Imperial College London

PROJECT REPORT

DEPARTMENT OF COMPUTING

IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

Cross-corpus Speech and Textual Emotion Learning for Psychotherapy

Author:
James Tavernor

Supervisor:
Abbas Edalat

June 17, 2020

Submitted in partial fulfillment of the requirements for the MEng Mathematics and Computing Degree of Imperial College London

Abstract

Virtual Reality therapy for treatment of anxiety and depression is a scalable and effective therapy. Remote mental health assessments are becoming a necessity due to the current COVID-19 pandemic. These therapies and remote assessments would benefit from accurate, live (real-time), automated emotion recognition.

Existing emotion recognition neural network models have limited application to the subset of emotions they are trained on. The best models available require data pre-processing and cannot operate on live inputs or have limited performance capability. This project aims to create a model which addresses those issues.

In this project we discuss a live speaker-independent emotion recognition model that can be applied to any subset of the emotions it was trained on. We also introduce contextual memory of the speaker's emotional state across numerous utterances while still supporting the live evaluation of utterances. Through this, we achieve state-of-the-art results on prediction of the emotions angry, happy, sad, and neutral and improve on other models that don't require data pre-processing.

Acknowledgments

First, I would like to thank my supervisor, Prof. Abbas Edalat, for his consistent support throughout this project and exceptional guidance. I also want to specially thank Georgios Rizos for his technical expertise in machine learning, terrific feedback, and valuable ideas.

I would like to acknowledge my parents, Consultant Psychiatrists Dr. Rosalyn Tavernor and Dr. Simon Tavernor, for their contributions to interview preparation, data collection, and annotation of emotional labels. Not to mention their constant and overwhelming support during my project and throughout my life.

Thank you as well to all the volunteers who took part in the data collection.

Finally, I would like to thank all of my other friends and family, all over the world, for all they do for me.

Contents

Chapter 1

Introduction

Depression and anxiety are very common mental health disorders with an estimated global prevalence of 4.4% and 3.6% respectively [1]. These percentages only take into account major depressive episodes and major anxiety disorders, and do not consider mild or moderate rates which will be even higher. Face to face therapy can have long waiting times, or people can often feel too ashamed, due to the stigma surrounding mental health, to seek out help.

The COVID-19 pandemic has brought the idea of remote mental health assessment and automated therapy into the spotlight, if not necessitated it. Along with this, it is argued that both the rates and severity of anxiety and depression have increased among the global population due to COVID-19 and the resulting lockdowns [2].

The utilisation of computers to assist in the administering of therapy is therefore essential. The use of Virtual Reality (VR) in therapy has the potential to improve the accessibility of therapy. With only a mobile phone and some hardware, VR therapy is possible without a therapist or visit to the doctors. However, a lot of existing VR therapies require a therapist to administer the therapy itself, and the VR is merely a tool used by the therapist. An Artificial Intelligence (AI) therapist would, therefore, be useful to make therapy entirely accessible for everyone with no waiting times. However, recognising emotions is often useful for effective face to face therapy. As it stands existing VR therapies do not utilise automatic emotion recognition or AI therapists.

In this project, we work with the goal of emotion recognition within an existing VR application in mind. This application will be developed to support the treatment of mental health disorders through Self-Attachment Therapy [3]. An essential aspect of this VR application, and of the therapy itself, is the user's emotions and projecting these emotions onto a self-child avatar.

The next steps towards making therapy fully automated, and therefore more accessible, is to build an AI therapist which can ask questions that elicit specific emotions from the user. Then the AI must continuously recognise the emotions in order to

modify the VR therapy experience. I focus my work on continuous and accurate emotion recognition that is compatible with a range of tools such as AI therapists, VR therapy, and remote mental health consultations.

Chapter 2

Background

2.1 Virtual Reality Therapy

Virtual Reality (VR) is a powerful tool which can be used to improve the quality of therapies offered to patients. An intuitive example of VR for therapy is exposure therapy. Patients can be exposed to their phobias in a gradual and controlled setting with a therapist able to monitor and potentially control the virtual environment as appropriate.

2.1.1 Attachment Theory and Self-Attachment Therapy

A child, whilst growing up, develops attachments with their primary caregiver. There are four types of attachment; [4; 5].

1. Secure Attachment
2. Ambivalent Attachment
3. Avoidant Attachment
4. Disorganised Attachment

Insecure attachments such as ambivalent, avoidant, and disorganised attachment can lead to a vulnerability to mental health conditions later in life, such as anxiety and depression. [4].

Self-Attachment therapy is, therefore, based on attachment theory. Attachment theory theorises a person has an adult self and an inner child, which through therapy sessions, can bond and form a secure attachment to allow for better emotional self-regulation [4; 5]. The result of being able to regulate emotions better should lead to improved mental health.

A currently developed VR environment supports self-attachment therapy [3; 6]. At the moment, the VR application presents the user with information regarding self-

attachment therapy. The user then has a virtual avatar which is customised to resemble the child-self. When the user then comforts the child and virtually embraces the child avatar, the application improves the child-self's mood from sad to neutral and then happy [6]. The addition of emotion recognition and an AI therapist will allow the child-self to accurately represent the feelings of the user and project their feelings onto the avatar when they enter the VR environment. As the user comforts the child-self, extracted emotion could then change the child-self avatar's emotion as the user's emotions improve.

2.1.2 Existing VR Therapies

VR therapy has proven to be successful in the treatment of Post-Traumatic Stress Disorder (PTSD) in soldiers by exposing them to phobias, for example, explosions or helicopter fly overs [7; 8]. By using VR, it is possible to expose patients to phobias which a therapist realistically cannot expose them to in real-life.

More interestingly VR can be used to implement self-counselling [9]. A patient can talk about their problems, and can then switch to a counsellor body where they listen to their problems and provide advice to themselves. They have experimented with using either a lookalike body or a Sigmund Freud body as the therapist. They found that mood was improved the most when the therapist was Sigmund Freud, and the patient had the illusion of body ownership of the therapist.

Other research has found that self-compassion towards oneself can be improved using VR [10]. The application recorded the patient providing compassion towards a distressed child. The patient then embodied the child and experienced the recorded compassion themselves. The VR environment was a reconstruction of the real-life experiment room, and this, combined with a first-person perspective and a mirror to reflect the patient's virtual body, most likely strengthened the illusion of body ownership.

A comparison of VR exposure therapy and VR cognitive therapy showed that cognitive therapy could lead to an improvement in patients with persecutory delusions more than just exposure therapy alone [11].

2.1.3 Mental Health Assessment

An ambitious extension to the project following successful emotion recognition could be to implement an AI therapist that utilises the emotion recognition. Existing VR therapy applications use a multitude of methods to assess improvements in mental health. The AI therapist component would intend to probe the user and elicit emotions from them to understand their feelings. However, existing applications often use a collection of scales and questionnaires. Most patients can present neutrally until the therapist starts to talk with the patient about specific experiences and feelings they have. The AI therapist will, therefore, need to continue to assess and delve deeper into subjects when talking to the patient dependent on their responses.

The World Health Organisation Schedules for Clinical Assessment in Neuropsychiatry (WHO-SCAN) [12] provides a series of questions which help with the assessment of many mental health disorders. The WHO-SCAN interview provides a series of questions, which often build on top of previous questions based on symptoms present in the patient that the therapist can notice from previous answers. If the AI therapist can recognise affirmative or negative statements, even just "yes" and "no", then using the relevant parts of the WHO-SCAN the intention would be for the AI therapist to then elicit emotion from the patient.

2.2 Machine Learning

Machine learning (ML) is the process which enables a computer to automatically perform a specific task without being explicitly programmed to do so. The computer "learns" how to perform this specific task, and the learning model is programmed at a higher level.

There are three primary models for machine learning: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning, in particular, concerns the use of labelled data. That is for every input we have the corresponding expected output, or classification, of that input. In this project, supervised learning will be used with labelled data to train an ML algorithm to classify the emotion currently present in the user of a Virtual Reality (VR) therapy application.

Related work in this area uses a range of different models and methods to approach a solution, and the next few sections will outline these underlying networks and methods to enable us to better approach the existing literature.

2.2.1 Artificial Neural Network (ANN)

The idea behind ANNs is to imitate the networks of interconnected neurons found in a human brain. A graph, therefore, represents the ANN with each node representing a neuron. Each neuron will take a fixed number of inputs which are each multiplied by a weight associated with that input and then summed together. An activation function is then applied to the resulting value to determine how activated the neuron is and thus decide the signal this neuron should output.

The neurons in the network create layers. If one compares this network to a human brain, then related neurons should activate each other. With this in mind, each layer connects only to the layer immediately before and after.

The network should have an input layer, an output layer, and any number of hidden layers in between. A network made using only a combination of these linear layers and activation functions is called a Multilayer Perceptron (MLP).

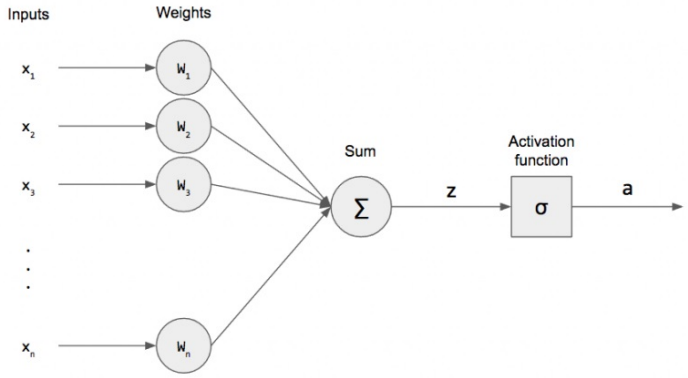


Figure 2.1: An example of a single neuron in an ANN [13]

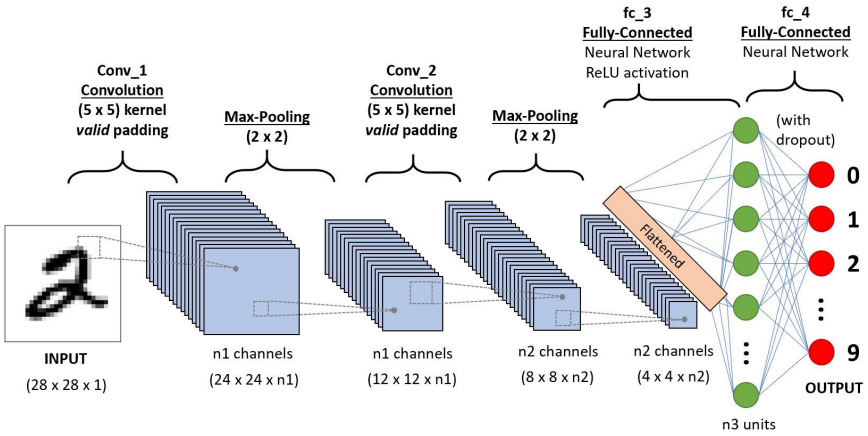


Figure 2.2: An overview of a full CNN for digit recognition [14]

2.2.2 Convolutional Neural Network (CNN)

A CNN is an adaptation of an ANN and is commonly used for learning from images. The predominant adaptation is the addition of convolutional layers and pooling layers.

The convolutional layer uses matrices which act as filters. They move across the height and width of the two-dimensional input computing the dot product of the filter and the underlying inputs. Each filter then introduces an extra dimension (depth) after passing over the two-dimensional inputs. The ML algorithm then learns the values in each filter. The filters then learn to activate upon passing over some feature.

The pooling layer reduces the size of the two-dimensional inputs while maintaining the depth. The idea behind pooling is to split the two-dimensional inputs into a grid of regions and then extract features identified by the convolutional layer, which are in this region. An example of a pooling layer is max pooling, which extracts the maximum value in each region.

Standard linear layers used in MLPs are called fully connected layers in a CNN.

2.2.3 Recurrent Neural Networks (RNN)

An RNN uses a memory of previous inputs to influence the output of a given input. If a sequence of inputs is likely to be related, then the memory of previous inputs in the sequence will be useful for accurate prediction. Facial expressions, body language, and voice are continuous inputs. If these modalities were segmented into consecutive slices and used as inputs, then the context of this sequence will likely be useful for the prediction of emotion, which may explain why we will see the use of RNNs in some of the previous work.

In essence, given a series of inputs, an RNN will work as a standard ANN on each input consecutively. However, it also keeps track of a context, and each input updates the context used in calculating the output for following inputs in the sequence.

2.2.4 Deep Neural Networks (DNN)

A DNN is a type of ANN which contains many hidden layers. The DNN may be a standard neural network or a more complex model such as a CNN. Since the layers in neural networks extract vital features that relate the input to the output, by having many layers, it causes the network to be more successful at finding more sophisticated features. Intuitively this is because each of the layers can learn partial features which create the final output of the network.

2.2.5 Generative Adversarial Network (GAN)

A GAN uses two neural networks in an adversarial competition against one another. The networks consist of a discriminative network and a generative network. The discriminative network attempts to determine whether or not the generative network generated the input, or if it is a real data sample. The generative network trains to increase the error of the discriminative network resulting in the generation of inputs which are more likely to be mistaken as real.

The discriminative network is trained on real samples until it has good accuracy. The generative network then trains by treating inputs which fool the discriminative network as a success and other inputs as failed.

2.2.6 Domain Adaptation

An area of difficulty in ML can arise due to significant differences between the data available for training the model and the data that will be available at the time of testing. When the time comes to predict using the model, the performance will be impaired due to the difference in data. We include discussion of this because a potential risk is encountering an issue of domain mismatch between the actor conversation dataset, and real interview dataset.

If the target data we want to predict and the source data we trained on are related, then we can apply domain adaptation to help tackle the issue. In this project, we

may find that training on high-quality data from a select group of individuals differs too much from real data which the VR app records from the user.

Adversarial Discriminative Domain Adaptation (ADDA) [15] showed promising work in using labelled photographs of numbers, for example from houses, as a source domain and adapting the domain to work with unlabelled handwritten numbers.

The way this was done was by training a CNN and classifier on the photographed digits. As discussed earlier, the CNN should have learned the useful features to extract from the photographs. They then implemented a GAN, where the extracted features from the trained CNN were used as the "real" data values to train the discriminator. The generative part of the GAN is a CNN which was applied to the handwritten digit images. In this way, the generative CNN should have learned to extract the same features from handwritten digits as the features extracted by the CNN trained on photographed digits. Since both CNNs extracted the same features from different domains, the classifier trained on the photographed digits will work in conjunction with the CNN trained on handwritten digits without training a classifier specifically for the handwritten digits. This method is beneficial because the handwritten digits are unlabelled and training a classifier would be inconvenient without any labels to identify the true value of the digits.

If we find that the existing labelled data that we train our emotion recognition on fails to generalise well to real-world data, then it would be worth investigating this approach. Failure to generalise could be due to the difference in the domain between the existing high fidelity data and the potentially noisy data obtained from real-world observation by the VR application. The experimental results found in this paper show a promising improvement of the accuracy by using ADDA over just the source domain's classifier.

Alternatively, domain adaptation can also be used to change the label space [16]. If the desired emotional states for the child avatar in the VR application differs from the labels available in existing training data, this would be a useful approach to take to improve the quality of these existing datasets. By using transfer learning and a multi-layer discriminator, this study shows that it is possible to modify the label space in classification and possibly use the trained network to perform a different task altogether.

2.2.7 Natural Language Processing (NLP)

NLP, at its core, is the study of how a computer can understand human language. With spoken language, an idea can be expressed through countless phrases, and minute inflections can change the meaning of a sentence entirely. NLP aims to study how a computer can navigate these intricate details and understand language.

Often, a pre-trained ML model is taken for NLP and applied to speech to produce numerical data. This data can then be incorporated into a larger network and fine-tuned to be specific to a particular task. The choice of words or the way that words are spoken and connected could be a reliable indicator of emotion. For example,

particular inflections on an otherwise neutral statement could be evidence of irritation.

2.2.8 Attention

Attention is a technique used in neural networks to reduce and remove sequences of values. As part of the network, an understanding is learnt about how to decide which parts of a sequence are important. The model can then 'pay attention' to those parts of the sequence.

The idea behind attention is to create an alignment score, which scores the relationship between the features and final output of the model. Applying softmax to these alignment scores creates a weight which can be used as the attention. For example, these could be used in a weighted-sum manner to the features. This alignment score, for example, could be a learned parameter in the form of a single fully connected layer with softmax activation.

2.2.9 Transformer

The core component of the Transformer is Multi-Head Attention. Transformers use several attention layers in parallel. The layer outputs are concatenated and passed through a linear layer in order to get the output of the Multi-Head Attention. They also use Scaled Dot-Product Attention for the attention layers [17], along with interpreting attention as a mapping of three vectors, which represent the query and a key-value pair to an output. The Scaled Dot-Product Attention is then the matrix multiplication of the query and key vectors. Softmax is applied to the (optionally masked) scaled output of this multiplication, and the resulting vector is matrix multiplied with the value vector [17]. Relating this to the explanation above the query and key are aligned. This alignment is used to pay attention to the value.

The model works in an encoder-decoder fashion with these Multi-Head Attention layers used in between. The values for the queries come from the decoder network, while keys and values are from the encoder [17]. Additionally, both the encoder and decoder models use self-attention, whereas the queries, keys, and values, all come from the previous layer in the model [17].

As discussed in the original paper, the Transformer can be trained faster than traditional recurrent/convolutional layers [17].

2.2.10 Bidirectional Encoder Representations from Transformers (BERT)

BERT is a language representation model that creates a numerical vector representing useful features from some input text. This vector can then be used like normal numerical input data to any type of neural network to tackle NLP tasks [18]. BERT achieves state of the art performance on several NLP tasks by simply fine-tuning the output of the model to fit the NLP task [18].

BERT makes improvements to previous pre-trained language representation models by introducing bidirectional learning allowing the model to look left as well as right down the text sequence [18]. The authors also show that a fine-tuned implementation of BERT can often be even better than task-specific models. Considering this, BERT should work well at emotion prediction with potentially less time needed to construct the model while still achieving excellent performance [18].

The BERT architecture uses blocks of bidirectional Transformers pre-trained on two different language tasks [18].

BERT has previously been fine-tuned to tackle emotion recognition of text with good results, and so it could be an excellent starting point for NLP on transcripts [19].

2.3 Emotion Recognition

2.3.1 Databases

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database

When creating models for emotion recognition, there is a limited amount of useful data available. The IEMOCAP database is specifically designed to deal with this issue.

The database contains well-labelled data on speech, facial expressions, and hand gestures of a particular subject during conversations. The subjects consisted of ten actors in two-party conversations [20]. The actors performed three scripted sessions and two improvised sessions which were designed to express specific emotions. In each session, an actor wore a headband with 53 facial markers and a wristband in order to track their motion during the actions [20].

The discrete labels used in the database represent the general emotion and will be relevant to this project. It may be hard to get this level of quality from face tracking performed without markers on the face. However, it should be possible to downsample the data in IEMOCAP to match the real data. It will also be essential to ensure that there is no overfitting. With only ten subjects in the database, it would potentially be easy to incorrectly fit the model to the mannerisms of the specific subjects.

Database of Elicited Mood in Speech (DEMoS)

DEMoS is an extensive labelled audio database in Italian from a large sample of 68 speakers. Similar to IEMOCAP, it contains relevant discrete categories [21]. The database uses Mood Induction Procedures to capture real emotions as opposed to using acted speech [21].

The large number of speakers in this database could be useful to prevent overfitting on IEMOCAP. Additionally, the Italian language could help identify overfitting. The

network should extract emotion from the way someone speaks, not solely from the language used, so a network trained on multiple languages has the potential to generalise better.

Some emotions are more easily detected through speech, while others are conveyed through facial expressions and hand gestures. This concept could be a limitation of DEMoS which only contains audio. This limitation could be overcome by training multiple networks, one to interpret the visual communication channel, and another for audio. However, the visual and audio communication channels work together to convey emotion; therefore, a network that interprets them together is ideal. As discussed earlier, it may be possible to adapt a network trained on the DEMoS audio and transfer it to work on combined audio and visual input with domain adaptation.

Extended Distress Analysis Interview Corpus (E-DAIC)

The E-DAIC was released as part of the AVEC 2019 challenge, and it is a subset of the DAIC dataset [22; 23]. The DAIC contains a series of interviews to assess both the severity of PTSD and depression of the interviewee. The type of interviews conducted include a portion which is conducted by an automated speaker, both acting autonomously and remotely controlled from another room [22]. This dataset does not contain emotional labels on the utterances; however, the dataset does contain real interviews with unscripted dialogue [22]. Additionally, the interviews which were conducted by an automated speaker represent the exact environment in which emotion recognition would be used.

The E-DAIC consists of interviews conducted by the automated interviewer, and they provide the Personal Health Questionnaire (PHQ) 8, PHQ 9, and PTSD severity scores [23]. The transcripts provided are automatically transcribed using Google Cloud [23], and therefore mimic the data that would be available to an AI interviewer.

2.3.2 Previous Work on Emotion Recognition

Attention-Augmented End-to-End Multi-Task Learning for Emotion Prediction from Speech

Previous studies have been designed to avoid hand-picked features and learn better and more useful features by implementing an End-to-End network along with multi-task learning to improve issues with overfitting [24]. This is achieved by using the raw waveform of speech in the IEMOCAP database. A CNN operates on the raw waveform with that idea that it will perform feature extraction. This CNN then feeds the extracted features into an RNN to perform sequence modelling. The network then uses an attention layer to evaluate which parts of the sequence contribute the most to the prediction. Finally, the network outputs the predictions of arousal, valence, and dominance. The reasoning for predicting three properties with one network is to prevent overfitting. The results of this experiment demonstrate that this

multi-task learning approach does increase the performance of the network.

The network is not directly applicable to the problem as they only use audio as inputs, and they do not predict the general emotion, but instead related properties of the speech. The paper also shows that the use of attention layers and multi-task learning can improve the generalisation of the network.

End-to-End Multimodal Emotion Recognition using Deep Neural Networks

An applicable approach to the problem at hand uses both visual and audio inputs to a network to predict emotion [25]. The network is split into three distinct parts, the visual network, the speech network, and a two-layer RNN which consumes the output of the visual and speech networks. The visual network is a DNN which takes the raw video cropped down to the face as input. The speech network works on the raw waveform of speech, In order to prevent overfitting, they use dropout with a 50% probability. The output of the visual and speech network have their output features concatenated and fed into a two-layer RNN.

This network provides an excellent example of prediction using both visuals and audio together, although they predict arousal and valence instead of general emotion. Since only raw pixels of the face are used, it might be possible to improve on this network by performing an estimation of the positions of the facial features relative to one another and using this as inputs.

Multimodal Emotion recognition on IEMOCAP Dataset using Deep Learning

Speech, text, and motion capture data for the hands and face can be used to perform classification of emotion on the IEMOCAP dataset [26]. The approach taken is to train a network for each of the inputs and then merge their features with concatenation and feed it forward into a standard dense layer with a Rectified Linear Unit (ReLU) activation function and then an output layer to combine them.

We may want to expand on the label space used here as they only feature four emotions. The inputs to the motion capture network may be too precise for it to function well without modification. The video feed of the user of the VR application is unlikely to be able to accurately track the motion of parts of the face and hands as precisely as IEMOCAP. They use Global Vectors for Word Representation (GLoVe) on the text transcript, this converts words to vectors which enables an ML algorithm to learn features such as word similarity. The use of GLoVe seems like a good idea as we will be doing speech analysis, and adding some kind of speech to text recognition to the software should not add much complexity. It would also enable the use of Natural Language Processing (NLP) at a later stage.

In their results, they also managed to obtain relatively high accuracy on emotion recognition, predicting anger, excitement, neutral and sadness; they manage to score 71.04% accuracy. This accuracy may decrease due to the addition of more emotions and the constraints that will need to be implemented to the motion capture inputs.

Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions

Similarly to previous discussed work, they only predict four emotions; in this case, anger, happy, sad, and neutral. The results may not extrapolate well when predicting a larger subset of the emotions available in IEMOCAP [27].

Several different combinations of feature inputs, models, and combinations of feature inputs are experimented with, providing helpful insight into which features will be especially useful for emotion recognition [27]. The paper discusses that the context of a sentence is significant for making the correct prediction of the emotion of utterances. One example they provide is that a word such as "good", may seem positive, but surrounding context could change that positive connotation to a negative one [27].

