Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG

# Multi-Modal Representation Learning for Emotion Recognition in Continuous Domain

**Master thesis**
im Arbeitsbereich Knowledge Technology, WTM
Prof. Dr. Stefan Wermter

Department Informatik
MIN-Fakultät
Universität Hamburg

vorgelegt von
**Navneet Singh Arora**
am
26.03.2023

Gutachter: Prof. Dr. Stefan Wermter
Dr. Di Fu
Dr. Hugo Cesar de Castro Carneiro

Navneet Singh Arora
Matrikelnummer: 7374546
Rauschener Ring 26B
22047 Hamburg

# Abstract

Multi-modal emotion recognition is of interest for human-robot interaction, given its relation to how humans express emotional cues across various modalities and its complex nature. Two widely applied approaches for emotion recognition involve identification through (i) categories (e.g., anger, happiness, sadness, etc.) or through (ii) dimensions (i.e., the continuous state through valence and arousal). However, predicting the continuous state of emotion has remained challenging despite the recent progress in the field of emotion recognition. Most multi-modal approaches leverage the audio-visual (AV) features in a highly correlated form, creating an implicit bias while doing feature learning. Additionally, current methods focus on data-hungry supervised learning requiring annotated data collection. In this proposed study, we present a multi-modal dimensional emotion recognition approach to leverage salient AV features by keeping these modalities isolated from each other. We curate two different models, one for the audio and the other for the visual aspect, training them using the TED-LIUM corpus (dataset for speech recognition in English) and FaceSynthetics (a synthetic dataset for facial landmark embeddings), respectively. Finally, we use IEMOCAP (dyadic conversation-based audio-visual dataset) to evaluate the models for dimensional emotion recognition. Using an iterative pre-training approach by training the encoder with non-contrastive learning and contrastive learning, the model achieves a mean performance gain of 11% across the three dimensions of valence, arousal and dominance over previous state-of-the-art models while being on par with the current best model. Furthermore, the study reveals a strong correlation between speech and arousal and a positive correlation between valence and visual aspects. Finally, we discuss the effectiveness of synthetic data, its challenges and future work. Code for all the models and experiments is available at: https://git.informatik.uni-hamburg.de/0arora/audio_visual_emotion_recognition.

# Zusammenfassung

Die multimodale Erkennung von Emotionen ist für die Mensch-Roboter-Interaktion von Interesse, da sie mit der Art und Weise zusammenhängt, wie Menschen ihre Emotionen über verschiedene Modalitäten ausdrücken, und weil sie sehr komplex ist. Zwei weit verbreitete Ansätze zur Erkennung von Emotionen beinhalten die Identifizierung anhand von (i) Kategorien (z. B. Wut, Freude, Traurigkeit usw.) oder anhand von (ii) Dimensionen (d. h. der kontinuierliche Zustand anhand von Valenz und Erregung). Die Vorhersage des kontinuierlichen Zustands von Emotionen ist jedoch trotz der jüngsten Fortschritte auf dem Gebiet der Emotionserkennung eine Herausforderung geblieben. Die meisten multimodalen Ansätze nutzen die audiovisuellen (AV) Merkmale in einer stark korrelierten Form, was zu einer impliziten Verzerrung beim Lernen der Merkmale führt. Darüber hinaus konzentrieren sich die derzeitigen Methoden auf datenintensives überwachtes Lernen, das eine kommentierte Datensammlung erfordert. In dieser Studie stellen wir einen multimodalen dimensionalen Emotionserkennungsansatz vor, der auffällige AV-Merkmale nutzt, indem diese Modalitäten voneinander isoliert werden. Wir erstellen zwei verschiedene Modelle, eines für den Audio- und eines für den visuellen Aspekt, und trainieren sie mit dem TED-LIUM-Korpus (Datensatz für die Spracherkennung in Englisch) bzw. FaceSynthetics (ein synthetischer Datensatz für die Einbettung von Gesichtsmerkmalen). Schließlich verwenden wir IEMOCAP (dyadischer, konversationsbasierter audiovisueller Datensatz), um die Modelle zur dimensionalen Emotionserkennung zu evaluieren. Unter Verwendung eines iterativen Pre-Training-Ansatzes, bei dem der Encoder mit nicht-kontrastivem Lernen und kontrastivem Lernen trainiert wird, erreicht das Modell einen durchschnittlichen Leistungsgewinn von 11% über die drei Dimensionen Valenz, Erregung und Dominanz im Vergleich zu früheren State-of-the-Art-Modellen, während es mit dem derzeit besten Modell gleichzieht. Darüber hinaus zeigt die Studie eine starke Korrelation zwischen Sprache und Erregung und eine positive Korrelation zwischen Valenz und visuellen Aspekten. Abschließend diskutieren wir die Effektivität synthetischer Daten, die Herausforderungen und die zukünftige Arbeit. Der Code für alle Modelle und Experimente ist verfügbar unter: https://git.informatik.uni-hamburg.de/0arora/audio_visual_emotion_recognition.

