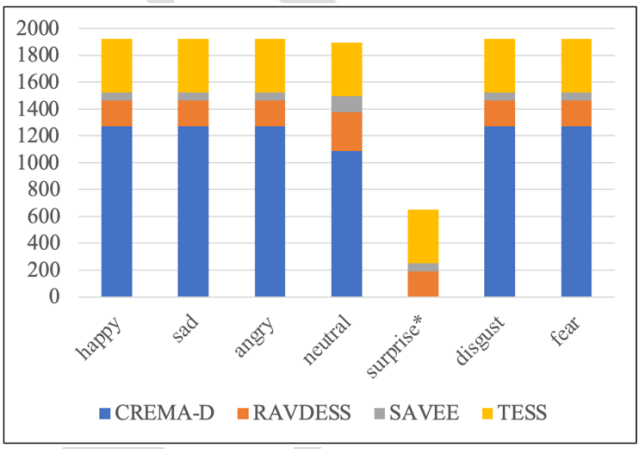

Fig. 2. Original dataset emotion counts (**pleasant surprise)*

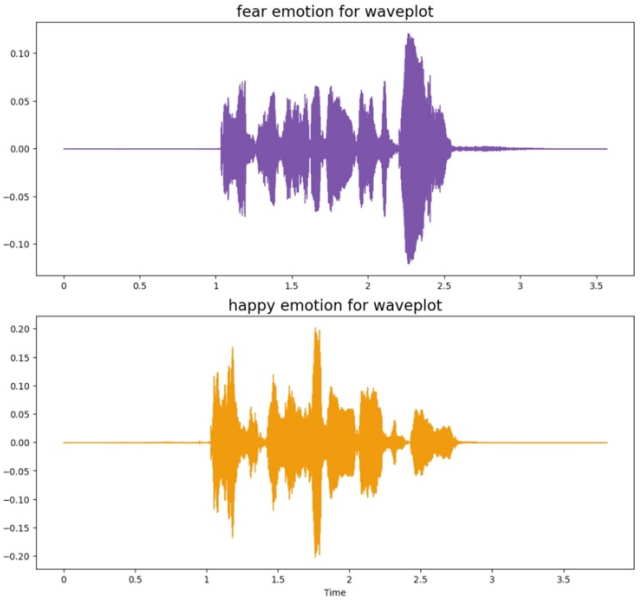

Fig. 3. Sample "fear" and "happy" waveform comparison

3.2 Data exploration and assessment

Each dataset was individually explored and assessed for balance in advance of generating predictions via our CNN model. Data augmentations are applied later in the workflow to correct imbalances as required. As can be observed from Figure-2, the datasets are generally balanced across all emotions. The two exceptions are "neutral" and "surprise", which are slightly and grossly imbalanced, respectively. In fact, "surprise" is missing entirely from the Crema dataset. From Table-1 we see that the combined dataset is also gender and age imbalanced.

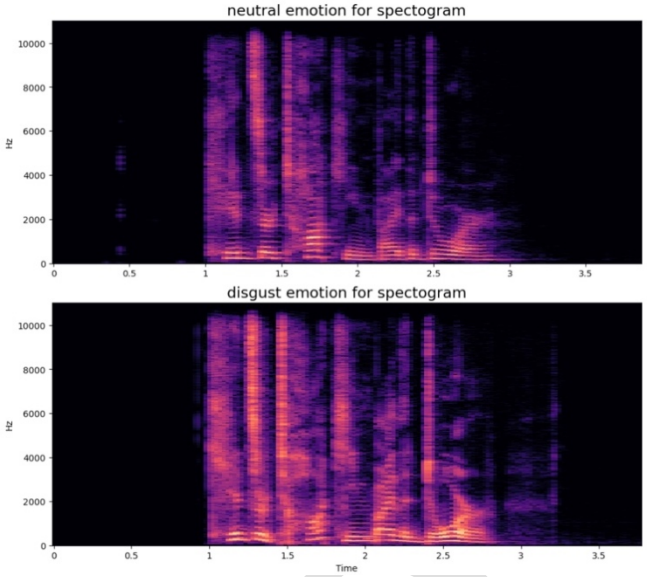

Fig. 4. Sample "neutral" and "disgust" spectrograms

To facilitate loading data for further exploration / assessment the team created a custom data loader class with functionality to load individual datasets or the combined dataset into a Pandas dataframe. Once the dataframe was obtained, we checked that the files were loaded correctly and with the expected counts.