In their first model, utterances are embedded to vectors, and a CNN is used to try and learn the surrounding context of words in the utterances. In the second and third models, they use a CNN with max pooling on the audio spectrograms and Mel Frequency Cepstral Coefficients (MFCCs) respectively. An additional model 2B was trained which experimented with additional parallel convolution layers. The following models are then fusions of models 1 to 3. Fusion is done by concatenating the output layers from each model before passing it through a fully connected layer. These fused models are denoted as 4A - fused spectrogram and MFCC models, 4B - fused spectrogram and text models, 4C - fused MFCC and text models.

Performance is then calculated on the 4 emotions. The reported overall accuracies are

- Model 1 - 64.4%
- Model 2A - 71.2%
- Model 2B - 71.3%
- Model 3 - 71.6%
- Model 4A - 73.6%
- Model 4B - 75.1%
- Model 4C - 76.1%

This information could suggest that on single modalities, using extracted features from audio as the input is going to have a better performance than text. However, a combination of both is even better.

Multimodal Speech Emotion Recognition Using Audio and Text

In line with other work, we again consider the use of RNNs on both text and audio features from the IEMOCAP dataset.

Initially, models are produced on the single modalities of audio and text. For audio, the use of MFCC as the input to the RNN is implemented, which is then combined with prosodic features using concatenation to make a prediction. Similarly, for text,

words are embedded as vectors and are encoded using an RNN just as they did with the MFCC features [28].

More interesting is the multimodal aspect of the paper. They perform standard fusion using concatenation on the output of these models for the first model, but in the second, they use attention to do this fusion. For the fusion part, they take the concatenation of the prosodic features and MFCC RNN outputs, along with the sequence of hidden states from the text RNN, and multiply them to get the attention weights that are applied to the hidden states of the RNN. This attention on the text hidden states is then concatenated with the output of the audio model (same as the single modality model) to make a final prediction [28].

They assessed the models using the weighted average precision on angry, happy (merged with excitement), sad, and neutral. They found the normal concatenation (0.718 WAP) to outperform the attention-based fusion (0.690 WAP) when considering the Weighted Average Precision (WAP). They, however, suspect it is due to the lack of data the more complex model parameters cannot be learned as easily [28]. It could be because of the way they have chosen to fuse the parameters. They are certainly learning to pay attention to the most relevant parts of the textual RNN, but they still use the audio vector as usual.

Multimodal Speech Emotion Recognition and Ambiguity Resolution

Consideration has been taken into how handcrafted features in combination with a series of different non-neural network ML techniques compare with end-to-end deep learning implementations. The assessment of the different models is based on IEMOCAP utilising upsampling techniques to balance the dataset predicting six emotions:- anger, happy, sad, fear, surprise, and neutral. They find that non-neural network models perform similarly to deep learning models depending on the feature inputs they are given [29].

Additionally, they can conclude the most important features from the audio, text, and video features they extracted and rank them. For a model which learns from pre-defined features, this could be incredibly useful to reference which features will likely give the best results. They also found that fusing audio and text modalities saw a good improvement in the scores [29].

Multimodal Sentiment Analysis using Hierarchical Fusion with Context Modeling

By investigating methods of fusion, there is evidence that a hierarchical type of fusion outperforms plain concatenation of the different modalities [30]. Their approach first fuses two modalities by first ensuring both modalities features are the same dimension size, and then applying a fully connected layer mapping each pair of features to one output feature [30]. For each time step, they take the corresponding feature from each modality and learn weights that map it to one fused feature.

After fusion of each combination of the audio, video, and text features, these fused

features, at a later stage in the model, are then fused once again with the other fused vectors, in this respect the fusion is hierarchical. By using the emotions of anger, happy, sad, and neutral for evaluation, their results show that modalities involving text performed significantly better, and that fusion of all three modalities did not perform significantly better than the use of only audio and text fusion [30]. The best model was the audio, video, and text model with an accuracy of 76.5%, closely followed by the audio and text model (76.1%) [30].

DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation

DialogueGCN produces state of the art results on the prediction of the six emotions, happy, sad, neutral, angry, excited, and frustrated with a weighted average accuracy of 65.25% using only the single modality of text [31].

The key idea behind this work is that during a conversation, the emotions of a person is dependent on themselves. However, they also depend on the other parties involved in the conversation and how they influence the other parties' emotions [31]. This may not be entirely fitting for the model we aim to achieve, as an AI therapist may not model well with its influences on the other speaker, though the concepts may be useful for self-dependency.

As with other models, they encode the text transcripts to word vectors and extract text features of the utterances using a CNN. These utterance level features are then used throughout the rest of the model [31]. Since conversations are sequential, they attempt to model this with a Sequential Context Encoder, which is a bidirectional RNN [31]. They then try to model the speaker contexts, how a speaker influences their own emotions, and other parties' emotions, using a directed graph trying to capture the relationship between participants and their sequentially encoded utterances [31]. Graph convolution is then used to make their predictions. The two main issues with this implementation in our use case are that preprocessing of the audio is needed to construct the graph and that the theory behind it relies on the understanding of the conversation, which may not be true in a therapeutic setting, or when talking with an AI therapist.

2.3.3 Previous Work on Prediction of Depression

Multi-level Attention Network using Text, Audio and Video for Depression Prediction

This paper was the winner of the AVEC 2019 depression detection challenge. This model works on the three modalities of audio, video, and text. They use a selection of extracted features from the audio, such as MFCC, for the audio modality. For text, they use word embeddings with zero paddings for shorter utterances, so they are all of the same lengths. Since the AVEC dataset does not include the raw video of interview, but instead provides extracted features from the video, they use these features for the visual modality [32].

Similarly to emotion recognition models, they implement the use of bidirectional RNNs, specifically Long Short-Term Memory units (LSTMs), to learn from each modality. Their method of fusion is by applying attention at different points throughout the network, both to learn which parts of audio and video features are important, but also in the fusion of the three modalities [32].

Their final result on the prediction of PHQ 8 scores is a Root Mean Squared Error (RMSE) of 4.28 [32]. This is state of the art for PHQ 8 regression. It is interesting to note that they looked at combinations of only two modalities and found that audio and text performed better than video and text, with the full three modality fusion only slightly improving on the audio and text fusion [32].

A Multimodal Hierarchical Recurrent Neural Network for Depression Detection

This was the runner up to the AVEC 2019 challenge, and it is worth considering how they achieved their impressive results. Once again the use of bidirectional RNNs is apparent, with a hierarchical format to the RNN layers. The first hierarchy is the feature extraction performed on the audio and video features, using each second as a timestep [33]. In order to use 1 second as the time step, they average each of the extracted audio and video features, provided as part of the challenge, over 1 second. The audio features are reduced in dimension to be only of size 50 [33]. They use an encoder-decoder style network to reduce the dimensionality to keep the utility of the data while removing redundancy contained within it [33].

When it comes to the text processing, they use a mean-max autoencoder which takes advantage of multi-head self attention [33], similarly to the Transformer [17]. Then they use a lexicon to look up the emotional connotation and the strength of this connotation of words in the transcript. This word emotion vector could indicate an emotion recognition tool could improve quality of depression detection and may be inspired by the fact that word choice could be influenced by a person's underlying emotions [33].

2.3.4 Related Work on Audio and Text Tasks

Attention-based Atrous Convolutional Neural Networks: Visualisation and Understanding Perspectives of Acoustic Scenes

This paper uses MFCCs as the input instead of the raw waveform [34].

This paper intends to classify the acoustic setting that audio was recorded in, and visualise the feature maps, so it is not entirely relevant to emotion recognition from speech. The idea of using attention layers that occurred in the previous paper appears again. More importantly, we are introduced to the idea of atrous CNNs [34]. An atrous CNN works by dilating the filter by a particular rate. This rate introduces a gap between the pixels considered by the filter of the CNN. The use of atrous CNNs is done to remove the pooling layers which reduce the size of the feature maps which

would be detrimental for visualisation of the features. The atrous CNNs had a higher accuracy than standard CNNs.

Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification

The goal of classification of hate speech is a hard, and somewhat subjective task. In this study, the accuracy of classification is improved on tweets by augmenting the texts without changing the meaning [35]. One augmentation is to convert the words to vectors and use the cosine similarity to switch words with sufficiently similar words. Additionally, given sentences are of differing length, inputs to the neural network can end up with much zero-padding, which can distort a network.

This approach could potentially be applied to the text transcript to determine emotion. Similar to hate speech, it can be challenging to understand emotion through text. However, by following this approach, we should be able to achieve a reasonable understanding of emotion from the text. It could be possible to combine this with the approaches above to produce an even better method of emotion recognition.

Chapter 3

Databases and Data Collection

3.1 IEMOCAP

The IEMOCAP data is the best available dataset for emotion learning. The data contains raw audio, raw video, text transcripts, and motion capture information, along with rigorously labelled emotions on each utterance [20]. Additionally, other emotional information of speech is labelled, which would allow the model to also learn about the arousal, valence, and dominance of speech [20].

3.1.1 Data Pruning

The dataset contains ten emotion categories; these are labelled as neutral, happy, sad, angry, surprised, fearful, disgust, frustration, excitement, and other [20]. Since there is not always full agreement between the human evaluators, only the utterances where there is majority evaluator agreement is considered, the resulting data distribution can be seen in Table 3.1.

Emotion	Count
Angry	1103
Happy	595
Sad	1084
Surprised	107
Neutral	1708
Frustrated	1849
Excited	1041
Fearful	40
Disgust	2
Other	3

Table 3.1

We can see from the table that the dataset is relatively imbalanced. Since disgust has

such a small number of examples, and other is an undefined emotion, I have chosen to remove these from the dataset, and thus train the model on the eight remaining emotions.

Fearful is an important emotion for therapeutic settings, but it is still very uncommon with only 40 examples in this pruned dataset. Upsampling might be a useful method to combat the still imbalanced dataset (see Figure 3.1 for histogram) however I have opted to use weighting during calculation of the loss function while training instead. This is simple in PyTorch as a weight for each class can be provided when constructing the loss function. Each class during training is given a weight of $\frac{\text{size of class}}{\text{size of largest class}}$.

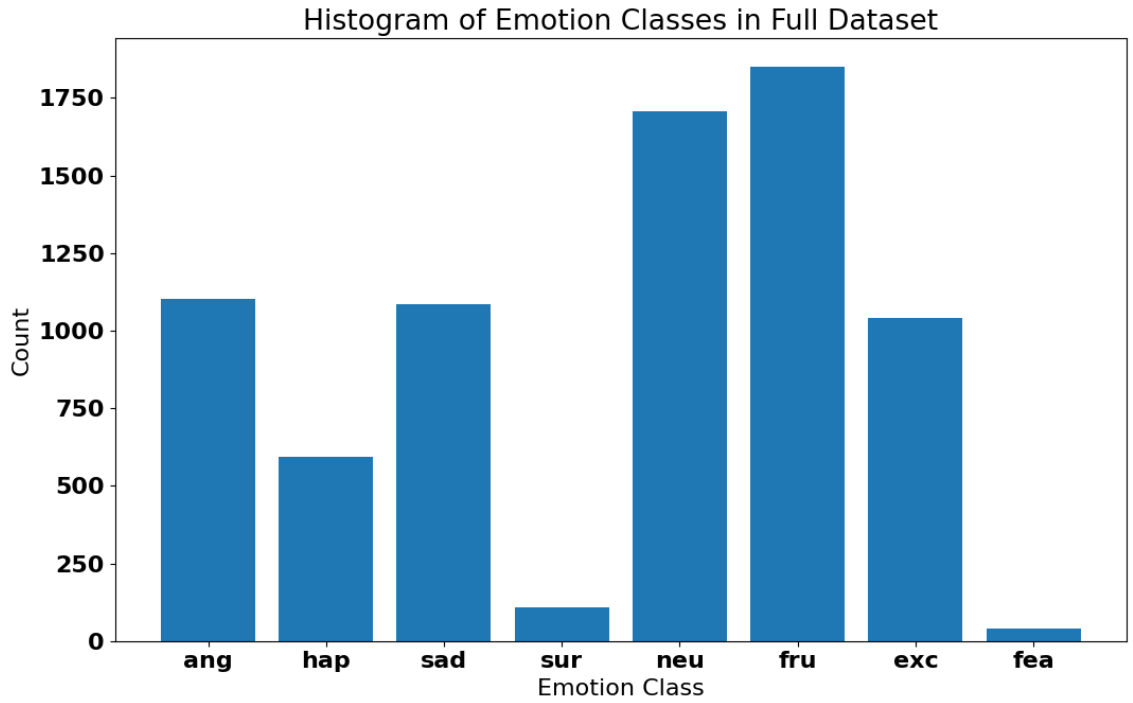
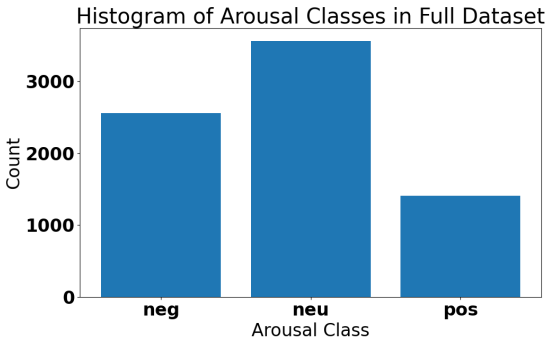


Figure 3.1: Balance of emotion in the pruned full IEMOCAP dataset.
The labels in the diagram represent ang = Angry, hap = Happy, sad = Sad, sur = Surprised, neu = Neutral, fru = Frustrated, exc = Excited, fea = Fearful.

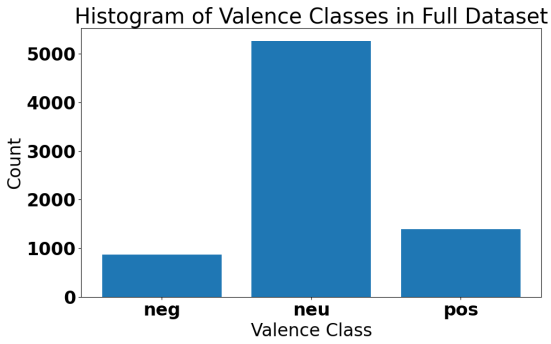
Similar imbalances are found in the arousal, valence, and dominance classes, as shown in Figure 3.2. Similarly, I have decided to weight these during training.

3.1.2 Training-Test Split

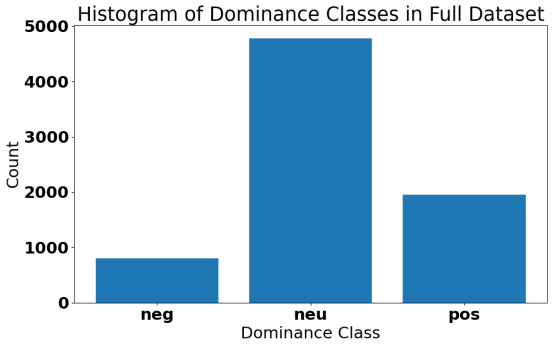
The IEMOCAP dataset does not define a pre-determined train-test split, so this must be established. The model must not overfit to the speakers it trains on. As such, the test training-test split is important. Each of the sessions in IEMOCAP consists of two speakers, with the full dataset containing a total of ten speakers [20]. With this in mind, the dataset is split, using the first four sessions as the training data and session



(a) Arousal



(b) Valence



(c) Dominance

Figure 3.2: Balance of arousal, valence, dominance classes in the pruned full IEMOCAP dataset.
The labels in the diagram represent neg = Negative, neu = Neutral, pos = Positive.

five as a test set. This split ensures that the results on the test set represent unseen speakers, and it has been done before to ensure speaker-agnostic results [26].

The resulting train and test splits follow similar distributions of the emotion labels, demonstrated in the Table 3.2 and Figure 3.3. The test set should, therefore, produce results which are a good representation of the dataset. Additionally, the results should be speaker-agnostic and represent the performance of how the model generalises to new speakers.

Emotion	Train Set Count	Percentage	Emotion	Test Set Count	Percentage
Angry	933	15.9%	Angry	170	10.3%
Happy	452	7.7%	Happy	143	8.7%
Sad	839	14.3%	Sad	245	14.8%
Surprised	89	1.5%	Surprised	18	1.1%
Neutral	1324	22.5%	Neutral	384	23.3%
Frustrated	1468	25.0%	Frustrated	381	23.1%
Excited	742	12.6%	Excited	299	18.1%
Fearful	30	0.5%	Fearful	10	0.6%

Table 3.2: Distribution of Emotion Classes in the Training Set and Test Set.

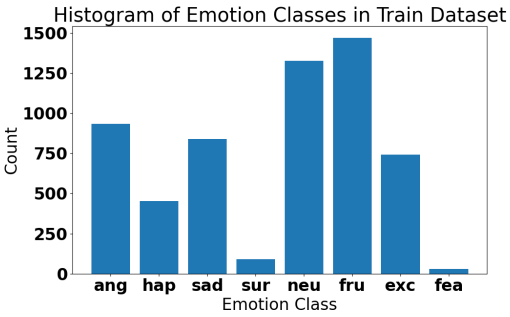
No cross-validation split was performed. However, in future work, to validate the results of the model, it would be possible to do cross-validation split across the five sessions using 80% for training and 20% for test. Taking the average across all of these would show that the model is reliable and reproducible.

3.2 AVEC 2019 E-DAIC

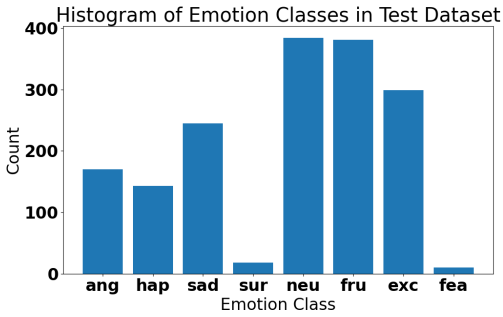
There is concern that the dataset will overfit to actors in some way. IEMOCAP consists of both improvised scenes and scripted scenes; however, in both cases, the participants are still acting [20]. There is a concern that the acted emotions are not the same, and may manifest differently, to real emotions. To make sure that this is not the case, we need to take advantage of other datasets for the evaluation of the model.

Initially, we had planned to collect interviews manually by leveraging mood induction [36] similarly to the DEMoS dataset [21]. Then we could verify the performance of the emotion recognition model on individual subjects speaking to a computer. The proposed method of collecting these emotional interviews was the use of film induction and self-statement induction, followed by specific questions to encourage the volunteer to speak naturally. However, based on initial experiments with a couple of volunteers, the results were not promising.

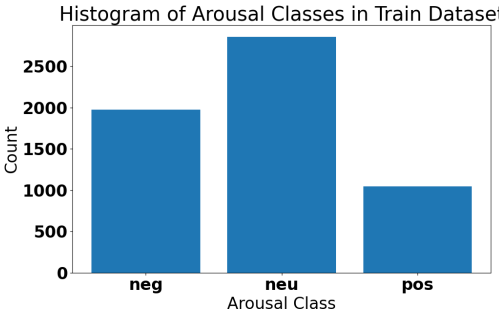
Fortunately, the E-DAIC dataset can be used in place of this dataset. While we cannot evaluate the emotions on this dataset, we can use the emotion recognition predictions to improve a baseline depression detection model. This will show that the



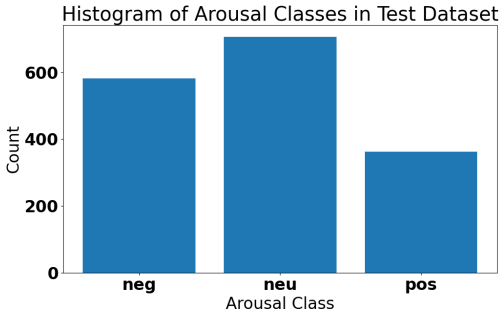
(a) Emotion Train Set Distribution



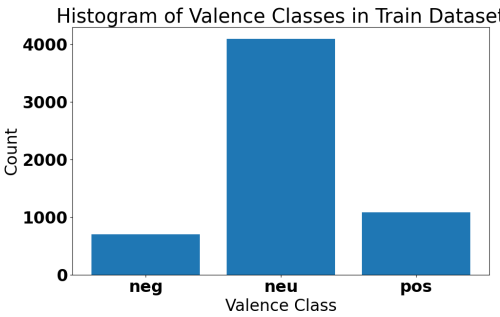
(b) Emotion Test Set Distribution



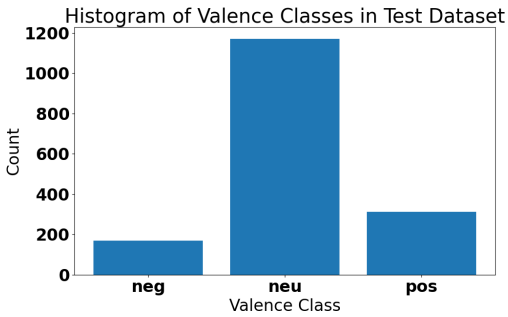
(c) Arousal Train Set Distribution



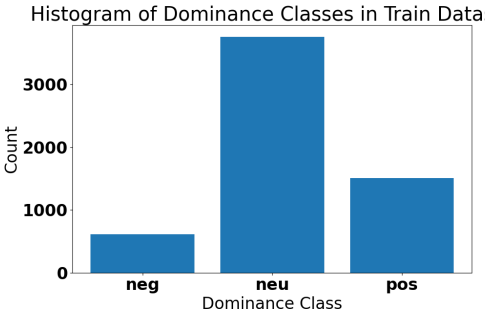
(d) Arousal Test Set Distribution



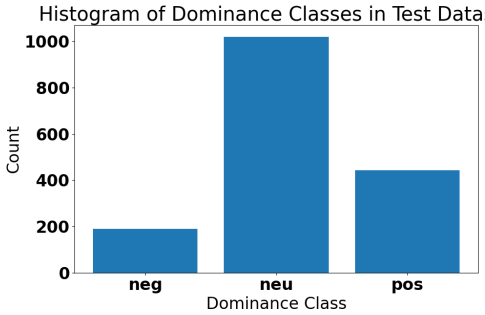
(e) Valence Train Set Distribution



(f) Valence Test Set Distribution



(g) Dominance Train Set Distribution



(h) Dominance Test Set Distribution

Figure 3.3: Balances of training and test sets side by side.

emotional understanding learned by the emotion recognition model is valid and that this understanding can be used to improve related emotional tasks.

3.2.1 Data Pruning

In this case, we have a series of interviews which contain numerical labels used for regression [22]. However some of the transcripts made by Google Cloud are inaccurate, or the timings do not match up correctly. Some timestamps are invalid; to mitigate this, we only consider utterances that are between one and ten seconds long, and with a confidence value of more than 0.85.

3.2.2 Data Usage

Loading of the data is somewhat tricky. We want to make predictions on the full interview, but we also want to utilise per utterance predictions using the emotion recognition model. This issue complicates how the data can be loaded.

The decided solution was to treat a single interview as a sample in a batch. When a batch is loaded, we return ten interviews and the corresponding labels. These interviews that are loaded within a batch contain a PyTorch DataLoader object which provides iteration of each utterance within that interview. By doing this, we can shuffle the interviews and can allow the use of the pre-defined training-test split of the data.

3.3 COVID-19 Interviews

There is limited data available on the study of modelling and prediction of general anxiety disorders by neural networks [37]. As part of the extension of emotion recognition, we look at depression and anxiety prediction with transfer learning from emotion recognition.

As discussed in the introduction, anxiety in the general population has increased during the COVID-19 pandemic. As such, we conduct interviews discussing various aspects of the COVID-19 pandemic to elicit various emotional responses.

3.3.1 Data Collection

31 healthy volunteers with no physical or mental health conditions agreed to participate in interviews through the medium of online teleconferencing software. The interviews consisted of discussions about the volunteer's feelings of sadness, anxiety, and anger regarding the current COVID-19 situation. To finish the interview, a brief discussion occurred about positive experiences and things to look forward to post-lockdown.

The volunteers provided GAD 7 questionnaire answers, and a Likert scale rating for how strongly they felt angry, happy, sad, and anxious, during the interview. In

addition to this, two Consultant Psychiatrists independently rated their opinion of the strength of emotion displayed during the interview using the same Likert scale rating.

3.3.2 Training-Test Split

Since the dataset is small, it can only be used to fine-tune the outputs of another model. With this in mind, we split the data to use five interviews to train, and the remaining 26 interviews for evaluation of performance.

Choice of the training set is made manually, as randomly selecting five interviews may not represent the full spread of values in the dataset well enough. For example if five interviews which each represent no anxiety are chosen, then any fine-tuning will not learn anything about anxiety.