# Contents


**1 Introduction** **1**
- 1.1 Motivation . . . 3
- 1.2 Approach . . . 4
- 1.3 Research Question(s) . . . 5
- 1.4 Thesis Outline . . . 6

**2 Related Work** **7**
- 2.1 Speech Emotion Recognition (SER) . . . 7
- 2.2 Visual Emotion Recognition (VER) . . . 9
- 2.3 Synthetic Data . . . 10

**3 Background Information** **11**
- 3.1 Emotions . . . 11
  - 3.1.1 Discrete Emotions . . . 11
  - 3.1.2 Dimensional Emotions . . . 12
- 3.2 Learning Methodologies . . . 13
  - 3.2.1 Self-Supervised Learning (SSL) . . . 14

**4 Dimensional Emotion Recognition** **17**
- 4.1 Speech Model - NCL Architecture . . . 17
  - 4.1.1 Attention-based Speech Encoder . . . 18
  - 4.1.2 Speech Projection Head . . . 19
  - 4.1.3 Speech Prediction Head . . . 20
  - 4.1.4 Speech Model - (NCL + CL) Architecture . . . 21
  - 4.1.5 Speech Network Fine-Tuning Architecture . . . 22
- 4.2 Visual Model . . . 23
  - 4.2.1 Mutual Contrastive Learning (MCL) Based Architecture . . . 23
  - 4.2.2 Visual Network Intermediate Training Architecture . . . 26
  - 4.2.3 Visual Network Fine-Tuning Architecture . . . 27

**5 Experimentation and Results** **29**
- 5.1 Datasets . . . 29
  - 5.1.1 Speech Datasets . . . 29
  - 5.1.2 IEMOCAP Dataset . . . 30
  - 5.1.3 FaceSynthetics Dataset . . . 31

5.2 Data Preparation . . . 32
5.2.1 Speech Data Pre-processing . . . 32
5.2.2 Speech Data Augmentation . . . 33
5.2.3 Visual Data Augmentation . . . 34
5.2.4 Audio and Visual (AV) Data Pre-processing . . . 34
5.2.5 Audio and Visual (AV) Data Post-processing . . . 35
5.3 Loss Functions . . . 36
5.3.1 Negative Cosine Similarity Loss . . . 36
5.3.2 Cross-Entropy Loss . . . 36
5.3.3 Mean Absolute Error (MAE) Loss . . . 37
5.3.4 Mean Squared Error (MSE) Loss . . . 37
5.3.5 Concordance Correlation Coefficient (CCC) Loss . . . 37
5.3.6 Triplet Loss . . . 38
5.3.7 InfoNCE Loss . . . 39
5.4 Training Setup . . . 40
5.4.1 Model Hyperparameters . . . 40
5.4.2 Evaluation Metric . . . 42
5.5 Results . . . 43
5.5.1 Contrastive over Non-Contrastively learnt Representations . . . 43
5.5.2 Comparison with SOTA . . . 45
5.6 Ablation Study . . . 46
5.6.1 Impact of Batch Size . . . 46
5.6.2 Impact of Gradient Accumulator . . . 47
5.6.3 Impact of equally weighted CCC Loss . . . 48

**6 Discussion** **49**
6.1 Dimensional Emotion Representation . . . 49
6.1.1 Training Results . . . 49
6.2 Emotional Dimension Correlation . . . 50
6.3 Synthetic Data - The Solution? . . . 51
6.4 Iterative Pre-Training . . . 51

**7 Conclusion** **52**

**Bibliography** **53**

**Appendix** **69**

# List of Figures


1.1 A 2D emotional space for discrete emotions showcasing valence and arousal [32, 86]. . . . 2
1.2 Step-by-step approach for dimensional emotion analysis using speech model. . . . 4
1.3 Step-by-step approach for dimensional emotion analysis using visual model. . . . 4