After identifying data imbalances, we created visualizations of sample speech clips using waveform (time domain) shown in Figure-3 and spectrogram (frequency domain) plots shown in Figure-4 to assess potential differences in emotion sound character. In addition, to satisfy quality considerations and curiosity, the team listened to a variety of data samples from the notebook using IPython functionality.

3.3 Data pre-processing

After familiarizing ourselves and assessing the data, we carried out pre-processing to generate our balanced dataset for SER model training and testing. To facilitate this work, custom AudioLoader and DataBalancer classes were built. The AudioLoader class has methods to load, pre-process, and extract features. The DataBalancer houses data augmentation methods and instantiates an AudioLoader object to start the pre-processing pipeline.

Data pre-processing included trimming the durations of the audio files (to 2.5s) and adding a small start time offset (0.6s). These steps were taken to reduce file sizes and compute time, while still capturing the central target speech data. The audio clips were also down-sampled to 22.05kHz – we felt this was a reasonable compromise between compute time and adequate WAV file resolution.

Next, the audio file features are extracted. Initially, we generated 3 features for each file: Root Mean Squared Error (RMSE), Zero-crossing Rate (ZCR), and Mel-frequency Cepstral Coefficients (MFCC). We felt a blend of time and frequency domain features would be worthwhile; MFCCs are considered powerful features and some SER models still use them [7,13]. Later we also added spectrogram features as an experiment but reversed this modification due to poor model results.

Data augmentation was not only used to correct imbalances, but also to synthesis additional data. Once the original audio files were processed, the DataBalancer class identified the emotion label with the highest count and then proceeded to generate randomized augmentations until each class count was 1.5-times the highest. Data augmentations implemented included adding noise, shifting, changing pitch, and stretching duration. The dataset dataframe was then saved to a csv-file for quick retrieval, and we listened to some augmented audio for quality checking. The final balanced dataset we used for our project contained 2885 samples for each emotion.

Next, the 1D labels tensor was one-hot-encoded and then the labels and features (2D tensor) data were shuffled and split into training and test sets (80/20 split). The training set was subdivided to create a validation set (90/10 split). Finally, the feature subsets were normalized using scikit-learn Standard Scaler (i.e., each feature vector is subtracted by its mean and then divided by its standard deviation). Normalization is recommended to reduce compute time (convergence) and improve model performance.

3.4 Classification model and training

For our mono (single channel) audio feature data we used PyTorch to implement a deep learning convolutional neural network. The network has 5x 1D convolutional layers with ReLU (Rectified Linear Unit) activations, 1 fully connected layer with ReLU activations, and a final fully connected layer with Softmax activations to obtain predicted emotion class probabilities. Each of the convolutional layers were followed by a batch-normalization layer to help improve stability and speed of training, reduce overfitting, and improve model generalization performance. After each normalization layer a max-pooling layer was used to extract features (padding was needed due to input tensor and filter sizes, and so dimensionality was maintained).

To enable and facilitate training and validation, we implemented iterative training and validation functions, an EarlyStopper class; a PyTorch optimizer (adam) and a learning rate scheduler, as well as a batch data loader. Based on memory limitations we restricted batch size to 8. The performance of the model during training was evaluated using cross entropy loss on the validation data. Model training and validation history data collection (average loss and accuracy) was accomplished using an outer for-loop bounded by our specified max number of epochs (100). During each iteration of the outer loop, training and validation history data is accumulated and the EarlyStopper instance is used to check whether a new best (lower) loss has been achieved, or whether model improvement is plateauing (i.e., one iteration closer to hitting the patience variable (12)). Using the history data noted, visualizations of training and validation progress were generated.

3.5 SER model investigations

The model was created using 1D convolutions in the Pytorch Library. The audio data was loaded with a duration of 2.5s and resulted in a total of 4590 features per sample being extracted from the methods described above. This 1 x 4590 input vector was initially expanded into 512 channels in the first convolutional layer, and then using progressive max pooling and convolutional layers reduced in size to a reasonable number of inputs for the final fully connected layers. The architecture is shown in Figure-5.