Chapter 4

Emotion Recognition Model

4.1 Model Design

4.1.1 Model Requirements

The main idea behind this project is that traditional therapy is not scalable, and AI that can reliably detect emotion can conduct therapies such as self-attachment therapy. With this in mind, the first clear objective for the emotion recognition model is for the recognition to work live, on a per utterance level of prediction. A model which only predicts utterance level emotions after the conversation has concluded is unable to feed information to an AI therapist while it speaks with a patient. Additionally, this model could be used to aid professionals in conducting remote assessments of patients, but again would need to be delivered in real-time to be clinically useful. Emotional predictions given after the conversation will not help a professional who may not remember exact utterances made by the patient.

For application in therapy, accuracy is important. A core aspect of self-attachment therapy is the projection of emotions onto the child avatar; poor emotion recognition will fail to map the correct emotions onto the avatar. Confusion of core emotions such as happy, sad, and angry, would be a fundamental flaw in the model. However, confusion of emotions such as disgust, or emotions that are very similar, such as angry and frustrated, would be undesirable, but still likely to be useful therapeutically.

The predictions need to be reliable and not speaker-dependent. When training the model, there is a risk that it will overfit to the speakers in the training data and fail to be reliable and valid when applied to new speakers. Additionally, as most available data consists of American speakers, there is a potential risk that the model may overfit to the American accent and dialect. In order to mitigate this, the model will need to model each new speaker's way of speaking somehow and how their emotions fluctuate. The predictions also need to be valid and accurately model emotions. There is a risk that the model could learn intonation as opposed to the underlying emotion.

Unlike other implementations, such as DialogueGCN [31], we do not want to rely on inter-speaker dependency. The way that parties in a conversation influence one another's emotions may be useful for modelling emotion in conversation. However, when we model the emotions between a single person and an AI therapist the dependency may not hold. The model needs to model one speaker without relying on the inputs of other parties within a conversation or any other outside factors.

It is important to consider the use of all the available modalities, such as the audio, video, and contents of the speech, which all factor into how we, as humans, recognise emotions. In order to achieve this, we must investigate the fusion of modalities and how to extract useful features from each modality used.

Finally, when the completed model is produced, the model should be easily applied to predict any subset of the emotions that the model was trained on. For example, as a therapy application is developed, there may only be a desire to predict angry, happy, neutral, and sad. Additionally, this allows us to compare the model's performance with previous attempts at emotion recognition which often report their performance on a different set of emotions.

4.1.2 Design Choices

A critical decision for this model was which modalities to take into account. Available for IEMOCAP, we had raw audio, raw video, transcripts of speech, and motion capture information [20]. Initially, we considered using audio and video primarily, and attempt to use the motion capture information if, and only if, the data could be reliably obtained in an interview session. Since the concept of emotion recognition is to work in a VR application as well, the modalities we train with need to be obtainable during the VR session. Of course, video data could be unrealistic with a VR helmet obscuring the view of the user's face.

Additionally, previous work on emotion recognition seems to suggest that audio and text are significantly more important than video. In papers where different combinations of modalities are considered, audio-text outperforms audio-video [30; 38]. This phenomenon seems to also apply in other, related, paralinguistic tasks such as depression prediction [32]. With all this in mind, we decided to prioritise the use of text/transcripts over the use of video.

In order to predict any subset of the emotions, we decided that the best approach is to train the model across all the available emotions and then predict only a subset. For example, if we only wanted to predict the emotions happy or sad, we would predict as usual and using a wrapper around the model, predict happy or sad based on which of the two had a higher probability out of all emotion predictions made by the model.

Emotions are complex qualia that cannot be reduced into a specific set of criteria to measure. In this way reducing the audio signal to a smaller handcrafted feature, such as MFCC or the audio spectrogram, might lose important low-level information contained within the raw audio. We aim to replicate the human understanding of

emotion. In this way, we have decided to construct the model end to end with as minimal reductions to input data as possible.

4.2 Audio Modality

4.2.1 Initial Model Decision

As part of the model requirement, we want to develop an end-to-end neural network to perform emotion recognition. One of the models used for arousal, valence, and dominance prediction was an end-to-end network operating on only the raw audio and had impressive results [24]. It was adapted for emotion classification [39]. Fortunately, we were able to obtain a version of this modified for emotion recognition model implemented in TensorFlow from Georgios Rizos. However, previous experience with PyTorch led to the decision to reimplement this code into PyTorch, which did require some changes.

For example, in this modified model, Peephole LSTMs are used instead of GRUs, and cell state clipping is performed on these LSTMs. Each of these is not implemented as default in PyTorch. The decision was made to use normal LSTMs since research into the usefulness of Peephole LSTMs indicates the performance difference is insignificant [40; 41]. We then implemented an LSTM stack using PyTorch LSTM cells and clipped the state manually after each state. In the end, however, cell state clipping was found to be relatively insignificant, and we opted to use the LSTM stack provided by PyTorch. The only other modification made by this implementation was the removal of the dropout layer. I also had to change the kernel size of the convolutions to an odd number in order to replicate the "same" padding found in TensorFlow.

Additionally, after investigation, a fairly significant improvement was found in the audio-only model by massively increasing the size of the feature maps in the convolution layers. The feature maps of the convolutions are 256, 512, and 1024, following each convolution with a max-pooling layer with kernel size and stride equal to 10. The bidirectional LSTM on top of this CNN then uses a hidden size of 1024. It then uses attention on the output just as in the original paper. The final CNN and bidirectional LSTM audio model components can be seen in Figure 4.1 and Figure 4.2 respectively. We will refer to these as the Audio CNN Model and the Audio BiLSTM Model from here.

4.2.2 Validation of Model

After constructing the pure audio model, we wanted to confirm that the knowledge learned by the audio model would generalise and work on interviews of real people, and not just the scripted conversations in IEMOCAP. Fortunately, the E-DAIC dataset almost perfectly replicates the hypothesis.

Since the E-DAIC [23] consists of real interviews with patients talking to an AI in-

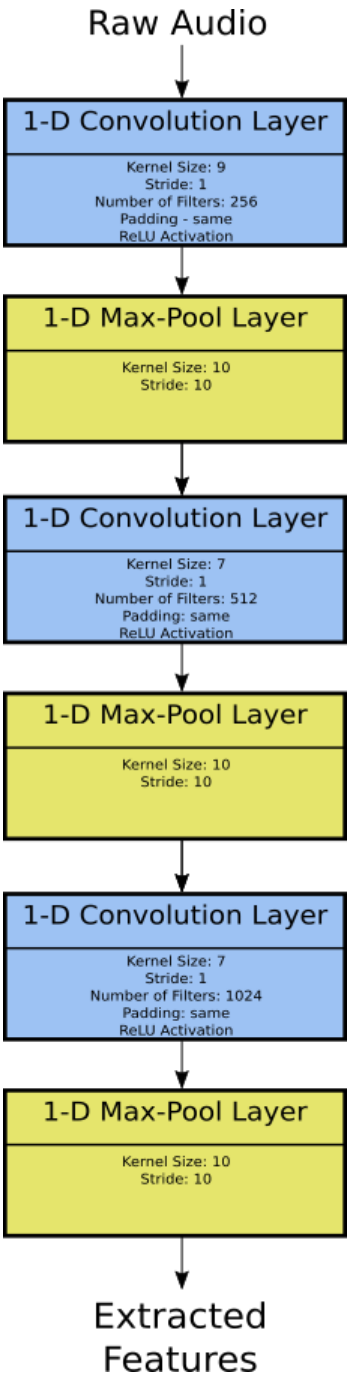


Figure 4.1: The final audio CNN model used.

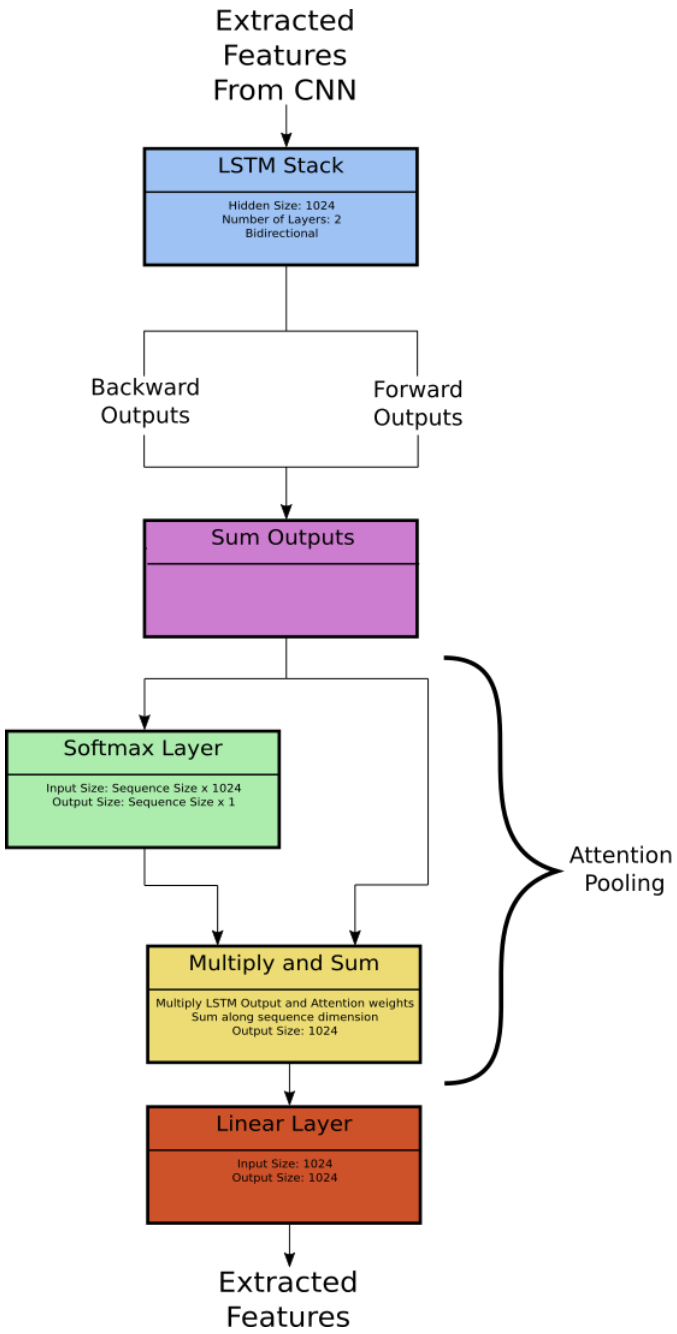


Figure 4.2: The final audio bidirectional LSTM model used.

terviewer, if we can successfully apply the emotion recognition audio model to this, we can conclude that it generalises well enough for this model to be accepted. The labels provided with this dataset, however, do not give details on the utterance level emotions. We decided the most reasonable decision was to apply the emotion recognition for happy, angry, and sad, on each utterance in the dataset, and manually verify that the model has good enough performance on the dataset. To do this, we took the top 30 most likely predictions for each utterance and investigated these. There was undoubtedly some overfitting to a specific speakers voice, but the performance was good enough to get a feeling that issues with overfitting to scripted IEMOCAP speakers were not an issue, and that this model would work in an unscripted interview setting with an AI.

4.2.3 Limitations of Model

Of course, the main issue is uncertainty whether the audio model is overfitting to the speakers in the training data. However, given the manual verification, when combined in the larger model utilising multi-task learning the model should generalise well.

Another issue is caused by the use of raw audio and the large feature map size in the convolution layers. The amount of memory required during training builds up very quickly; thus, a very small batch size with frequent loss calculations was required to free up the memory used for backpropagation. In order to mitigate this, the weights are only updated after enough small batches have been processed, adding up to the true batch size.

It was considered that a multi-modal autoencoder based on text and audio could produce an underlying latent space which could then be used for prediction. However, it became clear that to do this end to end would use far too much GPU memory, and ultimately was not possible. This autoencoder idea has been investigated in the past and produced impressive results [42]. The only way to make this work with the GPU resources available would be to run the model with half-precision. It is likely the performance would be much worse in this case, and it is not justified to follow through with this idea.

4.3 Text Modality

4.3.1 BERT

Choice of BERT

BERT is a relatively new language model developed in 2018, but its performance on a variety of language tasks is very impressive [18]. It has been successfully applied for emotion recognition on text/social media [19] and there is additional evidence that fine-tuning BERT on IEMOCAP will work well [43].

BERT should also be straight forward to implement since it is a pre-trained language model that only requires fine-tuning to the task [18]. The approach would consist of deciding on which BERT architecture to use, preparing it for fine-tuning, and then training as part of the multi-modal model. Additionally, since BERT is bidirectional, it should be able to learn the dependencies in both directions that influence the perceived emotions of a person.

4.3.2 Choice of BERT Architecture

In order to determine which variant of BERT would best suit emotion recognition, we constructed a straightforward model that could quickly be trained. We then chose the best performing of the architectures. The models all consisted of a pre-trained BERT architecture from the Python library Transformers by HuggingFace, followed by a ReLU fully connected layer, and an output layer with no activation function.

To follow is a brief description of each of the variants of BERT architecture tested in this simple model and how well they performed at the emotion recognition task over ten epochs.

BERT-base

BERT-base is the standard model outlined in the original BERT paper [18]. This model has 12 layers of Transformer blocks, consisting of 12 self-attention heads, with a hidden size of 768 resulting in 110 million parameters in the model [18].

BERT-large

BERT-large is the significantly larger model outlined in the original BERT paper [18]. This model has 14 layers of Transformer blocks, consisting of 16 self-attention heads, with a hidden size of 1024 resulting in 340 million parameters in the model [18].

RoBERTa

RoBERTa builds off and aims to improve BERT by analysing and modifying the pre-training methods used, carefully changing training configurations and the pre-training task. They manage to achieve a state of the art performance over BERT on a handful of language tasks [44]. A RoBERTa-base is used in this example, which has the same parameters and architecture as BERT-base, but following the RoBERTa training.

DistilBERT

DistilBERT is a distilled version of BERT, managing to retain almost all of its language understanding while being significantly smaller and faster than BERT [45].

Model	Epoch with Best Accuracy	Accuracy	Macro F1
BERT-base	9	45.9%	0.414
BERT-large	10	48.2%	0.424
RoBERTa	10	41.0%	0.339
DistilBERT	8	43.8%	0.370

Table 4.1: Table comparing the best performance across the different BERT architectures for emotion recognition.

Distillation is the process of training a smaller model on the output of larger models. The DistilBERT model uses similar parameters to BERT-base, with half as many layers.

Comparison and Choice

A comparison of the best performance achieved for each model can be seen in Table 4.1. From this we can see that quite significantly the best performance was with BERT-large. Surprisingly, RoBERTa had the worst performance across the architectures. DistilBERT does an excellent job being such a smaller model and is very comparable to BERT-base.

Additionally, the full graphs of Accuracy (Figure 4.3) and Macro F1 (Figure 4.4) show that BERT-large was not just an outlier with one good epoch and is consistently at the top after only a few epochs. The graphs suggest that RoBERTa was starting to look much better towards the end. However, unless we fine-tune the model and then freeze it and substitute this into the emotion recognition model, RoBERTa will not have time to mature to perform as well as BERT-base or BERT-large. DistilBERT shows satisfactory performance and will be much lighter on memory, so it is worth considering as a replacement for BERT-large or BERT-base if it becomes essential to reduce memory in the full model.

It looks like with further training RoBERTa may become more competitive, however, the full emotion recognition model will have began to overfit by 10 epochs. Since we are training in an end-to-end method RoBERTa is not applicable, however, if RoBERTa was fine-tuned outside the model and inserted into the emotion recognition model afterwards this may be even better than BERT-large.

Based on these results, we decided to use BERT-large. Fortunately, PyTorch is very modular, and future work on NLP would simply need to change the NLP module inside the emotion recognition model to replace BERT-large with another pre-trained NLP model completely.

The final NLP model can be seen in Figure 4.5. From now on any mentions of the NLP model will refer to this figure.

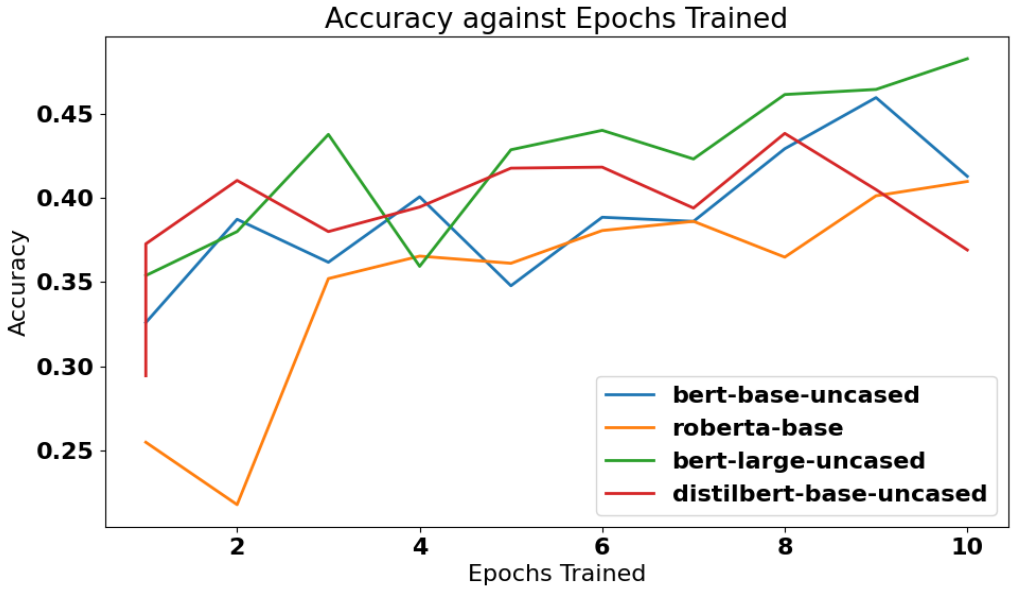


Figure 4.3: Accuracy of the different BERT architectures during training.

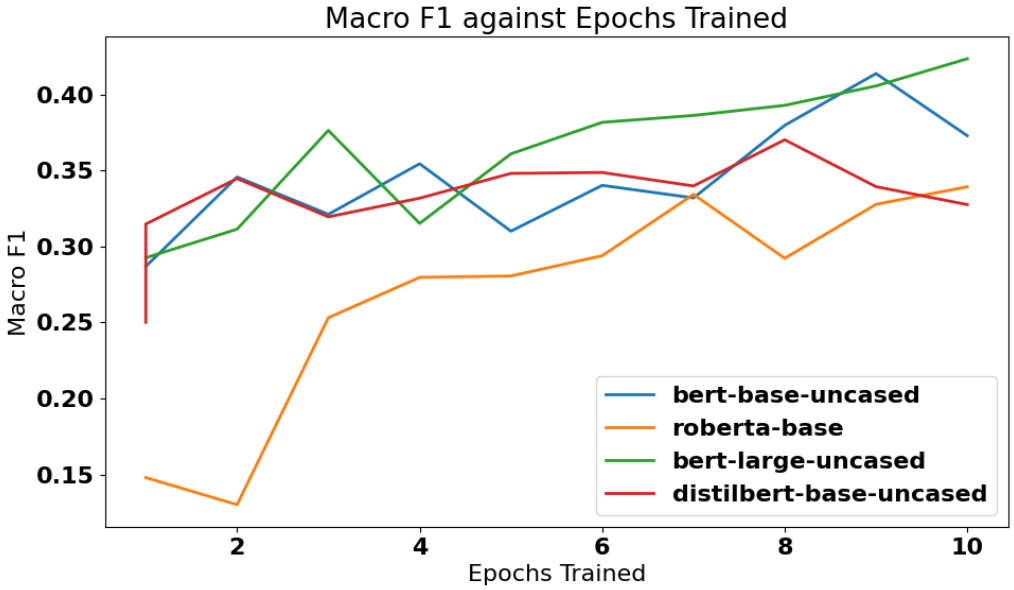


Figure 4.4: Macro F1 score of the different BERT architectures during training.

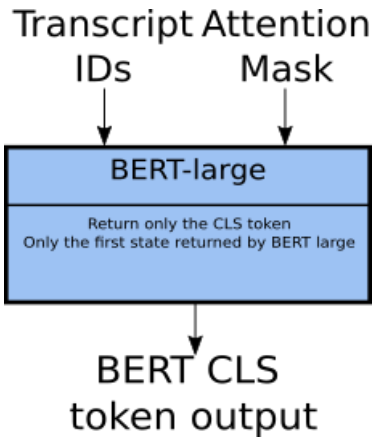


Figure 4.5: The final NLP model used, BERT-large can be swapped with DistilBERT to save memory.

4.3.3 Limitations

BERT-large is an incredibly heavy model, with 340 million parameters. Downloading the pre-trained model is over 1 gigabyte in size. Therefore the storage space of the trained model config, and the memory needed to run the model, are increased by a significant amount. It is found that during the evaluation of the model, this is not an issue, but while training the memory usage can be huge.

A solution to mitigate memory usage if storage space and memory becomes critical, would be to swap the NLP model with DistilBERT as performance should be comparable while significantly reducing the memory and storage required.

Since VR applications may be used with special goggles and smartphones, it is essential to preserve memory, and then in a DistilBERT implementation of the model may be superior to the BERT-large model.

4.4 Final Model

4.4.1 Fusion of Modalities

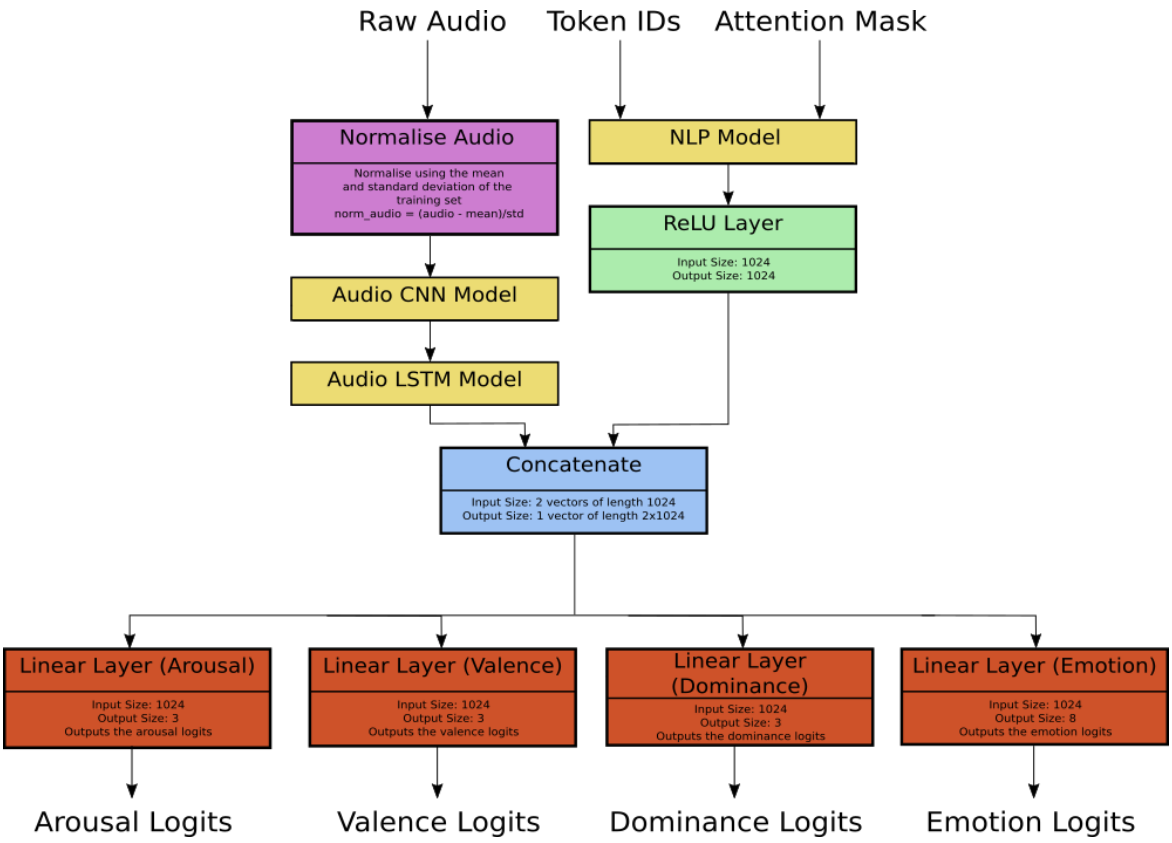
The fusion of the modalities is a fascinating problem, and gives an interesting insight into how speech and text can be utilised in tandem to facilitate emotional learning.