3.1 Siamese Network based architecture comparison [20]. . . . 14
3.2 SimSiam network with stop-gradient technique in one of the sub-networks [20]. . . . 15

4.1 Architectural overview of the network used for representation learning in the speech domain. The architecture consists of three main components, namely: (a) *Speech Encoder*, (b) *Speech Projection Head* and (c) *Speech Prediction Head*. The speech encoder is used to capture the representations, while both the speech projection head and speech prediction head are important components that help in optimising the representations learnt through the network. . . . 18
4.2 NCL speech model training process along with the architectural overview. As shown, an input waveform $X_s$ is passed through a stochastic data augmentation module $d_s(\cdot)$ forming a pair of augmented waveforms, passed onto the respective network branch for feature learning. . . . 20
4.3 CL training process along with the architectural overview. The process involves pre-processing the data using the EmoBERTa to create the required pseudo-labels for creating negative samples. These negative samples along with the anchor and the positive samples are processed respectively for improved representations. . . . 21
4.4 Fine-tuning process along with the architectural overview for the speech network using the CCC loss function. . . . 22
4.5 The figure exhibits the MCL . . . 24
4.6 Overview of the facial landmark recognition process and architecture using the pre-trained encoder network and cross-entropy loss. . . . 26

4.7 Architectural overview of the visual model used to fine-tune the pre-trained encoder embeddings for the downstream task. The network incorporates multiple layers of uni-directional and bi-directional Long Short-Term Memory (LSTM)'s to capture sequential dependencies. 27

5.1 Sample images from FaceSynthetics Dataset [128]. From top to bottom, the image shows different types of augmentation (a) Original Image, (b) Grayscale Image, (c) Noisy Image and (d) Brightness and Contrast formatted Image. . . . . 31
5.2 Mapping of 28 distinct emotions of GoEmotions dataset to 7 distinct emotions of IEMOCAP dataset. . . . . 32
5.3 IEMOCAP dataset pre-processing steps. Starting from the available conversations, each conversation is split into unique utterances. These utterances are further split based on the speaker to be further processed to extract the speaker-based keyframes. . . . . 34
5.4 2D attribute space for the dimensional emotions using the 4-way classifier. . . . . 47
5.5 2D attribute space for the dimensional emotions using the 6-way classifier. . . . . 47
5.6 (Left) NCL training loss plot with batch size = 128. (Right) NCL training loss plot with batch size = 32. . . . . 48

7.1 Sample images from FaceSynthetics Dataset [128]. From top to bottom, the image shows 6 different types of augmentation including the original image and the image with facial landmark points. . . . 69
7.2 Per-pixel semantic class segmentation mapping to the original synthesized image. . . . . 70

# List of Tables


5.1 TED-LIUM Corpus characteristics comparison across all three releases (v1 vs v2 vs v3 release). . . . . 30

5.2 Hyperparameter comparison across the three stages of the Speech Model Training. . . . . 41

5.3 Hyperparameter comparison across the three stages of the Visual Model Training. . . . . 42

5.4 Comparison of the two different approaches followed for learning speech representations. The analysis is based on the 4-way and 6-way emotion classifiers. The arrow indicator (⇑) represents better performance with a higher score. . . . . 43

5.5 Visual Model performance comparison on 4-way and 6-way emotion classifier. The arrow indicator (⇑) represents better performance with a higher score. . . . . 43

5.6 Comparison of ***Our*** model with the SOTA models in the visual or joint domain based on the CCC scores on the IEMOCAP Dataset. The arrow indicator (⇑) represents better performance with a higher score. . . . . 44

5.7 Comparison of ***Our*** models with the SOTA models in the speech domain based on the CCC scores on the IEMOCAP Dataset. The arrow indicator (⇑) represents better performance with a higher score. . . . . 46