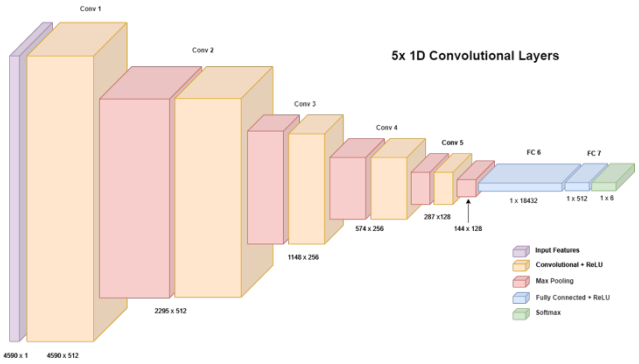

Fig. 5. Final convolutional neural network architecture

4. RESULTS AND DISCUSSION

The model was initially trained using all datasets. The results showed the emotion of surprised as an outlier being significantly more accurate than other emotions; we postulate that this was due to the largest dataset (Crema) missing the emotion. To balance the samples we performed excessive data augmentation. This means each surprise sample is included in the final dataset ~3 times with different augmentations and the CNN is effectively memorizing the samples. A solution to the problem would be to increase the sample size by recording more samples, however due to time constraints we elected to drop the emotion since it was artificially increasing the overall model accuracy. Further runs discussed include only the other 6 emotions.

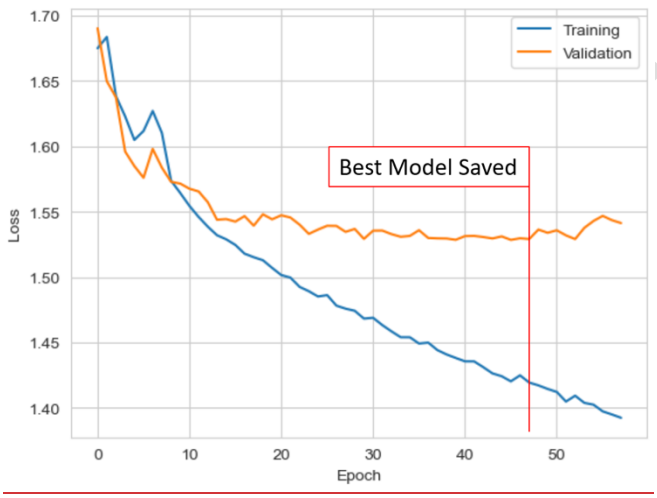

Fig. 6. Training and validation loss history plot

The model was retrained on all data sets together to evaluate performance and how augmentations affected the accuracy. The model trained for 57 epochs before early stopping after 12 runs without improvement. The loss chart from training is shown in Figure-6 above. The training loss was continuing to decrease even though the validation loss was stabilizing. Potentially the model was beginning to overfitting the training data singe the validation loss was stabilizing. We could potentially have improved this with a smaller patience and ideally pulled the best model somewhere around 30 epochs when the two began to diverge.

The results against the test set were reasonably strong with an accuracy of 47.9% which is significantly improved over random chance of 17% for a balanced 6 category problem. Looking at the confusion matrix shown in Figure-7, the model performs worst for the sad emotion only getting about half as many right as the other emotions. Overall, the errors are quite randomly distributed, it does not appear that any one label is being over predicted by the model. This is the ideal result since we are equally concerned about all emotions. The model seems better at differentiating the more negative emotions (disgust, fear, sad, and angry) from positive and neutral emotions. This is demonstrated by the majority of errors for Happy falling under Neutral and vice versa; and relatively few false predictions of negative emotions. This indicates there might be a more distinct difference between positive and negative, rather than the individual emotions; and the model may be better suited to differentiation the 2 opposing overarching categories.