Initial attempts to fuse the models was a simple concatenation and linear layer output, as we see in some of the previous models [26; 27; 28]. It is found that this late fusion-based model had almost no improvement over text and audio individually. It was investigated if this was due to a difference in the outputs. In order to test this hypothesis, it was ensured the vectors were normalised and the same length before concatenation. However, the model still failed to learn a relationship between the text and audio.

Since the model could not learn any meaningful relationship, it is clear the fusion needs to occur earlier. With this in mind we tried a kind of early fusion, where we concatenated the features early on in the model and then applied the bidirectional LSTMs and then fully connected layers after the concatenation. Then we fully connect layers after the concatenation. This implementation resulted in an improvement over the late fusion model and is seen in the previous work [25].

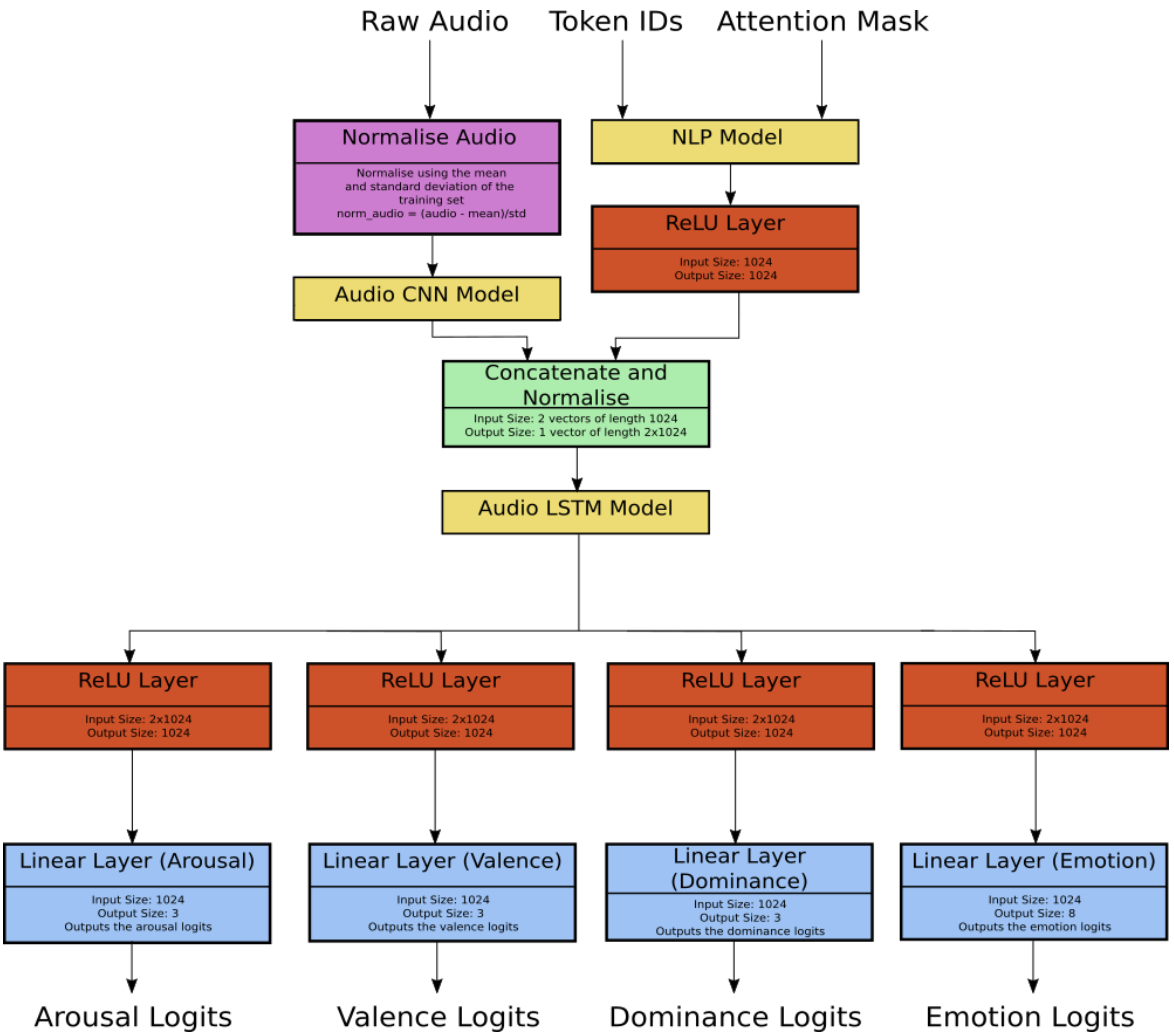
The model was further improved by making two changes. The first change was to use fusion to choose between the text feature vector, and speech feature vector, to produce a new feature space which hopefully learns to pick out the key aspects of each feature. This idea has been used in other fusion models for both emotion recognition [28] and depression prediction [32]. Additionally, a method similar to attention has been used as a Hierarchical Fusion method, where a standard linear layer is used with a single feature from each modality to produce one output feature. In this way they also learn which modality feature is important [30]. The second fundamental change was to apply the bidirectional LSTMs only to the speech vector and then use attention to choose between this vector and the text vector.

The different fusion architectures can be seen in Figure 4.6.



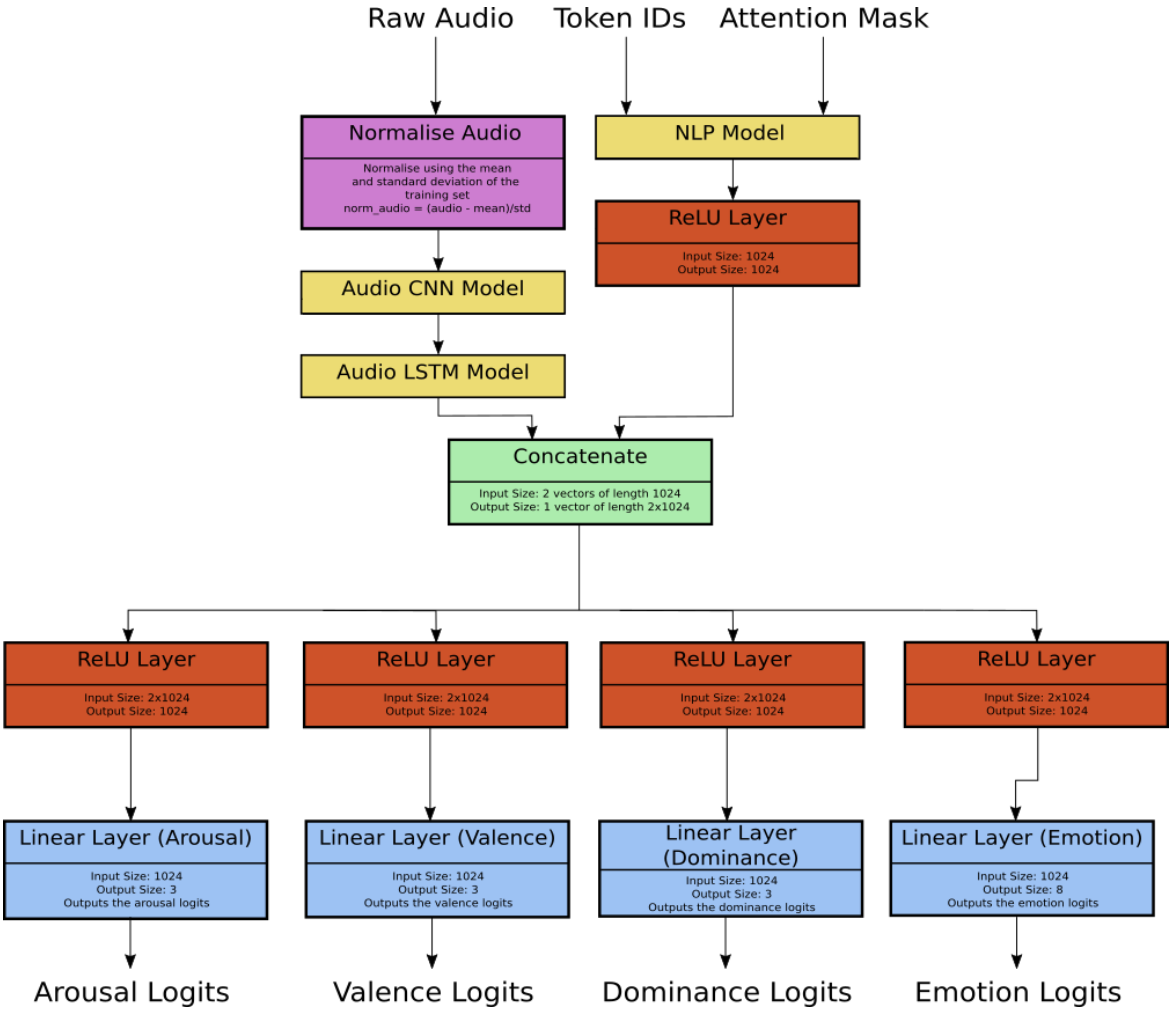
(a) Late Fusion

Figure 4.6: The 3 Different Fusion Architectures.



(b) Early Fusion

Figure 4.6: The 3 Different Fusion Architectures.



(c) Mid Fusion

Figure 4.6: The 3 Different Fusion Architectures.

4.4.2 Multi-task Learning

By training a model to solve more than one related task on the same data, we can prevent overfitting and improve the model's ability to generalise and adapt to new data. The prediction of arousal, valence, and dominance, on a single model, has shown to improve generalisation [24]. With this in mind, we implemented the prediction of arousal, valence, and dominance, as a secondary task to improve generalisation. The prediction of emotion is weighted such that it is loss function has three times the value of the loss for arousal, valence, and dominance.

Multi-task learning had little impact on the accuracy of the emotion prediction in the model, so it was kept. It did not make the model worse in any capacity, but it improved the potential use of the model as it can predict arousal, valence, and dominance, of speech with little difficulty.

4.4.3 Speaker-Level Memory and Self-Dependency

As discussed in previous work on emotion recognition, an essential task of modelling the emotions of speech is to consider the way each speaker influences their emotions [31]. Considering this, it seems important to have memory specific to the speaker, modelling their previous emotional states and how these will impact future predictions.

The context surrounding words in an utterance is crucial for understanding the emotions of an utterance [27; 30; 31]. This idea was extended to say that the emotional state of surrounding utterances is another critical aspect to the prediction of accurate emotion prediction. The original idea here was to use a bidirectional LSTM on the hidden state of the emotion prediction model since this can learn the dependencies and surrounding emotional contexts in the conversation. However, this could not run in real-time and would require a full conversation to generate all the hidden states which model emotions first.

Instead, we were able to implement a real-time compatible memory on previous emotional states by using a modified LSTM. When the model is provided with a batch of inputs, alongside this, a batch of speakers is given. Towards the end of the model, before the model head makes predictions on emotion state, the hidden state is passed through a custom LSTM. This custom LSTM treats the entire batch as a sequence, just as it would in a normal LSTM. But, when the speaker changes, the cell state is reset to 0. Since LSTMs have been shown to model short term, and long term memory, this LSTM enables the model to hang onto long term memory of the hidden state which models emotion inside the LSTM cell state. The cell state is forgotten when the speaker changes to improve the ability to train on multiple speakers.

This caused many complications in training since there needed to be a guarantee that each utterance from a conversation is loaded in the correct order, with only a single speaker. For example, given a conversation with two participants speaking intertwined with one another, uncoupling the utterances into the different speakers

was necessary, and ensure the same speaker's utterances were loaded together, in the correct order.

The significant advantage of this method is primarily that no pre-processing is required. Additionally, the memory is on a speaker level, and so it can be used on a single speaker. Other methods, such as the graph convolution, rely on both the speakers to accurately model the emotion fluctuations. These would rely on the AI therapist in an automatic therapy session [31].

The complete emotion recognition model can be seen in Figure 4.7.

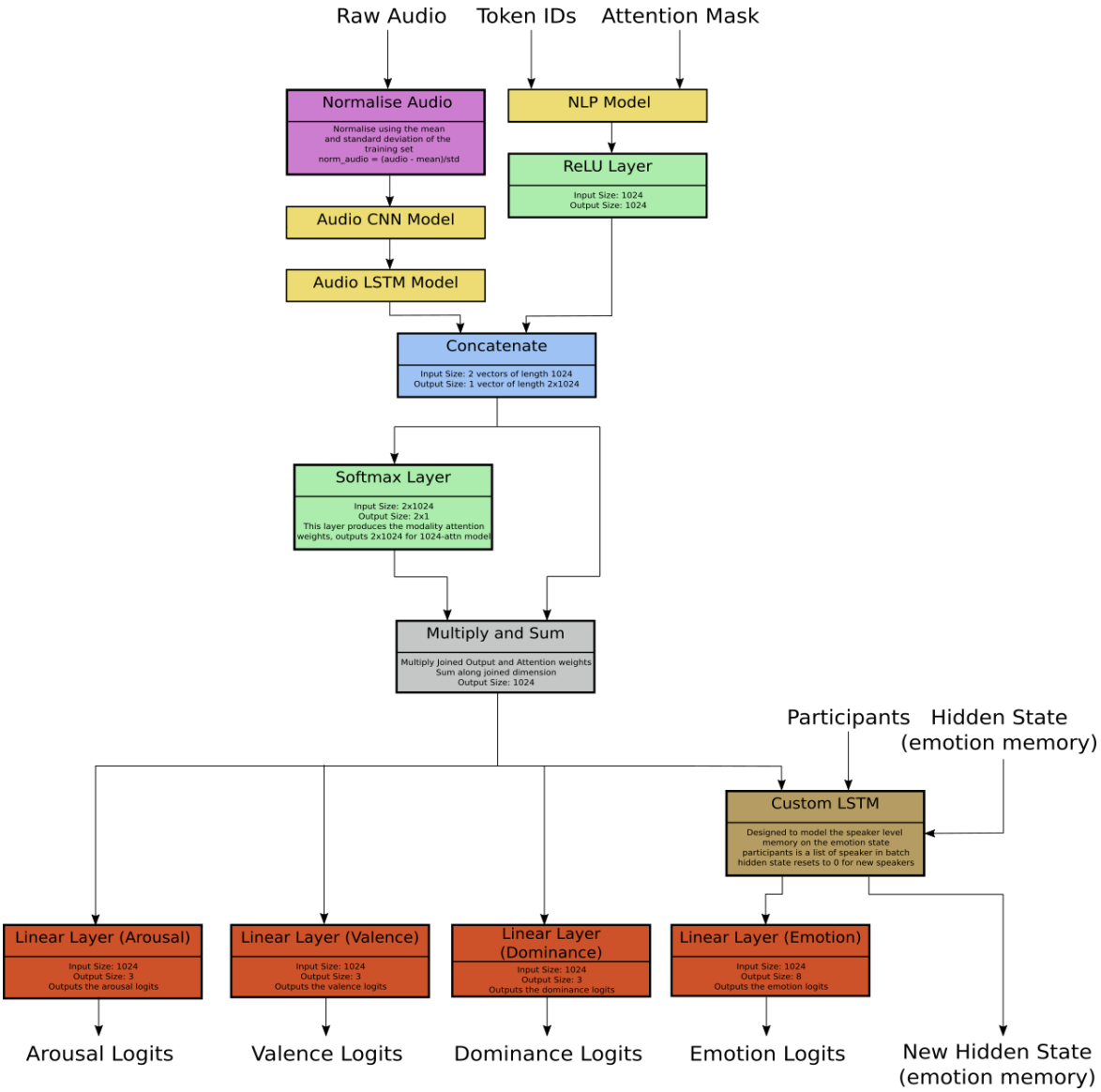


Figure 4.7: The full emotion recognition model.

4.5 Attention

4.5.1 Different Attention Methods

At the point of fusion, the model has a vector of length 1024, representing the text model output, and another vector of identical length representing the audio model output. An important decision to make regarding the attention is whether to generate separate attentions for the two main tasks. Shared attention might perform better due to improved generalisation as it prevents the attention from overfitting to the emotion prediction and tending to rely on text fully. On the other hand based on work reviewed during model design, it is reasonable to assume that text is much more valuable for emotion prediction and that audio is likely to be more valuable for arousal, valence, and dominance prediction. Perhaps performance would be improved by allowing the models to learn how much to focus on text/audio for each task.

A second option is whether the attention should choose just two weights, which it can then use for the entire vector, or should it predict 1024 pairs of weights allowing the model to learn which parts of the NLP/Audio features are more important than others, and vary these choices.

Based on both the accuracy and macro F1 score, the shared models outperform the split model for emotion prediction. The average accuracy and macro F1 can be seen in Figure 4.8.

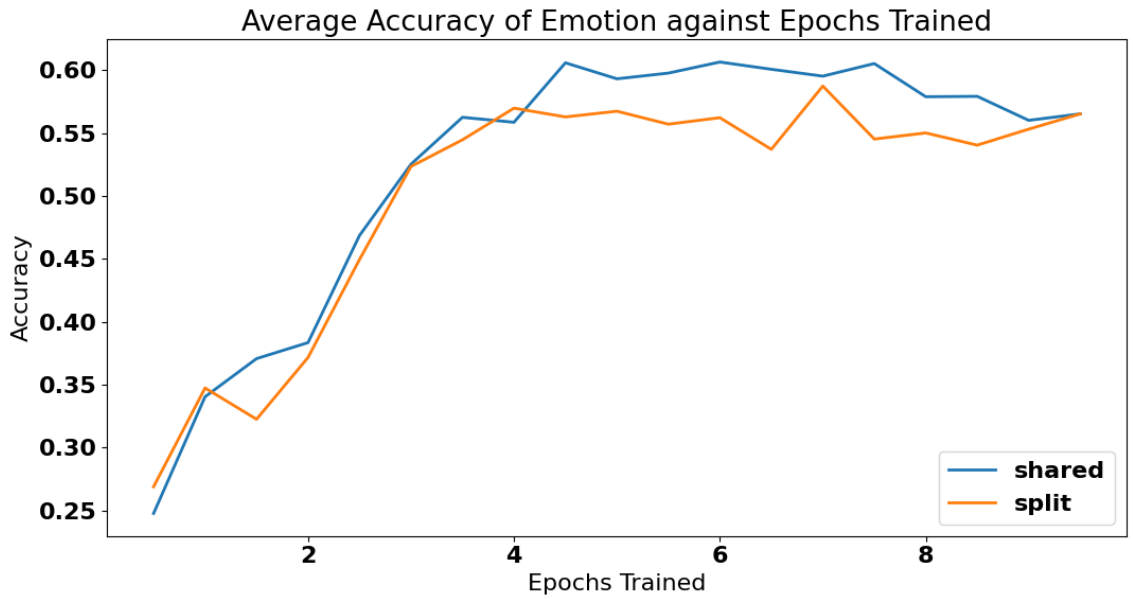
When considering only the shared models, with frequent evaluations over six training epochs, the full 1024 attention vector seems to trend just above the simple two attention vector. This can be seen in Figure 4.9. Regardless, we will explore how both the two attention vector, and the 1024 attention vector, choose between the text and audio models.

4.5.2 Attention Representation

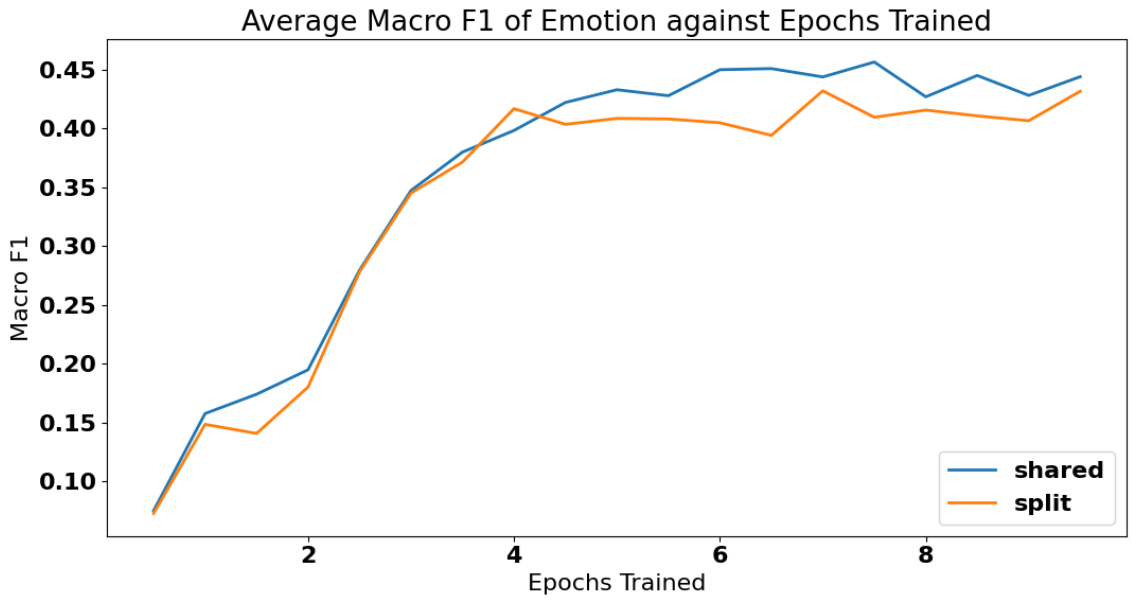
The attention in the model should represent how important text and audio are for different emotions. Investigating this, we should be able to see if any emotions trend towards text or audio. Since visualisation will be more transparent with the model that only uses two attention values, we will start by initially discussing this model, before briefly considering the full 1024 attention model.

By considering each utterance exclusively in the test set (Figure 4.10), we can see that angry utterances are much more likely to increase the weight of the audio modality significantly. Additionally, excited and surprised, are somewhat higher in the average weighting of the audio modality. This may not be a perfect representation; however, since incorrect predictions may also cause the attention not to be applied as expected.

If we instead filter this to only the instances where the predicted emotion in the model was correct (Figure 4.11), we can get a more accurate picture of what the

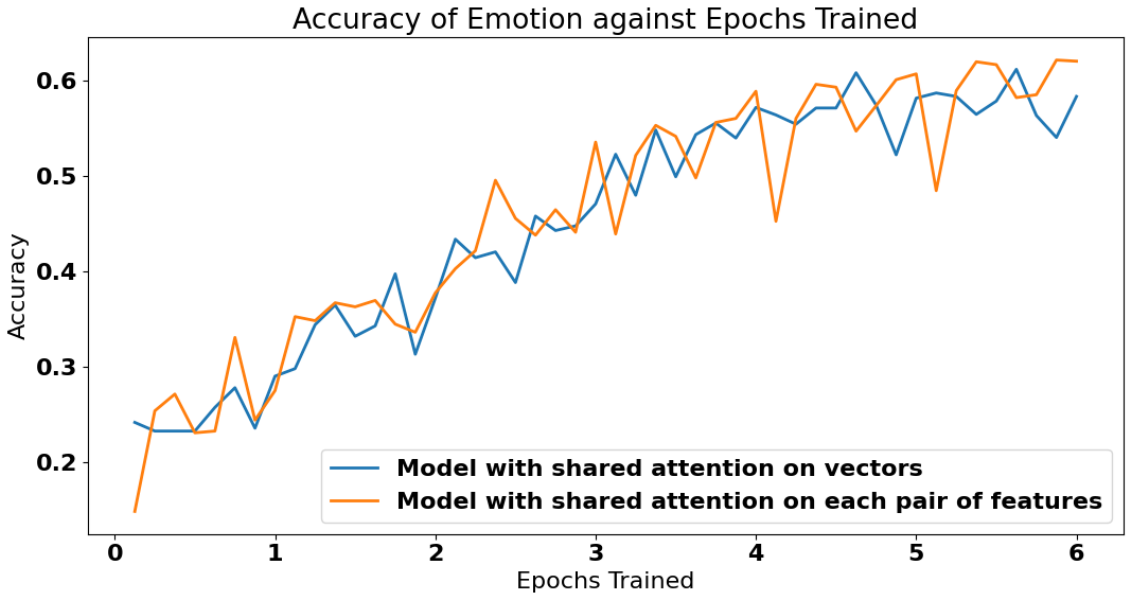


(a) Average Accuracy

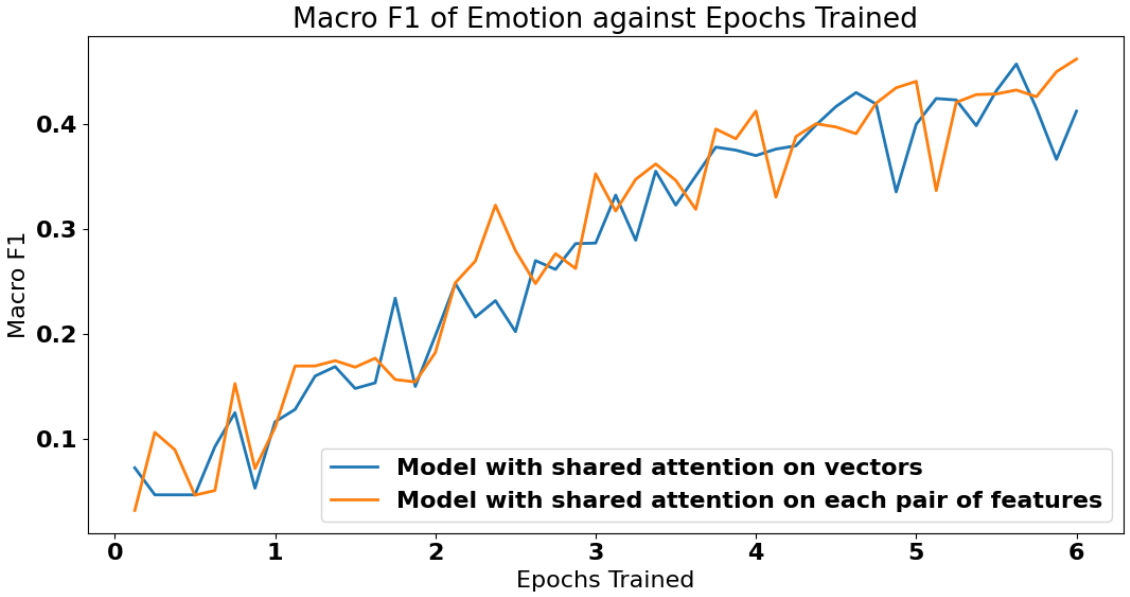


(b) Average Macro F1

Figure 4.8: Averages for the shared and split attentions on emotion prediction.