6.1 Effect for penalising the dimension of the emotion equally as well as heavily within the loss function. . . . . 50

6.2 Effect for penalising the dimension of the emotion equally as well as heavily within the loss function. . . . . 50

# List of Abbreviations

**ABAW** Affective Behavior Analysis in-the-wild.

**AI** Artificial Intelligence.

**AV** Audio and Visual.

**BERT** Bidirectional Encoder Representations from Transformers.

**BN** Batch Normalisation.

**BYOL** Bootstrap Your Own Latent.

**CCC** Concordance Correlation Coefficient.

**CL** Contrastive Learning.

**CMN** Conversational Memory Network.

**CNN** Convolutional Neural Network.

**Conformer** Convolution-augmented Transformer.

**CPC** Contrastive Predictive Coding.

**CV** Computer Vision.

**DER** Dimensional Emotion Recognition.

**DERM** Dimensional Emotion Recognition Model.

**DL** Deep Learning.

**DNNs** Deep Neural Networks.

**EmoBERTa** Speaker-Aware Emotion Recognition in Conversation with RoBERTa.

**ER** Emotion Recognition.

**ERC** Emotion Recognition in Conversation.

**FER** Facial Emotion Recognition.

**GANs** Generative Adversarial Networks.

**GCN** Graph Convolution Network.

**GLU** Gated Linear Units.

**GNN** Graph Neural Network.

**GoEmotions** A Dataset for Fine-Grained Emotion Classification.

**HCI** Human-Computer Interaction.

**IANN** Interaction-Aware Attention Network.

**ICL** Interactive Contrastive Learning.

**IEMOCAP** The Interactive Emotional Dyadic Motion Capture.

**InfoNCE** Normalized Cross-Entropy.

**KD** Knowledge Distillation.

**KL** Kullback–Leibler Divergence.

**LOSO** Leave-One-Session-Out.

**LSTM** Long Short-Term Memory.

**MAE** Mean Absolute Error.

**MCL** Mutual Contrastive Learning.

**MFCCs** Mel-frequency Cepstral Coefficients.

**ML** Machine Learning.

**MLP** Multi-Layer Perceptron.

**MoCo** Momentum Contrast.

**MSE** Mean Squared Error.

**NCL** Non-Contrastive Learning.

**ReLU** Rectified Linear Unit.

**RL** Representation Learning.

**RoBERTa** Robustly optimised BERT.

**SC** Social Cognitive.

**SER** Speech Emotion Recognition.

**SGD** Stochastive Gradient Descent.

**SimSiam** Simple Contrastive Learning.

**SL** Supervised Learning.

**SOTA** State-of-the-Art.

**SSL** Self-Supervised Learning.

**SSNet** Single Stream Network.

**TL** Transfer Learning.

**UL** Unsupervised Learning.

**VAD** Valence-Arousal-Dominance.

**VCL** Vanilla Contrastive Learning.

**VER** Visual Emotion Recognition.

**WavAugment** A Time-domain Data Augmentation library.

**XLNet** Generalized Autoregressive Pre-training for Language Understanding.

# Chapter 1

# Introduction

Machine Learning (ML) and Artificial Intelligence (AI) paradigms have revolutionised over the past decade. With the introduction of Deep Learning (DL) approaches, the space of Human-Computer Interaction (HCI) has seen revolutionary progress in terms of performance. It has helped enable multiple real-world applications in this field [36, 40] to solve real-world problems. Most, if not all, State-of-the-Art (SOTA) AI systems use some Deep Neural Networks (DNNs) [46, 121]. Most of these systems train using large amounts of annotated data through Supervised Learning (SL), leading to logarithmic performance gain through annotated data set sizes [48, 119]. Dominant and confined to the field of Computer Vision (CV) [44], the investment and, thus, the availability of annotated data is finite in the language-related domains [88]. Furthermore, the cost of collecting and making the annotated data available is a significant and time-consuming barrier [33].

The lack of annotated data has led to increased attention towards Unsupervised Learning (UL) and majorly Self-Supervised Learning (SSL), where the learning takes place without the annotated data by exploiting the supervisory signals present within the data [67]. Recent successes [27, 105, 106] and some promising results [49, 118, 130] have motivated further research. Representation Learning (RL), a subset of SSL, has been the choice in recent years, leveraging the use of un-annotated data available in abundance to improve the efficiency of learning and quality of representations across modalities, including audio [11], visual [18, 37], and textual [28] domains.

The use of the SSL approach for learning effective Audio and Visual (AV) representations [9, 29, 87, 92, 93, 99] has enabled more human-like interactions between humans and computers [69]. Emotion Recognition (ER), being one of them, is a crucial research area for not only analysing human-to-human conversations but also for facilitating better HCI [123, 113]. It is a highly complex and challenging problem as humans emit and perceive emotions through gestures, facial expressions, body movements, and vocal tone [104], with the most expressive channels being the non-verbal modes of communication [115, 103]. Besides that, the conveyed emotions are very diverse across individuals and cultures [7, 103]. ER branches