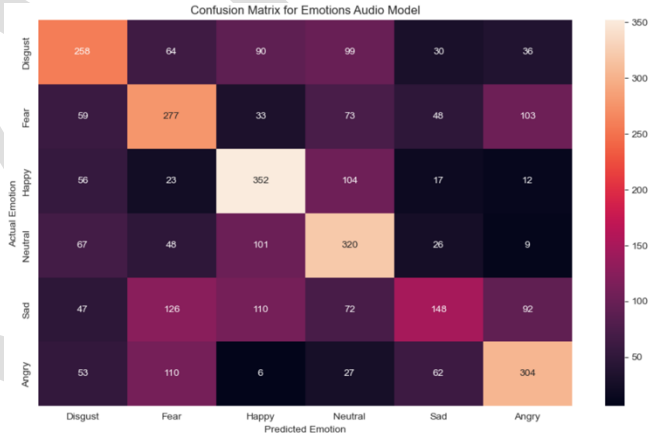

Fig. 7. Confusion matrix of full data set model

The final model was also evaluated to see how it performed against the different base datasets and augmentations that were used. These results for the test datasets are shown in Figure-8, the accuracies were evaluated against the test set only. Among the different datasets for most emotions the model is performing better on the Crema and Tess datasets. This is most likely because these two datasets contain the most samples, so the training is biased towards the actors and phrases used in these datasets. This result shows the major difficulty in SER of being able to generalize a model to people who have different ways of speaking and expressing emotions through speech and being able to develop high quality datasets.

Breaking the data down by augmentation results in some interesting outliers. The audio that was pitched has a very good accuracy for happy emotions and poor accuracy for most other emotions. This likely indicates that the pitch of an audio sample is a key feature the model is looking at to distinguish between

emotions and for happy emotions making them higher pitch brings samples more into the range the model expects. However, for other emotions the model expects lower pitches and increasing the pitch causes it to make mistakes. This would mean we are changing the emotion through augmentations and dropping pitching as a potential augmentation would result in a stronger model. We expected adding noise would make the samples harder to classify, that seems to be the case somewhat for sad, fear, and neutral emotions. Overall, the results are similar to non-augmented data, and since noise is very common in real-life recordings this would be a good augmentation to focus on for further model training to increase the amount of data to train on. The shifting and stretching emotions performed similarly to having no augmentation showing this was an effective way to increase the size of the dataset.

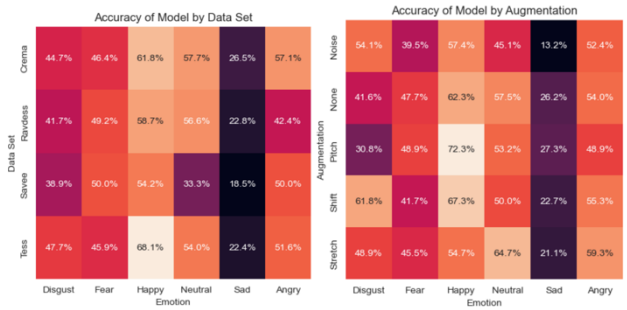

Fig. 8. Heat maps based on data set and augmentation

To better evaluate the different datasets, we additionally trained the model using each set individually to see if the results changed. The accuracies obtained from the models are shown in Table-2 below. The Tess dataset being trained on its own performed substantially better than any other run. Investigating this further to a human observer the samples for different emotions seem to be much more distinct between different emotions and similar within one emotion in the Tess dataset than the others. It is likely that the model is being severely overfit to the way the two actors are expressing emotion. Even though the results in our test set are very good since it comes from the same actors this model is very overfit and will not generalize well to different people talking in a normal way.

Table 2. Summary of dataset test results

Dataset	All	Crema	Ravdess	Savee	Tess
Accuracy	47.9%	43.3%	54.2%	46.8%	93.6%

The Crema dataset had the worst overall accuracy even though it has the most samples. Looking at the samples where there are misclassifications many of the mistakes seem to be poorly labelled. It is quite difficult in many samples to tell the difference between different emotions of the same phrase. This is likely due to the crowd sourced nature of the model, which captured a wider variety of lower quality data. This makes it difficult to build a strong model. The model performance could likely be improved by ignoring the crema dataset, since it seems to be of a lower quality then the others, however the combined model would likely generalize better since there are a lot more actors included.

Interestingly the combined dataset had a lower accuracy at 48% then the individual datasets, except Crema at 43%. This indicates that using the individual datasets for training, validation, and testing might be causing a bias towards the specific actors and sentences used when collecting the data. Although the combined model performs worse overall it has relatively consistent performance across the different datasets and is likely the most generalizable model to data it has not seen.