(a) Accuracy



(b) Macro F1

Figure 4.9: Accuracy and Macro F1 for emotion prediction on the shared attention models.

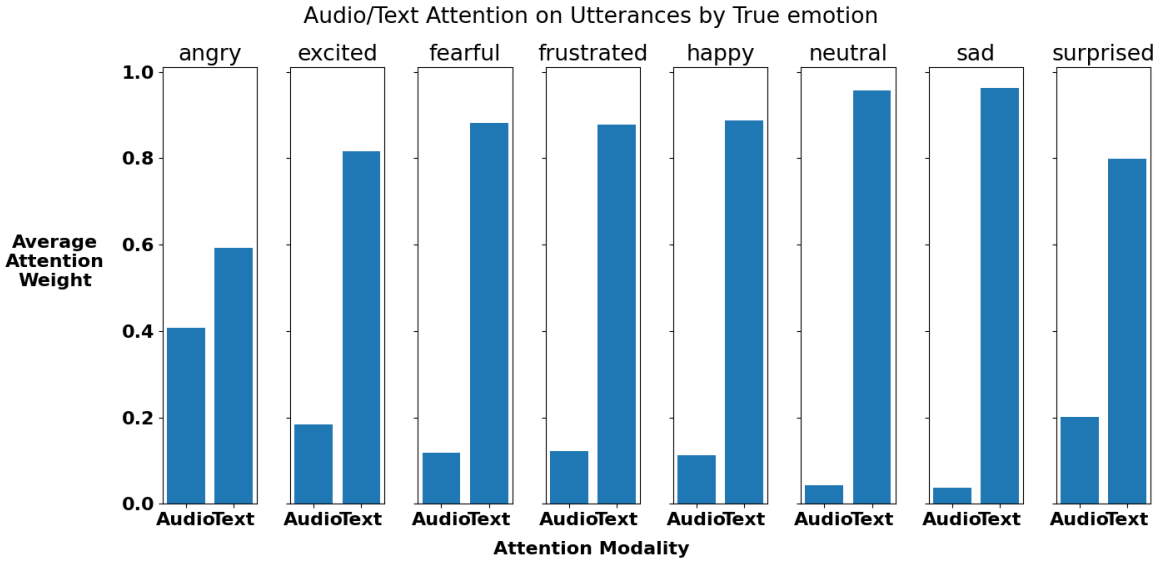


Figure 4.10: Average attention weights for each (true) emotion of the utterances.

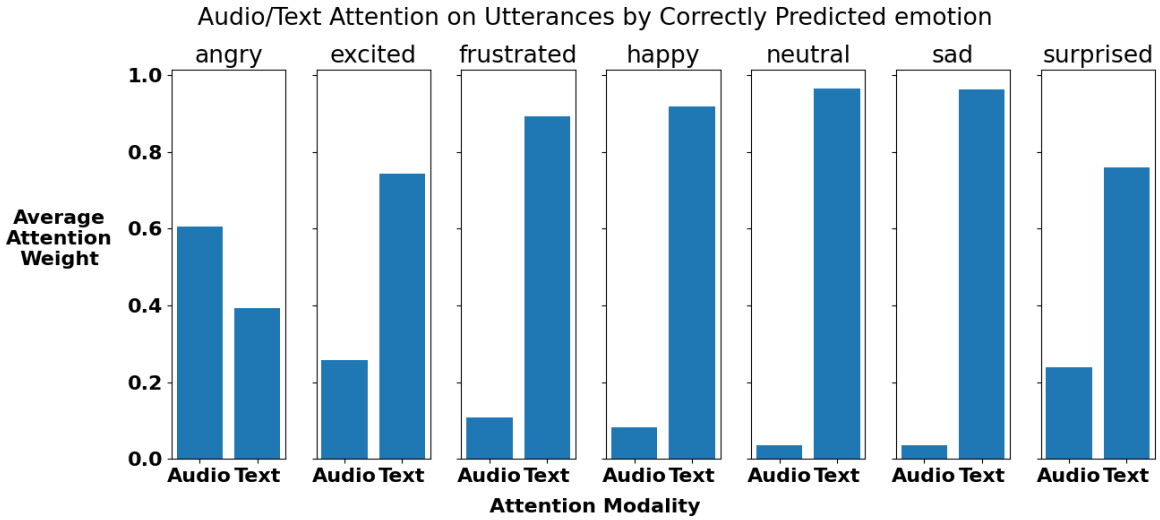


Figure 4.11: Average attention weights for each emotion when the prediction of utterance emotion by the model was correct.

model has learned. It is evident that angry, excited, and surprised, rely much more heavily on the audio modality for prediction, and something about these emotions must make it harder to detect purely from the text. On the other hand, we see that sad and neutral in particular are almost entirely reliant on the text modality.

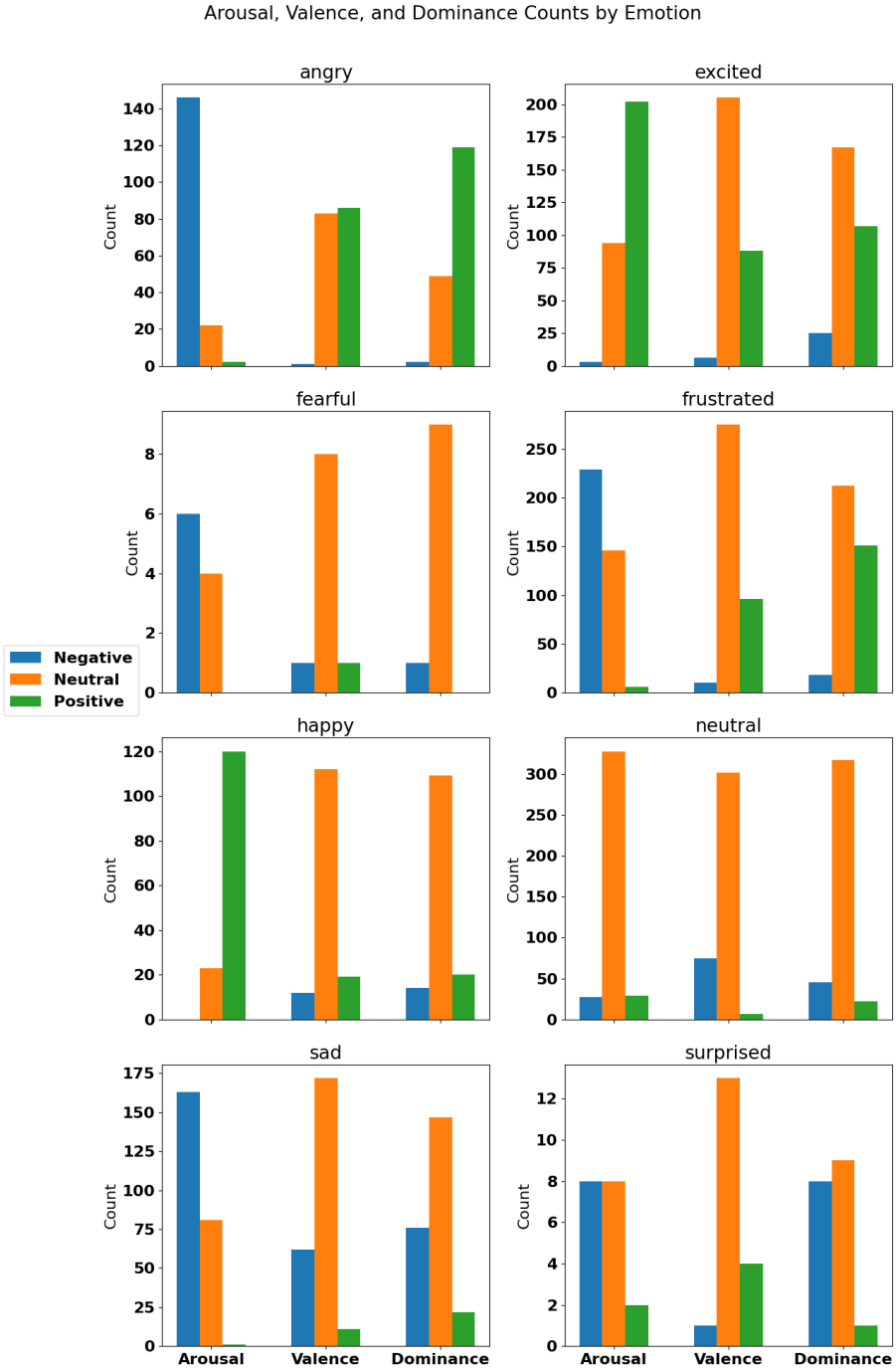


Figure 4.12: Arousal, valence, and dominance class counts per emotion in IEMOCAP dataset.

If we look at how common the arousal, valence, and dominance is for each emotion,

and consider that angry, excited, and surprised, required more reliance on audio, it may be possible to see what features of the speech cannot be detected from audio and are important for emotion.

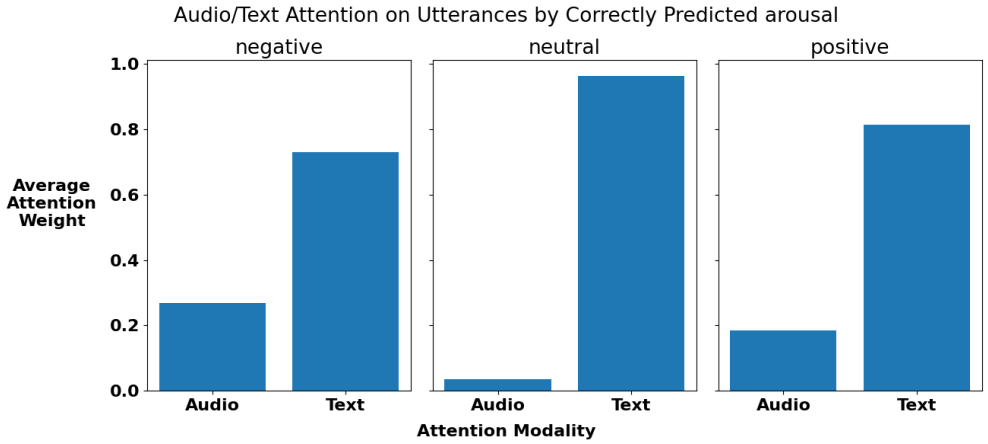
First, it is important to discuss what the negative, neutral, and positive categories mean based on the ANVIL ratings used in IEMOCAP [20]. For arousal, negative will represent that there is no excitement/activation in the voice, while positive will represent a lot of activity and excitement in the voice [20]. For valence, the rating is determined from a negative to positive sound in the voice. For example, sad should have a negative valence while happy will be positive valence [20]. Finally, for dominance, the rating uses negative to mean a small, or quieter, voice, while a positive dominance indicates a dominant and louder voice.

Graphing the counts for each category of the arousal, valence, and dominance separated by the emotion (Figure 4.12), we can see a similarity between the angry and excited emotions, especially. Angry, excited, and frustrated all tend to positive dominance and positive valence more often than other emotions. Additionally, surprised is more likely to tend to positive valence, but not positive dominance. It seems that in general, emotions with a stronger positive reaction in either valence and dominance tend to require more weight on the audio modality. Frustration is an exception to this rule, however.

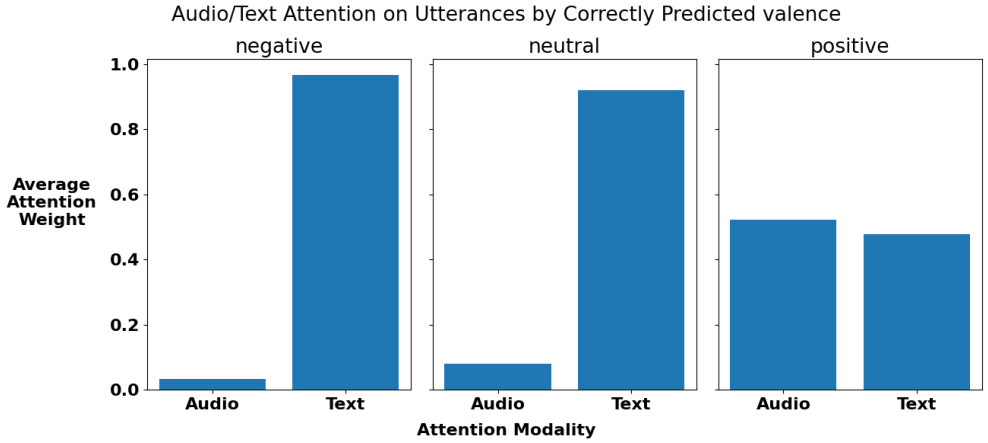
This is further verified when we consider the correctly predicted arousal, valence, and dominance, and their corresponding average attention weights (Figure 4.13). Evidently, audio becomes far more important when the valence or dominance is positive. Arousal seems to influence somewhat when it is negative or positive. However, this may be because angry is overwhelmingly negative in arousal, and excited is similarly very positive arousal.

Visualising the 1024 attention vector is somewhat less useful. The model is now free to pick and choose which aspects of each modality is important. Without precisely knowing what the 1024 audio features and 1024 text features truly represent, we may not be able to reach a conclusion. However, if we consider the box and whisker plot of the average 1024 weights for the audio modality (Figure 4.14), we can see that the range of common values fluctuates more for different emotions. We can see the emotions which needed to look more at the audio vector, especially angry and excited, have a larger interquartile range, and the third quartile point occurs at a higher audio attention weight.

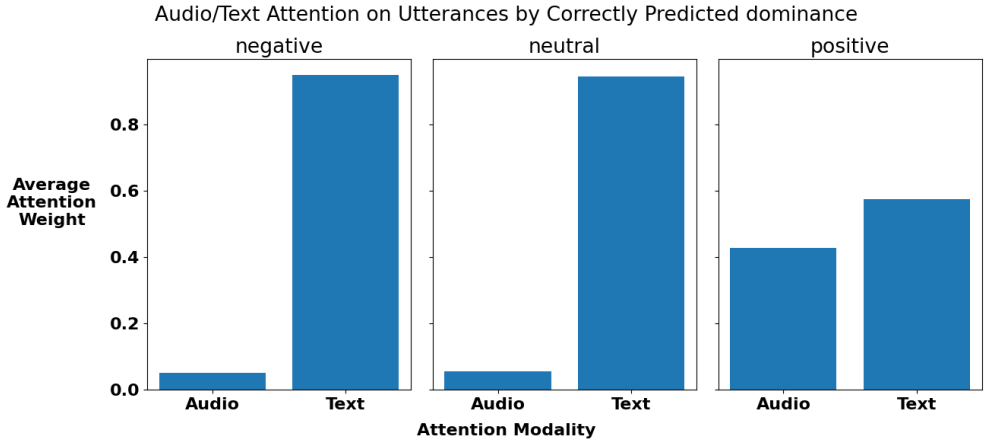
Considering both the 2 attention vector (Figure 4.11) and the 1024 attention vector (Figure 4.14) there is a significant disagreement. In the 1024 attention vector there seems to be no significant difference between emotions and how they utilise modalities. However, the 2 attention vector clearly favours certain modalities for each emotion. One could hypothesise that this is because particular features in the audio modality are important for reducing confusion, and when there are only 2 attentions the model is forced to increase the weight of the entire audio modality to understand these features. Meanwhile, the 1024 attention model is able to extract only the important features from audio, and can continue to rely on text for the



(a) Arousal



(b) Valence



(c) Dominance

Figure 4.13: Attention weights for each modality based on correct predictions of arousal, valence, and dominance.

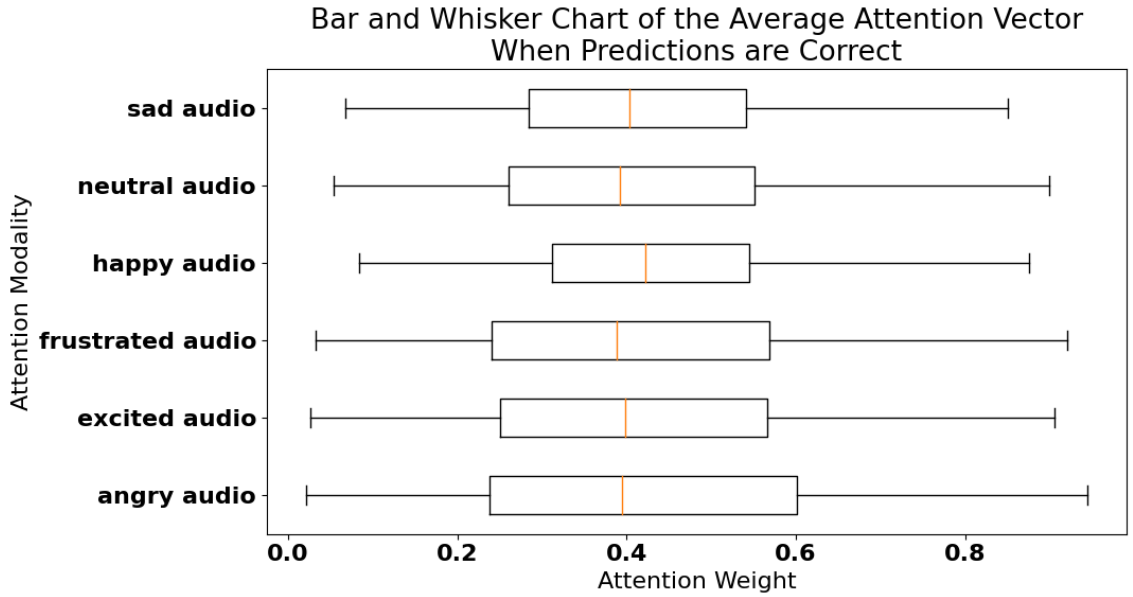


Figure 4.14: Average audio attention weight vector box and whisker plots for each correctly predicted emotion.

majority of the features.

It is worth noting that surprise does not feature in this graph, as the 1024 feature vector fails to predict surprise at all correctly. While the overall accuracy is slightly higher in the 1024 feature vector model, we lose understanding of how the modalities are fused, with relatively insignificant performance increase, as can be seen in Figure 4.9.

Chapter 5

Depression and Anxiety Prediction

5.1 Depression and PTSD

5.1.1 Theory

We intend to show that the emotion model can be used to improve the accuracy of depression prediction models and that the accuracy of the emotion model generalises to interviews with non-actor speakers.

To demonstrate the former, we intend to show that information learned about emotion can be applied in a cross-corpus manner to improve depression predictions. In this section we train a baseline model on the E-DAIC dataset [22; 23] to predict PHQ-8 scores primarily, but also the PTSD severity as a secondary objective. Following this we train the same architecture, but with a simple fusion with the hidden state from the emotion recognition model, and compare how this affects performance.

If information from the emotion recognition model makes improvements to the quality of depression prediction, we assume that the information extracted by the emotion recognition model is valid. If the emotion recognition model introduced meaningless noise it would not aid the model. Additionally, we investigate the relationship between the emotion recognition model predictions and the predictions on each utterance, as well as for the interview as a whole.

5.1.2 Model

In order to look at per utterance results, we intend to construct the model such that we can visualise per utterance predictions, but make a final prediction considering the entire interview.

Since we intend to keep the model straight forward, we opted to use a modification of the emotion recognition model to extract features from each utterance. Then, using the entire list of features extracted from each utterance in an interview, we use a second model to aggregate these and make a prediction for each utterance.

Then, we use attention based weighting with this to reach the final prediction for the interview.

Architecture

The first model is the feature extractor. This model extracts a feature vector from each utterance in the interview. Since the emotion recognition model showed very strong signs of modelling emotional states so well, we hypothesise that a similar architecture will have capacity to model features from speech about depression. Looking at the architecture in Figure 5.1 it is clear only minor changes have been made from the emotion recognition model in Figure 4.7.

We then have a model which can take the list of all utterance features, and the list of emotion recognition hidden states, to produce predictions for the full interview. For this model, we decided to approach it in a manner that would allow investigation into the relationship of utterance level predictions and final predictions, as well as into the weight applied to different parts of the interview. To do this, we use a linear layer to make utterance level predictions, and then use an attention based weighting to create a final prediction for the full interview. This can be seen in Figure 5.2.

As a baseline for comparison we use the same overall architecture as above, however, the aggregator takes only the extracted features as input. Thus, the concatenation layer is missing and the corresponding vector sizes throughout the model are halved.

Training

Ideally when training the model we could generate the feature vector for each utterance, then create a label with the aggregator and backward propagate through both models, and take this as a single sample. However, due to memory constraints, this is not possible as the gradients tracked in the calculation of the feature vectors rapidly builds up.

To mitigate this, I decided to train in sub-batches of utterances. Every four utterances in an interview, I use the aggregator to make a prediction, and then back propagate through the networks. This allows the gradients associated with the previous utterances to be freed. In order to make this as accurate as possible, I attach the previously freed feature vectors such that they influence the final prediction from the aggregator, but they do not influence the training of the feature extractor.

For example, after three sub-batches, during the fourth batch we will have twelve feature vectors with no gradient, and four feature vectors with gradients. The aggregator model takes these sixteen vectors to make a prediction for the “interview” consisting of sixteen utterances. The loss function then back propagates through the aggregator network, using the sixteen vector inputs to calculate updates made to the network parameters. Then, the four feature vectors with attached gradients, will continue to back propagate through the feature extractor network updating these.