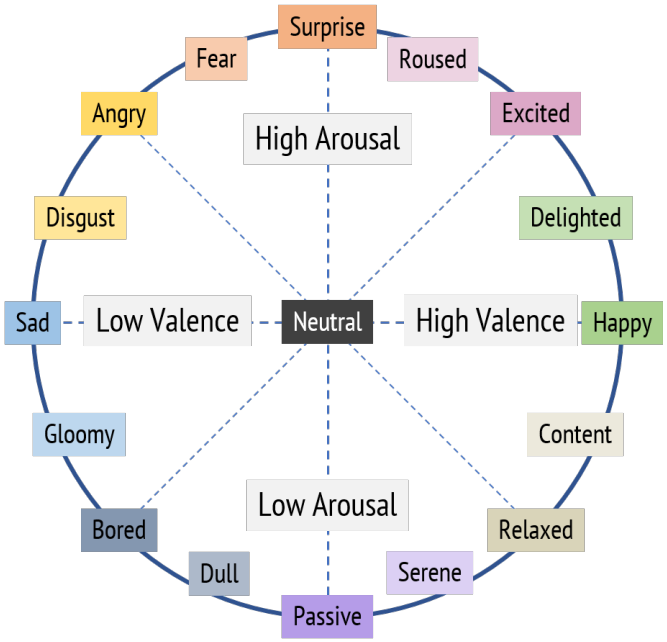


Figure 1.1: A 2D emotional space for discrete emotions showcasing valence and arousal [32, 86].

out in two different ways, (i) as a discrete classification problem (detection of discrete emotions, e.g., categorical emotions like happiness, sadness or anger), or (ii) as a continuous regression problem (predicting the distribution of the recognised emotion in a 3D space, e.g., dimensional emotions depicting continuous values of valence, arousal and dominance).

As shown in figure-1.1 [32, 86], valence and arousal, along with the third dimension of dominance, are widely used for estimating emotions in a continuous domain. Valence reflects pleasant to unpleasant responses, while arousal reflects the energy or intensity of those responses [104]. The third dimension of dominance (not represented in the figure) reflects the relative strength or loudness of the speech compared to background noise or other non-speech sounds. This continuity makes Dimensional Emotion Recognition (DER) challenging because it is harder to locate the emotions in these three dimensions of valence, arousal, and dominance than to predict the category due to the absence of temporal boundaries in human emotions [24].

Recent work on ER has shown substantial results, especially in the multi-modal domain. The McGurk effect [80, 91] suggests that visual information heavily influences the auditory experience. The findings of this strong association between the AV aspects (speech and facial features) [55, 13, 54, 59] led to the introduction of the Single Stream Network (SSNet) for shared latent space representation. Similarly, it gave rise to an AV fusion technique [79, 101, 103, 104], which jointly fuses the features from both modalities. Finally, Nagrani et al. [90] and Li et

al. [69] respectively show the self-supervised way of learning the cross-modal and speech embeddings. However, most learning techniques mentioned use clean and curated datasets [107]. Besides, most feature learning occurs through video frames with highly correlated AV feature sets. The words, phonemes and expressions are all in sync and aligned, forming a bias in the salient feature learning of the AV aspect. There is also a need for more variety regarding facial identities and vocal tones. Another aspect is the heavy reliance on large datasets, focusing on fully SL in already mentioned approaches [78]. It not only limits the learning capabilities, as creating such large datasets is expensive, but the downstream applications are also limited.

Hence, there is a need for approaches where smaller network architectures can learn quality representation using limited amounts of unannotated data. For instance, the Contrastive Predictive Coding (CPC) [94] technique can extract valuable representations from sequential data and achieves competitive performance on various tasks, including speaker classification and identification in speech [69]. On top of that, there is a need to put forward an alternative way of curating the datasets other than by capturing them in the real world, which, as already mentioned, is expensive and time-consuming. Creating synthetic data leads us towards that alternative solution, showing good generalising capabilities on real-world in-the-wild data [129]. In addition, the synthetic data has demonstrated great opportunities for learning embeddings which would be impossible to annotate through manual processes [129, 5].

## 1.1 Motivation

The main focus of this study is to evaluate the dimensions of discrete emotions when speech and facial information is not synchronised [62]. Thus, the work focuses on the perception of the emotion rather than its recognition. We base our analysis on considering the established achievements of SSL.