The model performs reasonably well overall with a 48% accuracy. There are several areas that could be looked at to improve the performance further. One of the most impactful improvements would be to improve the quality of the data sets. The Crema dataset in particular seemed to have samples that were hard to distinguish different emotions which could be a result of the crowd-sourced collection method. Going through and manually removing samples that do not convey much emotion could also improve the model performance. Another way to improve performance would be investigating different model architectures. Due to the time required to run the model we utilized a single 5-layer architecture which gave promising initial results for all runs, however there is almost certainly a more optimal architecture that could be obtained given more time to try different models. Finally including additional datasets could further improve the model performance and generalizability to real world situations. There are many more collections of labeled sound available that could be included given more time to collect them and integrate them into the pipeline.

5. CONCLUSION

The goal of this project was to investigate if emotion could be detected from human speech using deep learning methodologies. This is an interesting problem due to the many potential applications ranging from assisting autistic children in learning to understand emotion, to helping future AI/robotic systems have better interactions with humans. A literature review revealed this problem has been given a lot of thought in the past and is still an open area of research. The focus of this project is evaluating several existing datasets, and how augmentations affect performance of a SER convolutional neural network.

For this project four datasets were chosen that were collected in different ways. The datasets used are Crema-D, Ravdess, Savee, and Tess. A 5-layer convolutional layer neural network was trained on both the combined datasets, and each dataset individually to evaluate the differences. The combined model performed worse overall then most of the individually trained models. However, it performed similarly well against data from most of the sets indicating it learned a more generalized sample and is likely better for speech emotion data that is not from the actors used in the collected datasets.

Additionally, four randomized augmentations were performed on the audio samples before feature extraction, to increase the dataset size. The Noise, Shift, and Stretch audio did not substantially change the results, and were determined to be useful augmentations for this problem. The Pitch augmentation resulted in a substantial improvement for happy samples, and very poor performance for disgust and some other samples. This indicates that changing the pitch might be altering the underlying emotion in the sample, and Pitching is not a good solution for emotion augmentation and subsequent detection. Due to the high likelihood of having noise in recordings and the low impact of this augmentation on overfitting it is a good augmentation to use when the goal is increasing the amount of available information to train on.

6. REFERENCES

[1] M. Lech, M, Stolar, C. Best, and R. Bolia, “Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding,” Frontiers of C.S., vol. 2, May 2020, Art. no. 14, doi: 10.3389/fcomp.2020.00014 [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcomp.2020.00014/full.

[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124. [Online]. Available: https://ieeexplore.ieee.org/document/8805181.

[3] Z. Lian, J. Tao, B. Liu, J. Huang, Z. Yang, R. Li, “Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition,” Proc. Interspeech 2020, 394-398, doi: 10.21437/Interspeech.2020-1705. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2020/lian20b_interspeech.html.

[4] M. Spezialetti, G. Placidi, and S. Rossi, “Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives,” Frontiers in Robot. and A.I., vol 7, December 2020 Art. 532279, doi: 10.3389/frobt.2020.532279 [Online]. Available: https://www.frontiersin.org/articles/10.3389/frobt.2020.532279/full?utm_source=makeanapplike.

[5] A. Lausen, K. Hammerschmidt, “Emotion recognition and confidence ratings predicted by vocal stimulus type and prosodic parameters,” Humanit. Soc. Sci. Comm., vol. 7, Art. 2, June 2020. [Online]. Available: https://doi.org/10.1057/s41599-020-0499-z.

[6] J. de Lope, M. Graña, “An ongoing review of speech emotion recognition,” Neurocomput., vol. 528, pp. 1-11, April 2023. [Online]. Available: https://doi.org/10.1016/j.neucom.2023.01.002.

[7] M. B. Akçay, K. Oguz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” Speech Comm., vol. 116, pp. 56-76, December 19. [Online]. Available: https://doi.org/10.1016/j.specom.2019.12.001.

[8] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," in IEEE Transactions on Affective Computing, vol. 5, no. 4, pp. 377-390, 1 Oct.-Dec. 2014, doi: 10.1109/TAFFC.2014.2336244.

[9] Livingstone SR, Russo FA (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): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[10] P. Jackson and S. Haq. “Surrey Audio-Visual Expressed Emotion (SAVEE) Database.” http://kahlan.eps.surrey.ac.uk/savee/.