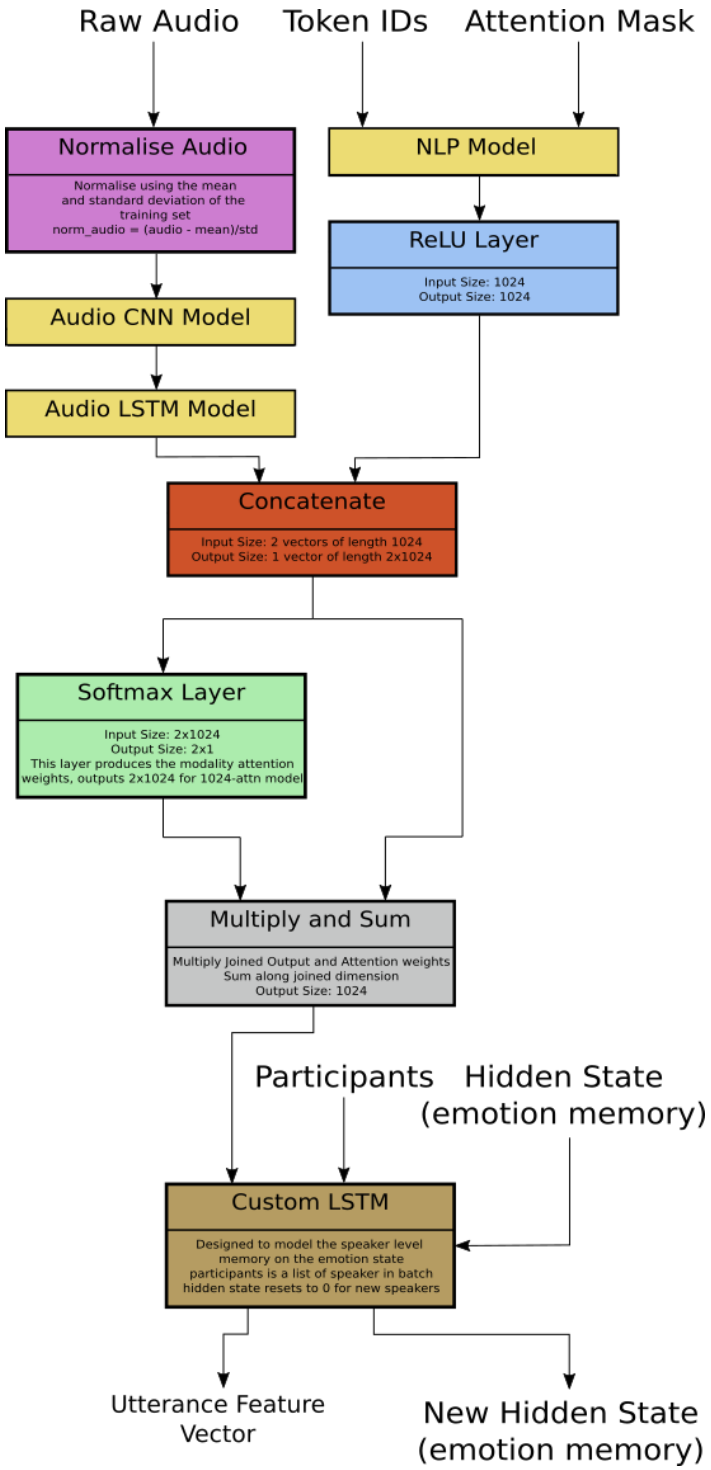


Figure 5.1: Feature extractor trained on PHQ 8 and PTSD severity labels.

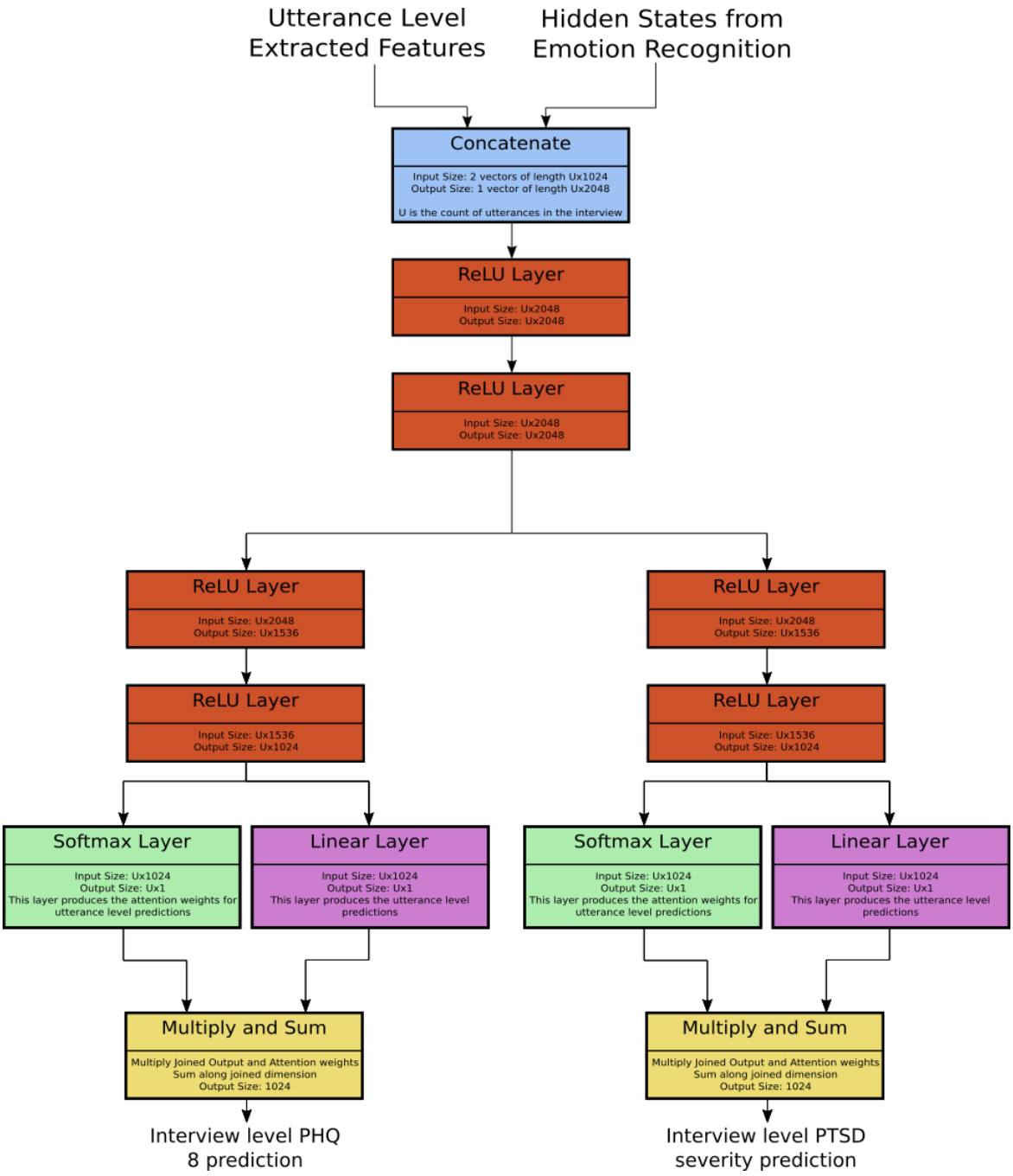


Figure 5.2: Aggregator used to make final interview level predictions.

This is best explained in Figure 5.3, a simple example using six utterances and sub-batch size of two.

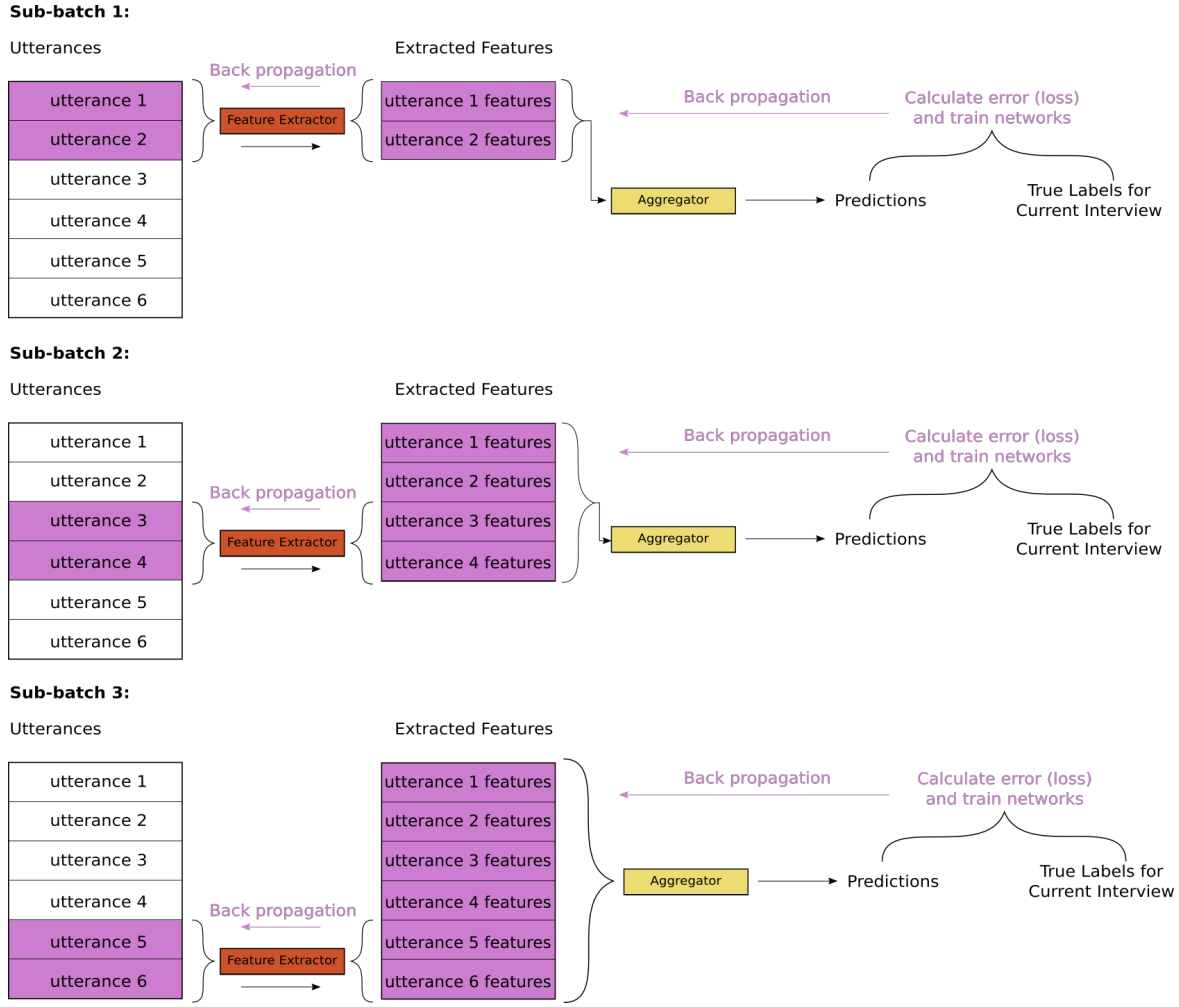


Figure 5.3: Training example with an interview consisting of six utterances, with a sub-batch size of two.

5.1.3 Results

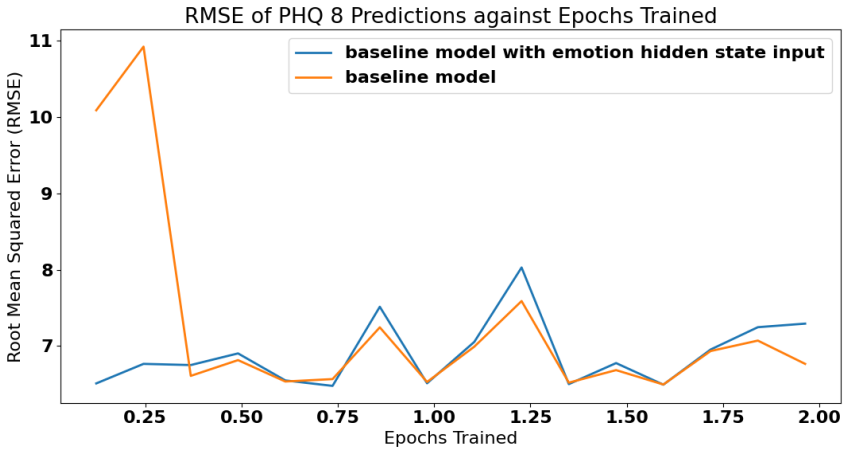
Comparison with Baseline

Ideally, we hope to see that the emotion recognition model does significantly better at the prediction of PHQ 8 scores. However, we can see from Table 5.2 that taking the average over five runs only achieves an improvement of 0.02 in the RMSE.

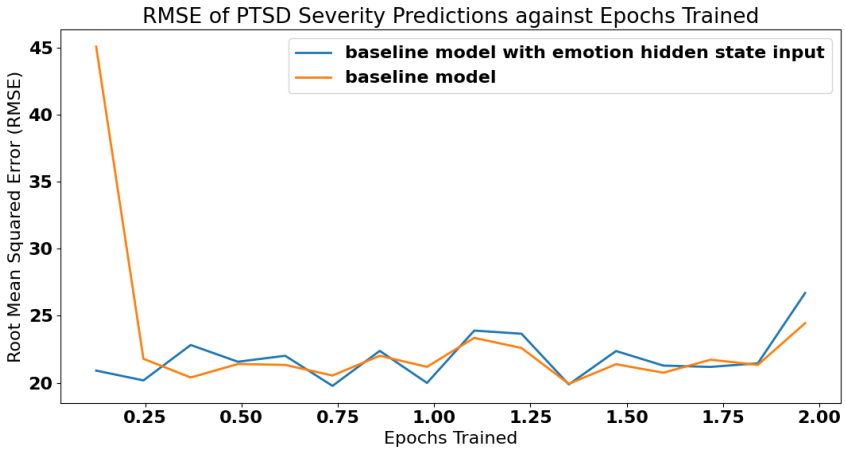
There is a consistent, slight, improvement here. In addition, if we graph the average RMSE at each evaluation point while training each model five times for two epochs (Figure 5.4), we can see that the amount of data required before the model with emotion states reaches the optimum is significantly less. To show this more clearly, I train each model again for 1 epoch using significantly more evaluation points during training; this can be seen in Figure 5.5.

Model	RMSE	MSE	MAE
Baseline	6.494	42.173	**5.564**
Baseline with Emotion States	**6.475**	**41.922**	5.569

Table 5.1: Table comparing the average performance on PHQ 8 score regression on the test split after training the model for two epochs five times.

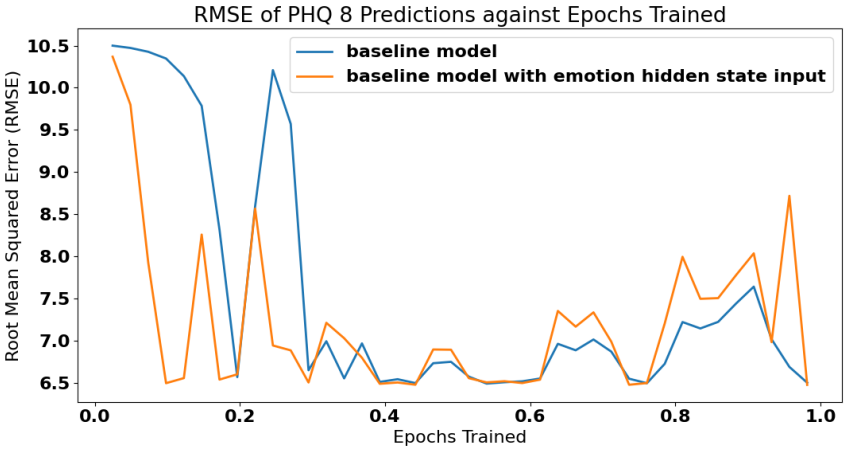


(a) PHQ 8

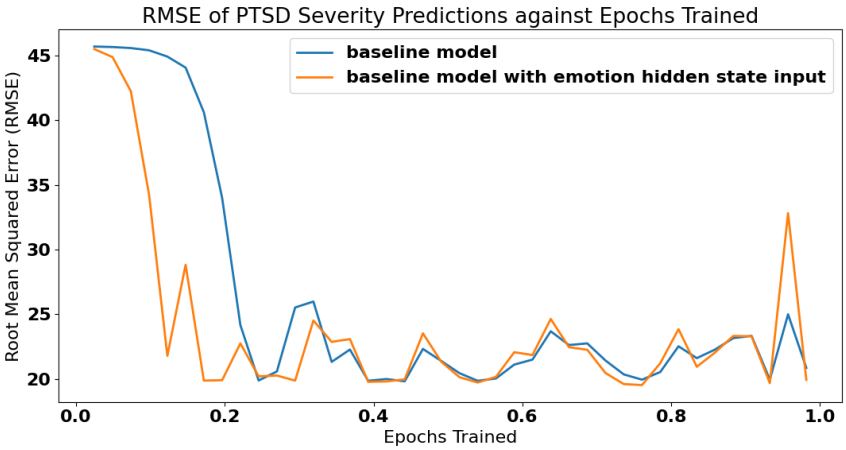


(b) PTSD Severity

Figure 5.4: RMSE on test set of both PHQ 8 scores and PTSD severity scores averaged over 5 runs. Results are reported on RMSE, Mean Squared Error (MSE), and Mean Average Error (MAE).



(a) PHQ 8



(b) PTSD Severity

Figure 5.5: RMSE on test set of both PHQ 8 scores and PTSD severity scores on 1 training epoch with frequent evaluations on the test set.

5.1.4 Limitations

There is evidence that the depression model benefits from emotion inputs, however, the quality of what is learned about depression is questionable. The range of predictions made, for both PHQ 8 scores and PTSD severity scores, is quite small, and seems to tend towards the average labels in the training set. The mean PHQ 8 and PTSD severity scores in the test set are 6.66 and 35.60 respectively. During evaluation on the test set, the mean PHQ 8 prediction is 7.56 and the mean PTSD severity prediction is 35.82. Evaluation on the validation set similarly produces mean predictions of 7.55 and 35.79. The similarity in these numbers indicates very limited understanding of the underlying data.

Additionally, if we consider the utterance level predictions made by the aggregator, there is some fluctuation, but as seen in Figure 5.6, the values are floating around the interview level prediction.

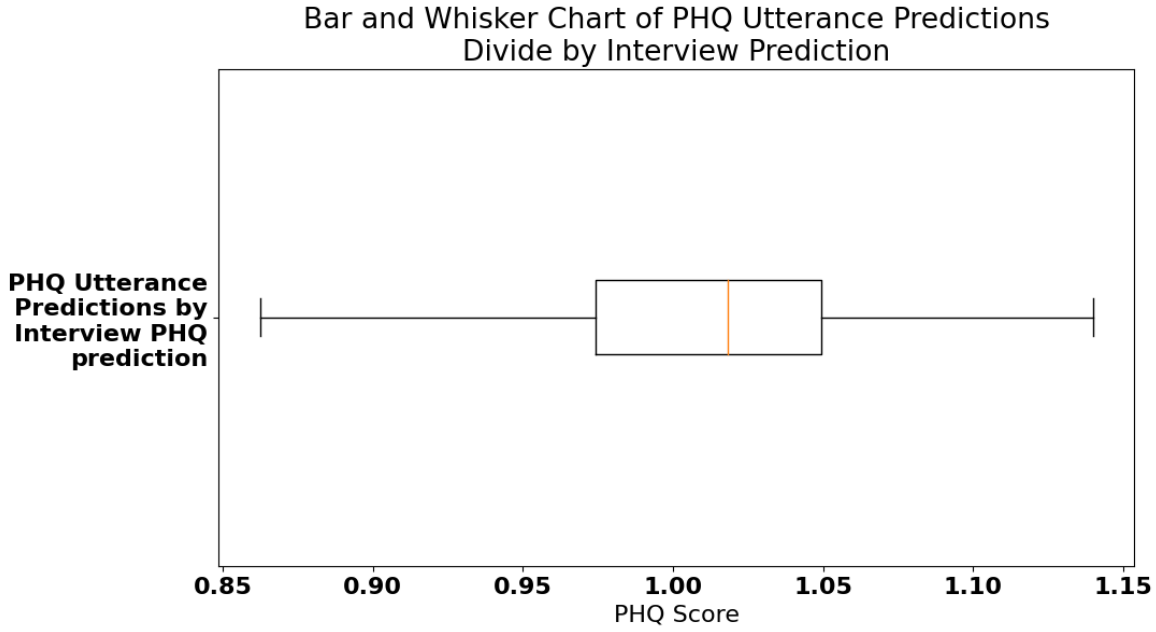


Figure 5.6: Box and whisker plot of the utterance level predictions, divided by the interview level prediction. A value of 1.0 indicates the utterance had the same prediction as the interview as a whole.

This behaviour is consistent for both PHQ and PTSD predictions. I would hypothesise that this is a consequence of training the feature extractor in conjunction with the aggregator. I suspect that the aggregator fails to extract adequate features from the data, but it gradually begins to learn something about PHQ, hence why we see some slight variation in PHQ prediction while PTSD just matches the mean value. The implementation which uses emotion states can use the information extracted from the utterances by the emotion recognition model to begin training the aggregator while the feature extractor is failing to extract anything meaningful. This might explain why the model with input from the emotion model trains much faster.

5.2 COVID-19 Interviews

5.2.1 Anxiety Fine-tuning

Since there is very limited data available for anxiety prediction, we take advantage of cross-corpus understanding and take the best performing depression prediction model defined above. We then add an additional head to this model, and freeze the feature extractor model and most layers in the aggregator. Information learned about depression, PTSD severity, and emotional states, should allow training on anxiety using a very small training set. The modified aggregator can be seen in Figure 5.7 and indicates which layers have frozen parameters preventing any training.

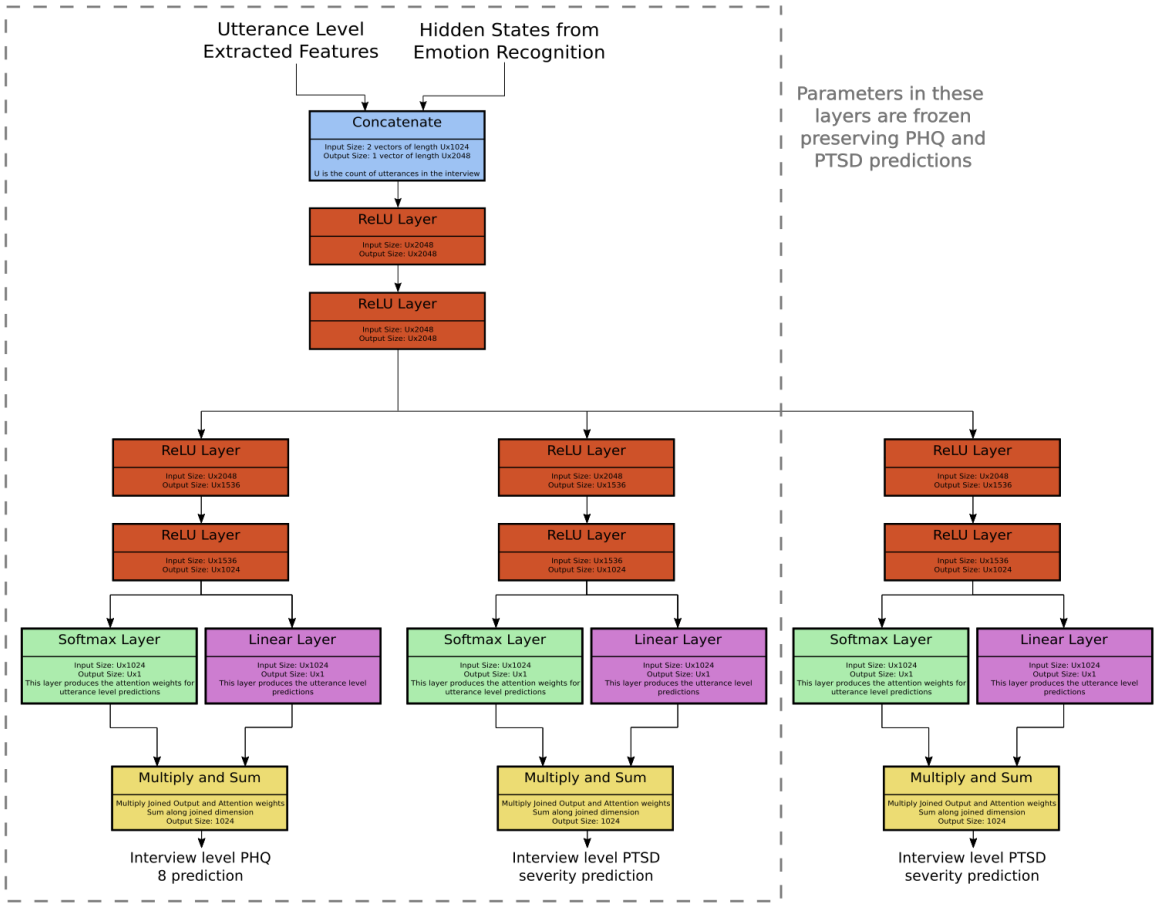


Figure 5.7: Aggregator with added head and frozen, pre-trained, layers.

Since the feature extractor is no longer being trained in this case, training is straight forward. First we freeze the layers as indicated above, then we use the feature extractor and emotion recognition on each utterance in the interview. The aggregator will then produce PHQ and PTSD predictions as normal, and an additional GAD 7 prediction.

Model	RMSE	MSE	MAE
Baseline	2.27	5.15	**1.73**
Baseline with Emotion States	**2.26**	**5.12**	1.77

Table 5.2: Table comparing the performance on GAD 7 prediction between the baseline model, and the model with emotion state input.