As dealing with raw audio is itself challenging, hence multiple recent works have taken the approach of converting the raw speech into some form of intermediate state like spectrograms [95], log-mel spectrograms [81, 25] or Mel-frequency Cepstral Coefficients (MFCCs) [42]. This intermediate state changes the approach and converts the problem of the speech domain into a problem in the visual domain. The primary task is to encode the emotional attributes in the speech without supervision, keeping the analysis constrained to the audio domain. The secondary task is to encode the facial emotional expressions using synthetic data and apply these encoded representations to real-world data for feasibility analysis. Finally, the aim is to focus on these tasks in an isolated manner while still being able to correlate the findings.

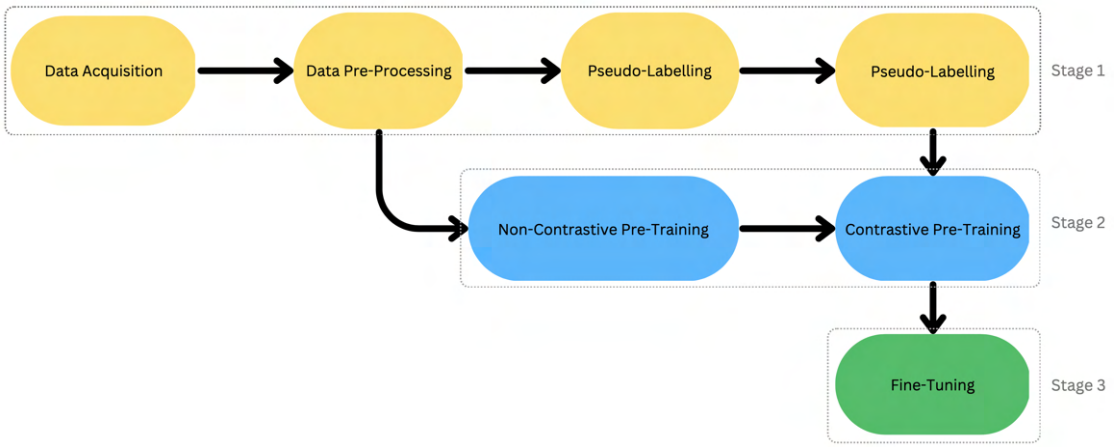


Figure 1.2: Step-by-step approach for dimensional emotion analysis using speech model.

With the motivation to find the dimensional emotion correlation in AV domains, this thesis aims to provide a valuable contribution towards understanding and improving the HCI.

## 1.2 Approach

The approach mainly consists of isolating the two modalities, vision and speech. The task involves learning the audio and visual feature representations in an unaligned manner, hence finding the association between the representations and the dimensions of the emotions as part of this study.

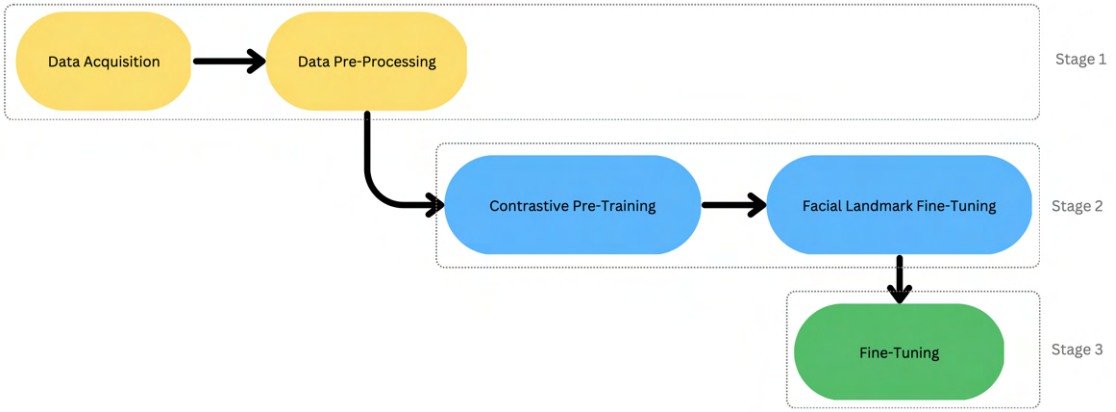


Figure 1.3: Step-by-step approach for dimensional emotion analysis using visual model.

The speech modality task is divided into three stages, as shown in figure-1.2. The first stage involves activities from data acquisition to data pre-processing.