[11] Pichora-Fuller, M. Kathleen; Dupuis, Kate, 2020, "Toronto emotional speech set (TESS)", https://doi.org/10.5683/SP2/E8H2MF, Borealis, V1.

[12] McFee, Brian, Matt McVicar, Daniel Faronbi, Iran Roman, Matan Gover, Stefan Balke, Scott Seyfarth, Ayoub Malek, Colin Raffel, Vincent Lostanlen, Benjamin van Niekirk, Dana Lee, Frank Cwitkowitz, Frank Zalkow, Oriol Nieto, Dan Ellis, Jack Mason, Kyungyun Lee, Bea Steers, … Waldir Pimenta. (2023). librosa/librosa: 0.10.0.post2 (0.10.0.post2). Zenodo. https://doi.org/10.5281/zenodo.7746972.

[13] A. A. Abdelhamid et al., "Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm," in IEEE Access, vol. 10, pp. 49265-49284, 2022, doi: 10.1109/ACCESS.2022.3172954.

[14] Project (Group-7) Team GitHub repository, Accessed April 2023[Online]. Available: https://github.com/mjspk/ENEL645_Final_Project.git.

[15] Russakovsky, O., Deng, J., Su, H. et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115, 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y.

[16] J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” Biomed. Signal Process. Control, vol. 47, pp. 312 323, Jan. 2019.

[17] M. Chen, X. He, J. Yang, and H. Zhang, “3-D convolutional recurrent neural networks with attention model for speech emotion recognition,” IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440-1444, Oct. 2018.

[18] J. Zhao, X. Mao, and L. Chen, “Learning deep features to recognize speech emotion using merged deep CNN,” IET Signal Process., vol. 12, no. 6, pp. 713 721, 2018.

[19] W. Zhang, D. Zhao, Z. Chai, L. T. Yang, X. Liu, F. Gong, and S. Yang, “Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services,” Softw., Pract. Exper., vol. 47, no. 8, pp. 1127 1138, 2017.

[20] Ali I. Siam, Naglaa F. Soliman, Abeer D. Algarni, Fathi E. Abd El-Samie, Ahmed Sedik, "Deploying Machine Learning Techniques for Human Emotion Detection", Computational Intelligence and Neuroscience, vol. 2022, Article ID 8032673, 16 pages, 2022. https://doi.org/10.1155/2022/8032673.

[21] C. Dalvi, M. Rathod, S. Patil, S. Gite and K. Kotecha, "A Survey of AI-Based Facial Emotion Recognition: Features, ML & DL Techniques, Age-Wise Datasets and Future Directions," in IEEE Access, vol. 9, pp. 165806-165840, 2021, doi: 10.1109/ACCESS.2021.3131733.

[22] Shin, D., Shin, D. & Shin, D. Development of emotion recognition interface using complex EEG/ECG bio-signal for interactive contents. Multimed Tools Appl 76, 11449–11470 (2017). https://doi.org/10.1007/s11042-016-4203-7.

[23] Kim, K.H., Bang, S.W. & Kim, S.R. Emotion recognition system using short-term monitoring of physiological signals. Med. Biol. Eng. Comput. 42, 419–427 (2004). https://doi.org/10.1007/BF02344719.

[24] Jaiswal, S., Virmani, S., Sethi, V. et al. An intelligent recommendation system using gaze and emotion detection. Multimed Tools Appl 78, 14231–14250 (2019). https://doi.org/10.1007/s11042-018-6755-1.

[25] J. Zhang, Z. Yin, P. Chen, S. Nichele, “Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review,” Information Fusion, vol. 59, 2020, pp. 103-126, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2020.01.011.

[26] Rob Mulla. Audio Data Processing in Python. (Feb. 24, 2022). Accessed: Feb. 28, 2023. [Online Video]. Available: https://www.youtube.com/watch?v=ZqpSb5p1xQo

[27] Ranya. “audio_analytics_speech_emotion_recognition.” Kaggle. https://www.kaggle.com/code/ranyajumah/audio-analytics-speech-emotion-recognition (Accessed: Feb. 28, 2023).

[28] Shivam Burnwal. “Speech Emotion Recognition.” Kaggle. https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition (Accessed: Feb. 28, 2023).