5.2.2 Results

To create a baseline to compare to, we use the baseline depression model and fine-tune this on GAD 7. The behaviour of the depression is replicated again, where we see that the model with emotion states trains faster, but has little improvement to the RMSE of the values.

Graphing the GAD 7 RMSE against epochs trained (Figure 5.9) confirms that the capacity of each model is very similar, but the additional information gained from the emotional states helps to train faster. This is particularly interesting, since the model pre-trained on depression without emotional state input takes longer to train.

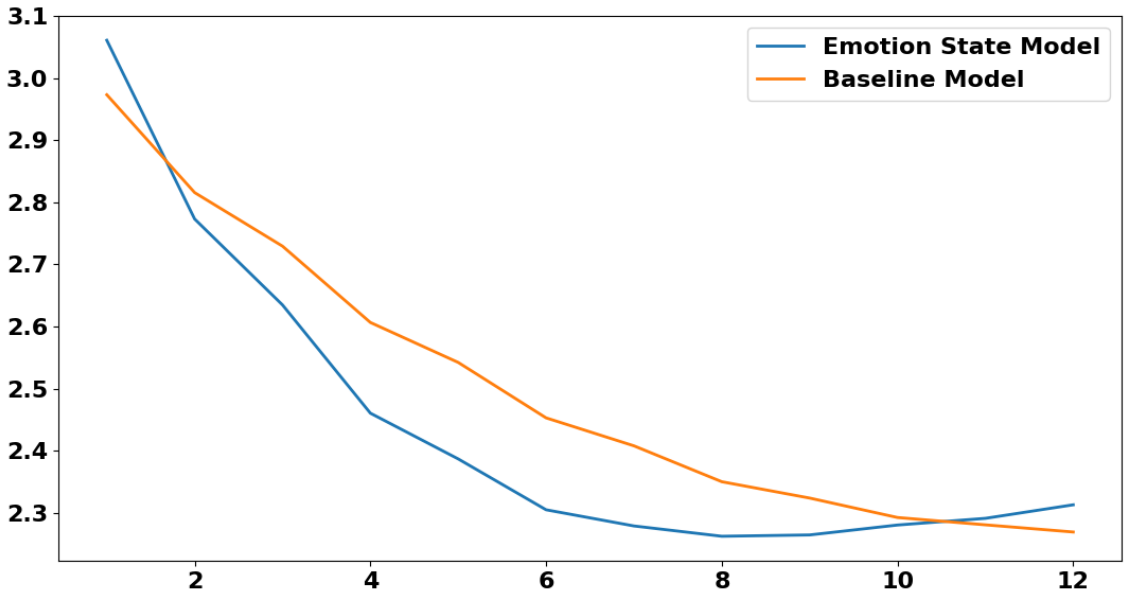


Figure 5.8: GAD 7 RMSE against epochs trained for anxiety fine-tuning models.

Similar behaviour to the depression model is seen in the anxiety model as well. I hypothesise that this is due to the feature extraction model not having learnt enough information about depression. The automatic transcription of these interviews was not perfect, and could have been a source of incorrect emotional understanding, which could also have contributed towards a worse understanding of anxiety by the model.

Considering the clinician assessed Likert scores of anxiety, it appears there could be a correlation indicating some level of understanding by the model. However, it may

be due to less data on the upper and lower ends of the Likert scales, and by chance the model has predicted higher/lower on these few samples.

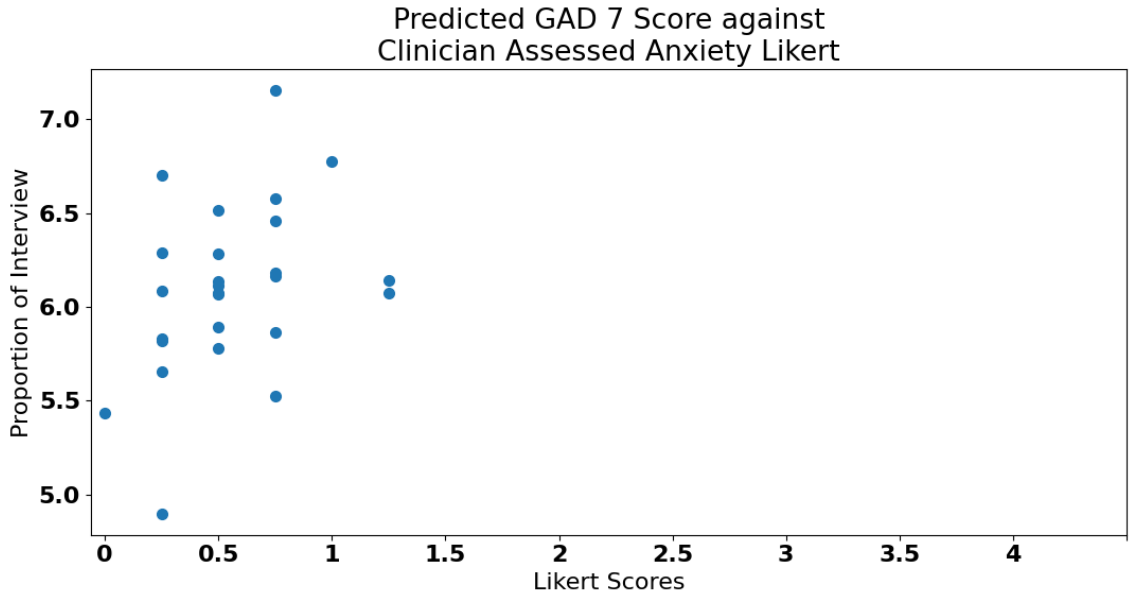


Figure 5.9: Predicted GAD 7 scores against clinician assessed anxiety Likert scores.

Chapter 6

Evaluation

6.1 Emotion Recognition

6.1.1 Comparison with Previous Work

Since many state-of-art implementations of IEMOCAP report their results on different subsets of emotions, the performance is considered in relation with each subset.

Additionally, results on both versions of modality attention are reported on in this model. We refer to the model which generates two weights for each audio and text modality as the `2-attn` model. Similarly, the model with 1024 weights choosing between each extracted audio and text feature will be referred to as `1024-attn` model.

When comparing with other models that have produced models for multiple modalities, we consider only their best-performing modalities.

Additionally, it is worth noting that the different models compared throughout this section have used different training-test splits on the IEMOCAP dataset.

In the following sections we adjust the model to predict subsets of the full 8 emotions from IEMOCAP. This allows us to compare our model against several different state of the art models which operate on different sets of emotions.

Anger, Happiness, Neutral State, and Sadness

Comparing my model with previous state-of-art emotion recognition models predicting anger, happiness, neutral state, and sadness, shows that both the `2-attn` and `1024-attn` models are competitive. Table 6.1 shows that the models significantly out perform previous models on accuracy and weighted average precision. Additionally Table 6.2 compares the class specific performance on previous models which report these results.

Model	Modalities	Accuracy	Weighted Average Precision
MDRE [28]*	Text + Audio	-	0.718
Combined Model [26]*	Text + Audio + Video	71.04%	-
bc-LSTM [46]	Text + Audio + Video	76.1%	-
Text and MFCC [27]	Text + Audio	76.1%	-
CHFusion [30]	Text + Audio + Video	76.5%	-
2-attn	Text + Audio	**80.3**%	**0.808**
1024-attn	Text + Audio	80.0%	0.801

Table 6.1: Table comparing overall performance with state-of-art models. Best performances are highlighted in bold.
* These models have treated excitement as happiness and either joined the two emotions, or used excitement as the happiness label.

Model	Anger Accuracy	Happy Accuracy	Neutral Accuracy	Sad Accuracy
MDRE [28]*	78.85%	79.08%	76.76%	66.67%
bc-LSTM [46]	77.98%	**79.31%**	78.30%	69.92%
Text and MFCC [27]	63.41%	49.24%	84.06%	81.30%
CHFusion [30]	79.6%	74.3%	75.6%	78.4%
2-attn	74.7%	66.4%	81.6%	**86.9%**
1024-attn	**83.5%**	62.2%	**84.4%**	81.2%

Table 6.2: Table comparing class performance with state-of-art models. Best performances are highlighted in bold.
* These models have treated excitement as happiness and either joined the two emotions, or used excitement as the happiness label.

Model	Modalities	Accuracy
E2 [29]	Text + Audio	70.1%
2-attn	Text + Audio	77.1%
1024-attn	Text + Audio	**77.7%**

Table 6.3: Table comparing overall performance with state-of-art model. Best performances are highlighted in bold.

Model	Anger Accuracy	Happy Accuracy	Neutral Accuracy	Sad Accuracy	Fearful Accuracy	Surprise Accuracy
E2 [29]	52.2%	59.0%	44.1%	77.7%	**100.0%**	**98.8%**
2-attn	74.7%	**66.4%**	83.3%	**86.9%**	0.0%	0.1%
1024-attn	**83.5%**	62.2%	**84.4%**	81.2%	0.0%	0.0%

Table 6.4: Table comparing class performance with state-of-art models. Best performances are highlighted in bold.

When considering both the overall accuracies and the individual class accuracies, the model remains highly competitive outperforming most models in both. However, it is worth noting that the happy accuracy is significantly lower than the other three classes, and of these four classes happiness was the least well represented in the dataset, suggesting that some kind of upsampling may result in a better performance for the happy class.

Anger, Happiness, Neutral State, Sadness, Fearful, and Surprised

Compared with the best-reported model on this emotion set, both models appear to demostrate an improvement, as seen in Table 6.3. However, when we consider the accuracy in each class, a significant disparity emerges.

As we can see in Table 6.4, there are significant improvements across angry, happy, sad, and neutral, however the fearful and surprise predictions are very inaccurate, with 1024-attn never predicting fearful or surprised. The data imbalance is a likely significant factor, as in the test set I have used there are only 18 examples of surprised and 10 examples of fearful. This would probably be improved if the examples of surprised and fearful could be upsampled.

Anger, Happiness, Neutral State, Sadness, Excited, and Frustrated

Considering both the overall and individual class performance, the models perform very similarly. The 1024-attn model shows much confusion with happiness and scores only 6.99% accuracy, which is significantly worse than both other models. However, it makes significant improvements in the excited accuracy. This confusion is likely between happy and excitement due to excited being much more common in the dataset.

Model	Modalities	Accuracy
DialogueGCN [31]	Text	**65.25%**
2-attn	Text + Audio	62.27%
1024-attn	Text + Audio	64.00%

Table 6.5: Table comparing overall performance with state-of-art model. Best performances are highlighted in bold.

Model	Anger Acc.	Happy Acc.	Neutral Acc.	Sad Acc.	Excited Acc.	Frustrated Acc.
DialogueGCN [31]	**67.53%**	**40.62%**	61.92%	**89.14%**	65.46%	64.18%
2-attn	57.06%	39.86%	**78.13%**	82.86%	51.84%	51.97%
1024-attn	50.59%	6.99%	76.56%	68.16%	**78.93%**	**64.30%**

Table 6.6: Table comparing class performance with state-of-art models. Best performances are highlighted in bold.

Model	Anger F1	Happy F1	Neu. F1	Sad F1	Excited F1	Fru. F1	Weighted Average F1
DialogueGCN [31]	**64.19**	**42.75**	63.54	**84.54**	63.08	**66.99**	**64.18**
2-attn	60.82	38.26	67.04	73.02	61.39	59.02	61.83
1024-attn	57.53	12.27	**67.74**	71.06	**72.28**	61.95	61.76

Table 6.7: Table comparing F1 class performance with state-of-art models. Best performances are highlighted in bold.

Model	Modalities	Accuracy
2-attn	Text + Audio	61.2%
1024-attn	Text + Audio	**62.9%**

Table 6.8: Table comparing accuracy over all trained emotions. Best performances are highlighted in bold.

While DialogueGCN slightly outperforms both 2-attn and 1024-attn, it requires pre-processing of the full audio/conversation to make its predictions, because it must first construct a graph where each utterance is represented as a node [31]. Both 2-attn and 1024-attn do not require any pre-processing of data and would be compatible with live, real-time audio. Given audio for an utterance and the transcript for this audio, which could be generated with Google Cloud, 2-attn and 1024-attn can return a prediction for this utterance and hidden state, which can be tracked throughout any potential AI therapy session or phone interview.

6.1.2 Full Results

Accuracy and F1 Scores

We can see in Table 6.8 that the 1024-attn model is slightly more accurate than the 2-attn model. But this is likely due to the poor balancing of the data. Looking at accuracy of individual classes (Table 6.9) we can see that the way 1024-attn achieves this increase in accuracy is by massively increasing the number of predictions of excited and frustrated at the detriment of happy, angry, and sad classes.

If we recall Figure 3.1, frustrated is the most over-represented class in the dataset, and additionally excited is much more common than happy, and it raises suspicions that the 1024-attn model has not learned as much about the emotions as the 2-attn model. With this in mind, it is possible to conclude that using a single attention value for each modality in the fusion (thereby forcing the model to consider all features in the vector to decide importance), actually improves its ability to learn multiple emotions, rather than learning individual emotional features.

There is much confusion between happy and excited, and similarly between sad, angry, and frustrated. By giving the model the freedom to pick combinations of attentions, the model seems to be detecting small indicators of the most common emotion and predicting this. For example, happy and excited both have a lot in common. By allowing the model to choose which features are important it may learn to only pay attention to features that are shared with happy and excited. At each point in the vector, we have one audio feature and one text feature. One can hypothesize that the model is paying attention to whichever of these two features is common in excited utterances due to the weighting of excited in the dataset. This attention results in features that differentiate happy and excited being ignored in favour of features that happy and excited share. The restriction imposed by listening to the entire modalities prevents the dataset overfitting to the larger classes in the

Model	Anger Acc.	Happy Acc.	Neu. Acc.	Sad Acc.	Excited Acc.	Fru. Acc.	Sur. Acc.	Fear Acc.
2-attn	**57.1%**	**39.9%**	**77.9%**	**82.9%**	51.8%	52.0%	**5.6%**	**0.0%**
1024-attn	50.6%	7.0%	76.6%	68.2%	**78.9%**	**64.3%**	0.0%	**0.0%**

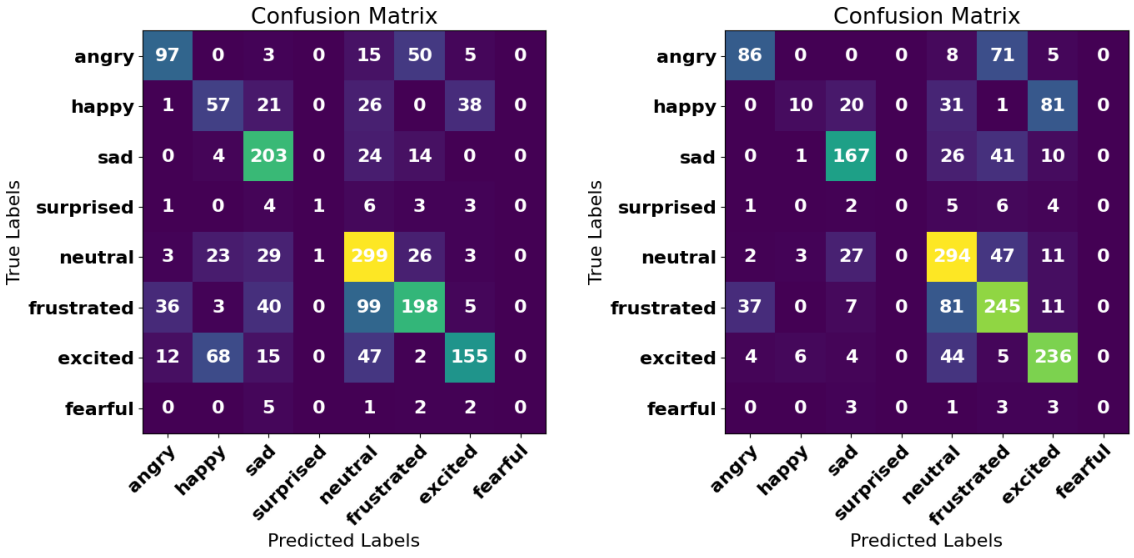
Table 6.9: Table comparing class performance over all trained emotions. Best performances are highlighted in bold.

Model	Anger F1	Happy F1	Neu. F1	Sad F1	Excited F1	Fru. F1	Sur. F1	Fear F1	Weighted Avg. F1
2-attn	**60.6**	**38.3**	66.4	**71.9**	60.8	58.6	**5.6**	**0.0**	**60.3**
1024-attn	57.3	12.3	**67.3**	70.3	**71.5**	**61.3**	0.0	**0.0**	60.2

Table 6.10: Table comparing F1 class performance over all trained emotions. Best performances are highlighted in bold.

dataset. We can see further evidence of this in the F1 scores in Table 6.10.

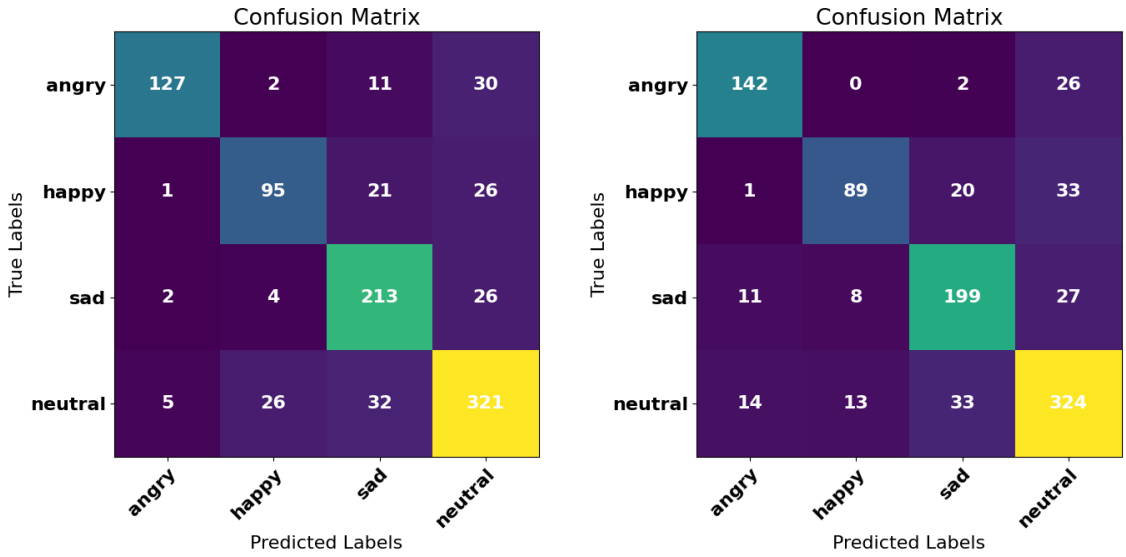
Confusion Matrix



(a) 2-attn **(b)** 1024-attn

Figure 6.1: Confusion Matrix over all trained emotions.

We can see a lot of confusion among the classes with very little weight in the original dataset when we consider the confusion matrix across all classes (Figure 6.1). Additionally, from this matrix, it is clear that the 1024-attn model is predicting the most common classes more, rather than learning something emotionally meaningful.



(a) `2-attn` **(b)** `1024-attn`

Figure 6.2: Confusion Matrix over angry, happy, sad, and neutral.

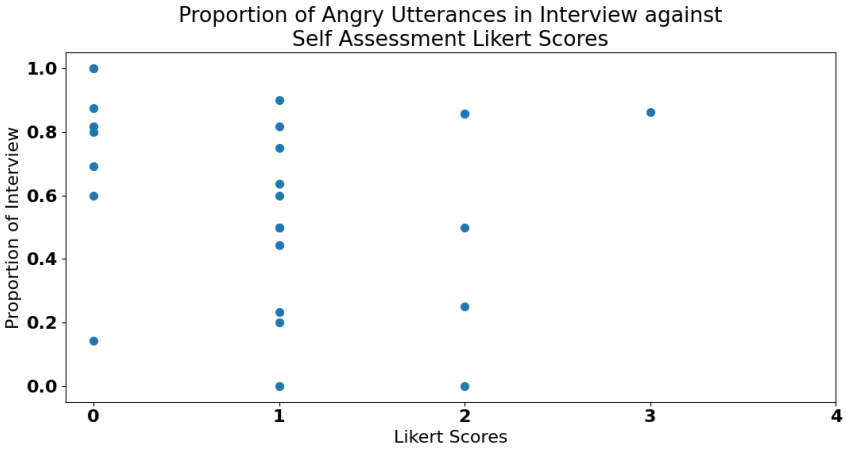
When we reduce the confusion matrix to only angry, happy, sad, and neutral, we see that the confusion is massively reduced and the model is very accurate (Figure 6.2). As discussed earlier, we reduce our model's prediction to a subset of the emotions by predicting all trained emotions and then only using the predictions for the emotion subset we are predicting. The performance is remarkably similar in this case. It indicates that the `1024-attn` model usually predicts happy as the second-highest probability after excitement.

6.1.3 COVID-19 Interviews

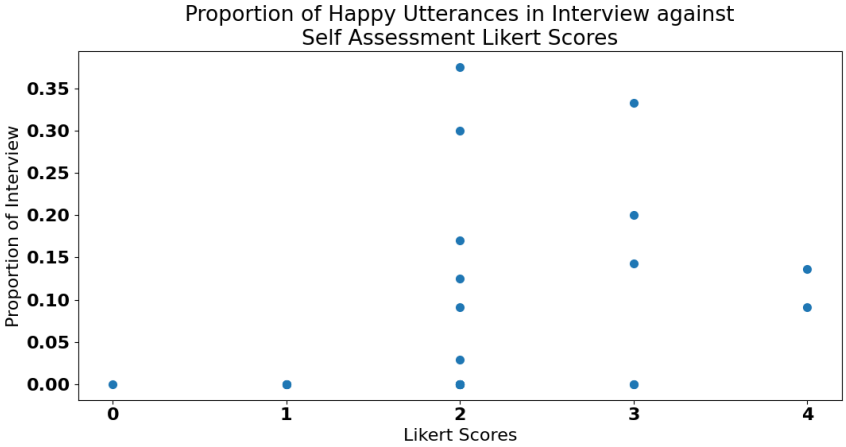
The collected COVID-19 interviews are a small corpus of data, designed in the hope of eliciting core emotions. We have both self-assessed and clinician-assessed Likert scores which rate how strong emotions are felt throughout the interviews. We use these scores to evaluate the performance of the model.

In order to directly compare utterance level emotion predictions and Likert scale emotion ratings, we have opted to plot the proportion of utterances spoken in the interview against the Likert scores for each emotion rated. For example, if half the utterances are angry, then the y-value will be 0.5, and the x-value will be the Likert scale value.

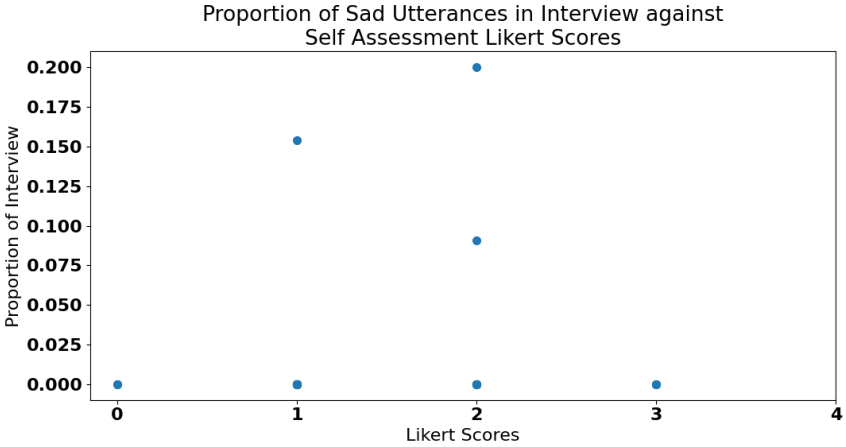
Other approaches were considered, such as using the probabilities of emotions to equate likelihood to emotion strength, or splitting the interviews into sections associated with the elicitation of different emotions. However, these approaches are not as simple to implement, and there is no guarantee that they would provide any real relationship between the Likert ratings.