Once the data pre-processing concludes, it passes as input into a pre-trained model to create pseudo-labels, a pre-requisite for the second stage. During the second stage, the execution of the pre-training of the model designed for learning speech representation takes place using the Non-Contrastive Learning (NCL) method through the pre-processed data without the pseudo-labels. The model is then further pre-trained through Contrastive Learning (CL), but this time with negative samples created using the pseudo-labels. Finally, the third and last stage involves fine-tuning the model encoder created in the previous stage on a cross-domain dataset for the downstream task of DER.

Similarly, for the visual modality (figure-1.3), the task is divided into three stages, where the first stage involves data acquisition and data pre-processing. Compared to the speech model approach, this stage of the visual model does not involve creating negative samples. Hence, no pseudo-labelling is required. So, the second stage directly involves pre-training through CL exploiting the synthetic data and data augmentation to learn representations. These representations are further trained with the facial landmarks for more robustness before fine-tuning them on a real dataset for the downstream task of DER in the third and final stage. Finally, the dimensions of the emotions are analysed through the respective representations to understand further the role each modality plays in understanding human emotions.

## 1.3 Research Question(s)

The research done through this thesis addresses four individual research questions as below.

1. ***Research Question 01***: Is the emotional dimension of "arousal" better represented through speech modality, whereas the visual modality correlates highly with the "valence" dimension of emotion?

2. ***Research Question 02***: Do emotional dimensions correlate with facial attributes like facial landmarks?

3. ***Research Question 03***: Can synthetic data help learn universal and generalisable facial features and bridge the gap with real-world data?

4. ***Research Question 04***: Can iterative pre-training of the models using unsupervised learning improve the performance on the downstream tasks?

To answer above mentioned questions, we propose a Multi-Modal Representation Learning for the Dimensional Emotion Recognition model.

## 1.4 Thesis Outline

The first three chapters of this thesis provide the necessary introduction, background information and an overview of related research. Chapter-1 of the thesis provides the basic introduction and outlines this research work's motivation, approach and objectives through the research questions. Chapter-2 presents an overview of existing research on general emotion recognition in specific domains of speech and vision. Chapter-3 discusses the relevant conceptual information about the thesis.

The following chapters detail the methodology and experiments carried out in the thesis. Chapter-4 describes the architectural details of the entire Dimensional Emotion Recognition Model (DERM), including both the speech and visual networks, respectively. Chapter-5 lays out the experimentation details, including the pre-training and fine-tuning approaches. It also describes the process of acquiring, preparing and pre-processing the dataset while presenting results and lays down points related to the ablation study.

The final chapters summarize the research work and discuss the achievements and shortcomings. Chapter-6 describes the methodology's achievements, limitations, and approach bottlenecks. Chapter-7 provides a concluding remark.

# Chapter 2

# Related Work

ER and identification are challenging, especially considering the unavailability of largely annotated data. Even from the human perspective, it is challenging as the perception of emotion can differ entirely for individuals. Thus, finding a robust feature extraction method to represent the emotional state of speakers through linguistic, acoustic or visual signals is the major challenge faced by the researchers [89]. In order to simplify the problem, many studies try to find salient and discriminative features by focusing on a single modality, either text, speech or visual domain. In comparison, recent success in the linguistic domain has leap-frogged through the advent of social media [4]. Not only limited to the availability of textual content through social media, but the success can also attribute to the breakthrough proposal of transformers [121] and transformer-based language models [6] with models like BERT, Robustly optimised BERT (RoBERTa) [74] and Generalized Autoregressive Pre-training for Language Understanding (XLNet) [133] achieving SOTA.

## 2.1 Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) is a sub-field of ER and has many challenges due to the variations introduced by the speaker's individuality and the noises in the background while processing the audio. Despite the challenges, audios include many acoustic features, including prosodic features like pitch and spectral features like MFCCs that contribute to transmitting emotional content in voice [114]. For this reason, much effort stows upon finding the optimal set of features to utilise and represent these acoustic properties. RL helps extract low-level to high-level features through linear and non-linear transformations to the underlying signals. To garner robust representations of emotions, techniques like pre-training-based auto-encoders [35], CL-based pre-training [67] and recently, GANs are under investigation for SER.

Recently, DL networks deploying CNNs can extract the low-level variations and complexities compared to the conventional methods [60] by using 2D spectrograms.

Regardless, the methods fail to capture the sequential context and temporal dependencies. To counteract this gap, CNN-LSTM-DNN [111] uses LSTM and helps capture long-term dependencies and frequency variations using the utterance-level feature space [60].