(a) Angry

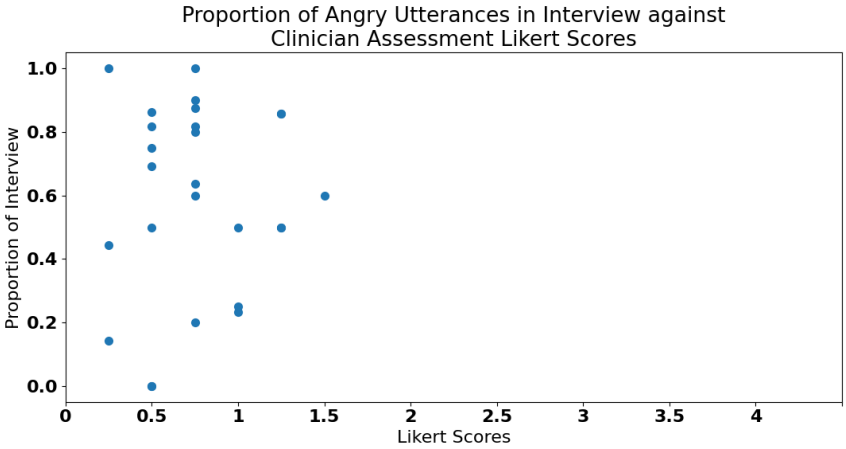


(b) Happy

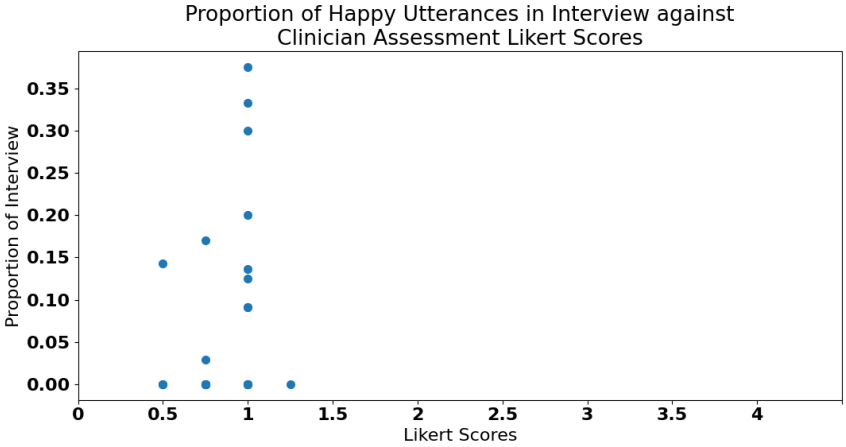


(c) Sad

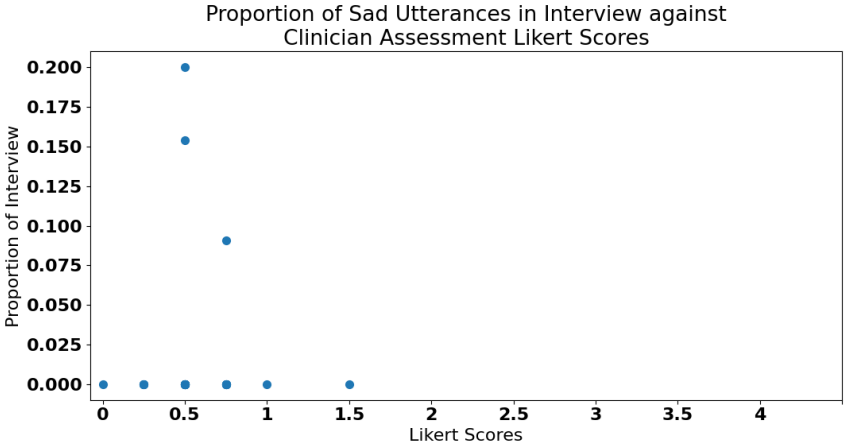
Figure 6.3: Proportion of Utterance Emotion Predictions against Self-assessed Likert Ratings.



(a) Angry



(b) Happy



(c) Sad

Figure 6.4: Proportion of Utterance Emotion Predictions against Clinician-assessed Likert Ratings. The clinician assessments are averages between two assessments.

The relationships in the graphs, both clinician-assessed and self-assessed, are not very strong. We hypothesize this is due to the automatic transcription of the interviews and the actual setting of the interviews. The format requires a single speaker to talk for a long time, and this does not make sense for utterance level emotion prediction.

A lot of utterances in this dataset are over 60 seconds in length, and emotions will fluctuate over that length of time. The proposed emotion recognition model supports long sequences; however, the results are less accurate, as shown by these interviews.

In future, it is important to consider the length of utterances or to manually transcribe interviews such that small pauses can split the utterances.

6.1.4 Limitations and Improvements

In an attempt to deal with the imbalance in the training data, weighting to the loss functions was provided in PyTorch which weights the contribution of different classes during training. The weight applied to each class is defined as $\frac{\text{size of class}}{\text{size of largest class}}$. This solution did not mitigate the issues with the class imbalance. Thus in future, using upsampling methods or supplementing the smaller classes with data from other datasets may be more fruitful.

The reported results are reproducible, evidenced by the similar performance between `2-attn` and `1024-attn`. However, in order to rigorously prove that the model is reproducible and consistently better than existing models, one should either perform 10-fold validation on the ten speakers or 5-fold validation on the five sessions in the dataset. Then, one could take the average performance across all of these runs as the model performance.

The size of the attention vector seems to make a significant difference to the confusion of the model. Further investigation into the meaning of the features, and exactly why extra freedom in the attention causes a decrease in average class performance, would be fascinating.

Due to the use of BERT-large and the extensive feature maps in the audio part of the model, the memory usage during training is significant. While the memory usage seems to be fine when evaluating the network, there needs to be evaluation into how heavy the model is. If there is an intention to deploy this model onto mobile phones, memory may be incredibly valuable. Similarly running this on devices with weak GPUs or no CUDA compatible GPU may impact the speed. In future, it would be useful to benchmark the speed and memory usage of the model during evaluation. In future, it may be useful to investigate how much impact smaller NLP models, such as DistilBERT, have on the performance of the model and the overall memory usage.

An improvement to the training part of the model would be a modification to how we remove data from the training set. At the moment, utterances are removed if

there is no majority agreement of the human evaluators; and in addition, the minority votes are ignored. Training on the confusion between the evaluators may learn a notable relationship between emotions. A good example is between happy and excited. If two evaluators determine the audio to be excited, and one thinks it is happy, the current model only learns that this audio is excited. However, learning that it is two-thirds excited and one-third happy may be more useful and reduce confusion between happy and excited. Additionally, since memory between utterances is an essential part of the model, removing utterances from the dataset is potentially hindering the model.

Fine-tuning the hyperparameters and perhaps investigating which optimiser to use could improve the model. Implementing some kind of cross-validation would probably result in a better and more consistent model.

6.1.5 Strengths

The model achieves state-of-art performance on some subsets of emotion prediction, and when considering models that can be performed live, it achieves state-of-art performance across the board. With data imbalance mitigation, there is evidence to suggest this model will perform even better.

Attention fusion of modalities significantly outperforms concatenation based late fusion and early fusion. Additionally, by restricting how much control attention has and forcing it to choose the entire modality to pay attention to rather, than allowing attention to be calculated on a per attention basis, we find that there is less confusion in the model.

The modified LSTM improves speaker-independent accuracy and enables modelling the fluctuations in emotion without the need to process every utterance and then use an LSTM on this. The modelling introduced will only represent left to right dependency of emotion states. Additionally, memory on the emotion hidden state allows modelling the speaker's emotional state as learned and extracted by the model.

Training on all the emotions is valid, even if the particular emotion results in bad performance. When we use the output of the model to predict a subset of emotions, accuracy is very high, and it generalises to any subset of the emotions the model is trained on.

6.2 Depression and Anxiety Prediction

6.2.1 Depression Model

A comparison against the established baselines and the winners from the AVEC 2019 challenge [23] can be seen in Table 6.11.

The poor performance in the depression model is likely due to utterance level features extracting bad features due to the sub-batch training method implemented to

Model	RMSE (Test set)	RMSE (Dev set)
AVEC 2019 Baseline (fusion) [23]	6.37	5.03
AVEC 2019 Runner-up [33]	5.50	4.94
AVEC 2019 Winner [32]	4.73 (Text only)	4.28
Proposed Model	6.51	5.70

Table 6.11: Table comparing the performance on PHQ 8 prediction between a multimodal baseline established as part of the AVEC 2019 challenge [23], and the two best performers of the challenge.

avoid exceeding memory constraints during training.

Future Work

The primary limitation of the depression prediction model is the utterance level feature extraction. As demonstrated by the emotion lexicon used in the runner-up model [33] emotional context of words and statements should have a significant impact on the prediction of depression. However, emotional understanding of utterances is not sufficient alone, and the model needs to be able to gradually learn other underlying features which can be combined with the emotional understanding.

A possible solution would be training a model on an utterance level of depression prediction, and then taking the final hidden state of this utterance level model as the feature extraction model in our proposed depression prediction model.

A different output layer could also improve the performance of the model, using a sigmoid activation before multiplying by 28 would ensure the model outputs in the range of PHQ 8 labels.

6.2.2 Anxiety Model

Anxiety prediction is very similar to the depression prediction. The RMSE on the test set is only 2.27, though this could be an optimistic score due to the limited range of GAD 7 scores in the dataset.

The anxiety model does show a relationship between predictions when we graph the predicted GAD 7 scores against the true scores (Figure 6.5). However, this relationship may be caused by limited results on the higher end of the GAD 7 scores and coincidence has caused this relationship.

Future Work

Using an underlying depression prediction model with better results before fine-tuning would likely yield more informative results as to how well anxiety can be predicted using transfer learning from a depression model.

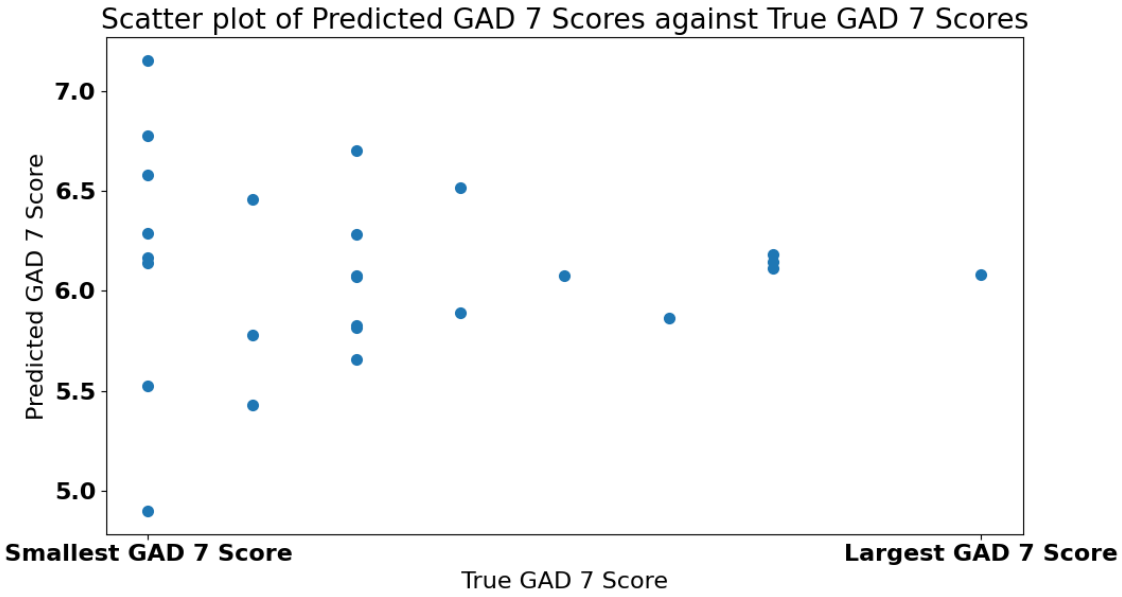


Figure 6.5: Predicted GAD 7 scores against true GAD 7 scores. True scores are hidden to ensure no re-identification of participants scores.

There seems to be evidence that information about anxiety can be learned using a model trained on depression. The average GAD 7 score in the training set is 2.6, with a median of 1, and the average GAD 7 prediction on the test set is 6.10, which is significantly different to the training set. There is evidence of overfitting after based on the RMSE graph (Figure 5.4), but this does not happen until a significant level of training is reached. With a small training set, fine-tuning is essential. Training for a long time will very rapidly overfit to the small dataset. With this in mind, further work with a better depression model should yield positive results with regards to the prediction of anxiety despite limited data.

Chapter 7

Conclusions and Future Work

This project tackles some problems associated with fully autonomous and remote therapy. For VR therapy applications, such as self-attachment therapy, accurate emotion recognition with the capacity for live evaluation is necessary. During remote auditory assessments, or when facial expressions are obscured by masks, audio-based emotion recognition during mental health consultations provides an additional useful tool to assist with people who may have emotional difficulties during remote consultations.

The proposed multi-task emotion recognition model introduces speaker-level dependency modelling without data pre-processing. Using a modified LSTM applied to the emotion prediction hidden state, we model the long-term memory of left-to-right self-dependency on the speaker's emotions and introduce emotion context from previous utterances. By doing this instead of first extracting features from all utterances and then learning the context between these features, we can make predictions per utterance. This difference allows emotion recognition to take place without the need for a conversation to conclude. This approach significantly outperforms other state-of-the-art models on the recognition of angry, happy, sad, and neutral emotions.

We found that attention-based fusion of audio and text modalities to be significantly better than concatenation based methods. By restricting the attention choice to be a binary choice as opposed to allowing attention to apply across all features in each modality, the confusion is reduced, and attention on the audio modality is increased.

Training on all the emotions available in the dataset and making predictions on a subset by only considering the logits from this subset produces considerable results without the need to retrain on each subset of emotions. This allows a heavier model, with significant memory requirements during training, to be trained once on all available emotion classes and be applied seamlessly to different applications with different requirements for emotion prediction.

It would be worth looking into improving the training of the model. Since we rely

on left-to-right speaker dependency, removing utterances from the IEMOCAP dataset during training is not desirable. Training in a multi-label fashion may produce better results as more items in the dataset would be considered, since human evaluator consensus would not be required. Additionally, human confusion can be learned. Human evaluators can confuse similar emotions, such as anger and frustration, or happiness and excitement. Rather than learning an input mapping to the majority decision of the evaluators, we could learn two thirds towards the majority, and one third towards the minority emotion.

Additionally, upsampling the smaller emotion classes in IEMOCAP would almost certainly improve the quality of the model, especially when predicting larger subsets of emotions. Supplementing these classes with another emotional dataset may provide even better results than upsampling methods.

An issue highlighted shows that longer utterances become much harder to classify correctly since a range of emotions may be experienced and cannot have a single label applied to them. Applications of the model should keep this in mind, and further work could be done to look into breaking these utterances down during emotion prediction, such that the model can return a range of more accurate emotions for a single long utterance input.

Future VR therapy applications, in particular self-attachment therapy, and AI therapists should make use of automatic emotion recognition tools, and we provide a powerful, live, real-time emotion recognition tool for use in these settings. Also, further work into the depression and anxiety prediction models could be extremely beneficial to therapy based applications.

Bibliography

[1] Organization WH. Depression and other common mental disorders: global health estimates; 2017. pages 1

[2] Torales J, O'Higgins M, Castaldelli-Maia JM, Ventriglio A. The outbreak of COVID-19 coronavirus and its impact on global mental health. International Journal of Social Psychiatry. 2020;66(4):317–320. Available from: https://doi.org/10.1177/0020764020915212. pages 1

[3] Cittern D, Edalat A, Ghaznavi I. An immersive virtual reality mobile platform for self-attachment. AISB;. . pages 1, 3

[4] Edalat A. Self-attachment: A holistic approach to Computational Psychiatry. In: Computational neurology and psychiatry. Springer; 2017. p. 273–314. pages 3

[5] Edalat A. Self-attachment: A new and integrative psychotherapy. In: Talk at the Institute of Psychiatry; 2013. p. 2–5. pages 3

[6] Ghaznavi I, Jehanzeb U, Edalat A, Gillies D. Usability evaluation of an immersive virtual reality platform for self-attachment psychotherapy. 2019;. pages 3, 4

[7] Rizzo A, Parsons TD, Lange B, Kenny P, Buckwalter JG, Rothbaum B, et al. Virtual Reality Goes to War: A Brief Review of the Future of Military Behavioral Healthcare. Journal of Clinical Psychology in Medical Settings. 2011 jun;18(2):176–187. Available from: https://doi.org/10.1007/s10880-011-9247-2. pages 4

[8] Rothbaum BO, Rizzo Aifede J. Virtual reality exposure therapy for combat-related posttraumatic stress disorder. Annals of the New York Academy of Sciences. 2010;1208(1):126–132. Available from: https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.2010.05691.x. pages 4

[9] Osimo SA, Pizarro R, Spanlang B, Slater M. Conversations between self and self as Sigmund Freud - A virtual body ownership paradigm for self counselling. Scientific Reports. 2015;5(1):13899. Available from: https://doi.org/10.1038/srep13899. pages 4

[10] Falconer CJ, Rovira A, King JA, Gilbert P, Antley A, Fearon P, et al. Embodying self-compassion within virtual reality and its effects on patients with depression. BJPsych open. 2016 jan;2(1):74–80. pages 4

[11] Freeman D, Bradley J, Antley A, Bourke E, DeWeever N, Evans N, et al. Virtual reality in the treatment of persecutory delusions: Randomised controlled experimental study testing how to reduce delusional conviction. British Journal of Psychiatry. 2016 jul;209(1):62–67. Available from: https://www.cambridge.org/core/product/identifier/S0007125000244371/type/journal_article. pages 4

[12] Wing JK, Babor T, Brugha T, Burke J, Cooper JE, Giel R, et al. SCAN: Schedules for Clinical Assessment in Neuropsychiatry. Archives of General Psychiatry. 1990 jun;47(6):589–593. Available from: https://doi.org/10.1001/archpsyc.1990.01810180089012. pages 5

[13] Deshpande M. Perceptrons: The First Neural Networks. 2017;Available from: https://pythonmachinelearning.pro/perceptrons-the-first-neural-networks/. pages 6

[14] Saha S. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way; 2018. Available from: https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a5 pages 6

[15] Tzeng E, Hoffman J, Saenko K, Darrell T. Adversarial Discriminative Domain Adaptation. CoRR. 2017;abs/1702.0. Available from: http://arxiv.org/abs/1702.05464. pages 8

[16] Luo Z, Zou Y, Hoffman J, Fei-Fei L. Label Efficient Learning of Transferable Representations across Domains and Tasks; 2017. pages 8

[17] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 2017-Decem; 2017. p. 5999–6009. pages 9, 16

[18] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018;Available from: https://arxiv.org/abs/1810.04805. pages 9, 10, 30, 31

[19] Luo L, Wang Y. EmotionX-HSU: Adopting Pre-trained BERT for Emotion Classification. CoRR. 2019;abs/1907.0. Available from: http://arxiv.org/abs/1907.09669. pages 10, 30

[20] Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation. 2008;42(4):335. pages 10, 18, 19, 21, 26, 45

[21] Parada-Cabaleiro E, Costantini G, Batliner A, Schmitt M, Schuller BW. DEMoS: an Italian emotional speech corpus. Language Resources and Evaluation. 2019;Available from: https://doi.org/10.1007/s10579-019-09450-y. pages 10, 21

[22] Gratch J, Artstein R, Lucas G, Stratou G, Scherer S, Nazarian A, et al. The Distress Analysis Interview Corpus of human and computer interviews. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik, Iceland: LREC; 2014. p. 3123–3128. Available from: http://ict.usc.edu/pubs/TheDistressAnalysisInterviewCorpusofhumanandcomputerinterviews.pdf. pages 11, 23, 48

[23] Ringeval F, Schuller B, Valstar M, Cummins N, Cowie R, Tavabi L, et al. AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition. In: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop; 2019. p. 3–12. pages 11, 27, 48, 69, 70

[24] Zhang Z, Wu B, Schuller BW. Attention-Augmented End-to-End Multi-Task Learning for Emotion Prediction from Speech. CoRR. 2019;abs/1903.1. Available from: http://arxiv.org/abs/1903.12424. pages 11, 27, 38

[25] Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller BW, Zafeiriou S. End-to-End Multimodal Emotion Recognition using Deep Neural Networks. CoRR. 2017;abs/1704.0. Available from: http://arxiv.org/abs/1704.08619. pages 12, 35

[26] Tripathi S, Beigi HSM. Multi-Modal Emotion recognition on IEMOCAP Dataset using Deep Learning. CoRR. 2018;abs/1804.0. Available from: http://arxiv.org/abs/1804.05788. pages 12, 21, 34, 60

[27] Tripathi S, Kumar A, Ramesh A, Singh C, Yenigalla P. Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions. 2019 jun;Available from: http://arxiv.org/abs/1906.05681. pages 13, 34, 38, 60

[28] Yoon S, Byun S, Jung K. Multimodal Speech Emotion Recognition Using Audio and Text. In: 2018 IEEE Spoken Language Technology Workshop, SLT 2018 - Proceedings. Institute of Electrical and Electronics Engineers Inc.; 2019. p. 112–118. Available from: http://arxiv.org/abs/1810.04635. pages 14, 34, 35, 60

[29] Sahu G. Multimodal Speech Emotion Recognition and Ambiguity Resolution. 2019 apr;Available from: http://arxiv.org/abs/1904.06022. pages 14, 61

[30] Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-Based Systems. 2018 jun;161:124–133. Available from: http://arxiv.org/abs/1806.06228. pages 14, 15, 26, 35, 38, 60

[31] Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In: EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. Association for Computational Linguistics; 2020. p. 154–164. Available from: http://arxiv.org/abs/1908.11540. pages 15, 26, 38, 39, 62, 63

[32] Ray A, Kumar S, Reddy R, Mukherjee P, Garg R. Multi-level attention network using text, audio and video for depression prediction. In: AVEC 2019 - Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop, co-located with MM 2019. Association for Computing Machinery, Inc; 2019. p. 81–88. Available from: http://arxiv.org/abs/1909.01417. pages 15, 16, 26, 35, 70

[33] Yin S, Liang C, Ding H, Wang S. A multi-modal hierarchical recurrent neural network for depression detection. In: AVEC 2019 - Proceedings of the 9th International Audio/Visual Emotion Challenge and Workshop, co-located with MM 2019. New York, New York, USA: Association for Computing Machinery, Inc; 2019. p. 65–71. Available from: http://dl.acm.org/citation.cfm?doid=3347320.3357696. pages 16, 70

[34] Ren Z, Kong Q, Han J, Plumbley MD, Schuller BW. Attention-based Atrous Convolutional Neural Networks: Visualisation and Understanding Perspectives of Acoustic Scenes. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019. p. 56–60. pages 16

[35] Rizos G, Hemker K, Schuller B. Augment to Prevent: Short-Text Data Augmentation in Deep Learning for Hate-Speech Classification. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. CIKM '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 991–1000. Available from: https://doi.org/10.1145/3357384.3358040. pages 17

[36] Martin M. On the induction of mood. Clinical Psychology Review. 1990;10(6):669–697. Available from: http://www.sciencedirect.com/science/article/pii/027273589090075L. pages 21

[37] Pintelas E, Kotsilieris T, Livieris I, Pintelas P. A review of machine learning prediction methods for anxiety disorders; 2018. . pages 23

[38] Poria S, Majumder N, Hazarika D, Cambria E, Gelbukh A, Hussain A. Multimodal Sentiment Analysis: Addressing Key Issues and Setting Up the Baselines. IEEE Intelligent Systems. 2018 nov;33(6):17–25. pages 26

[39] Rizos G, Baird A, Elliott M, Schuller B. Stargan for Emotional Speech Conversion: Validated by Data Augmentation of End-To-End Emotion Recognition. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2020. p. 3502–3506. pages 27

[40] Greff K, Srivastava RK, Koutnik J, Steunebrink BR, Schmidhuber J. LSTM: A Search Space Odyssey. IEEE Transactions on Neural Networks and Learning Systems. 2017 oct;28(10):2222–2232. pages 27

[41] Breuel TM. Benchmarking of LSTM Networks. 2015 aug;Available from: http://arxiv.org/abs/1508.02774. pages 27

[42] Latif S, Rana R, Qadir J, Epps J. Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study. 2017 dec;Available from: http://arxiv.org/abs/1712.08708. pages 30

[43] Heusser V, Freymuth N, Constantin S, Waibel A. Bimodal Speech Emotion Recognition Using Pre-Trained Language Models. 2019 nov;Available from: http://arxiv.org/abs/1912.02610. pages 30

[44] Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019 jul;Available from: http://arxiv.org/abs/1907.11692. pages 31

[45] Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019 oct;Available from: http://arxiv.org/abs/1910.01108. pages 31

[46] Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency LP. Context-Dependent Sentiment Analysis in User-Generated Videos. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada: Association for Computational Linguistics; 2017. p. 873–883. Available from: https://www.aclweb.org/anthology/P17-1081. pages 60