While many approaches can recognise these emotions, the target is to form a piece of good, generalisable information using utterance-level information independently and contextually within a conversation. These methods though effective, fail to capture the global context in the feature space. The Conversational Memory Network (CMN) [43] is one of the first approaches to capture the global information using separate memory banks for both interlocutors (speakers) in a dyadic conversation. Using the attention mechanism, an Interaction-Aware Attention Network (IANN) [134] integrates these memories for better performance. For more context and relations information, Graph Neural Network (GNN) explore the connections using the connections bettwen different interlocutors, utterances or even conversations [73, 34]. The limitation with these approaches is the fixed-length context, which limits the model to general and robust representations.

With the help of self-attention and multi-head attention modules, transformers capture the global context and dependencies. Transformer-based architectures have also proven successful in domains including vision and audio. The prime examples of this success in the speech domain are through the introduction of the models like wav2vec [112] and HuBERT [51], along with multiple variants of these models [117, 10, 110, 11, 75]. Despite the success, these large DL models are trained on large datasets, which is only possible for some research works.

To handle this problem, many studies focus on the utterance-level learning of speech processing, which is ideal for the task of SER and results in smaller model sizes [22], especially using CL. In this approach, the anchor's similarity to the positive and negative sample (refer to Chapter-3, Sub-section-3.2.1) drives the feature learning [114, 131, 136]. However, this creates another bottleneck as ensuring the random choice of the negative samples becomes challenging. These samples can belong to the same conversation, speaker, or emotion class and do not ensure a different semantic representation in the feature space.

Constructing the negative samples is even more challenging because the emotional description involves discrete or continuous states. On the other hand, models using the NCL do not experience the same problem compared to CL. It does not require negative samples and has achieved comparable, if not better, performance [22, 77]. There have been other techniques like Transfer Learning (TL) [125] that have also proven to be effective along with Knowledge Distillation (KD), where a more potent modality, like text or vision, guides the weaker modality, like speech, for feature learning [85].

## 2.2 Visual Emotion Recognition (VER)

While speech contains valuable features for emotion recognition, the visual domain expresses important emotional information [68] through facial expressions, micro-expressions, muscle movement and alignment. These are essential factors in human communication that help us understand the intentions of others. Facial expressions are one of the leading information channels in interpersonal communication through non-verbal communication. In contrast to the speech modality, where the research focuses on emotion recognition, research in the visual domain is distributed between emotion recognition and expression recognition. Conventional approaches follow facial component detection like feature extraction using Gabon filters and expression classification or landmark detection. Most of the work related to ER in the visual domain belongs to FER [53, 126, 21]. Recent studies use auxiliary information that can classify into two main categories:

- **Facial Landmarks**: These are the salient points, such as the end of the nose, the eyebrows, or the mouth.
- **Action Units**: These units code the actions of individuals or groups of expression-producing muscles.

This auxiliary information, with DL, has propelled the feature extraction through images with approaches as deep Convolutional Neural Network (CNN), having surpassed with more information retrieval and robust feature learning. To further streamline the VER, the problem divides into static (image-based) or dynamic (video-based recognition) sub-problems. While static images are relatively more straightforward to compute due to the lack of context, the dynamic approach depends on the temporal dependencies making it a challenging problem.

Transformer-based models like [31, 41] help solve the temporal problem along with LSTMs, surpassing the CNN-based networks. Despite the success of these approaches, as mentioned previously (Sub-section-2.1), less focus has been laid on making emotional predictions solely through the facial analysis, with the majority of VER through the conversion of raw audio into MFCCss or spectrograms or using multi-modal approaches, including the facial domain [68, 23, 82, 102].

Apart from this, the introduction of several new techniques, utilising 3D-CNNs or Graph Convolution Networks Graph Convolution Network (GCN)s has seen success, but in terms of practical applications, these models encounter various issues like faces' occlusion affecting the overall performance [68].

## 2.3 Synthetic Data

The majority of the research in the ML paradigm involves the acquisition of data from the real world, be it the images of objects or natural persons. This process of collecting the data and then annotating the data with the help of human annotators is not only time-consuming but also very expensive. Apart from being expensive, it is also error-prone due to different perceptions of the labels by the individual annotators or due to the flaw in the annotation process followed by multiple annotators. Even with a seamless annotation process and 100% alignment among the annotators, humans can only annotate little details with precision.

In contrast to this manual process, synthetically generating the data and annotating the same has proven to be a much more feasible solution. Not only is the variability of the data samples generated can be controlled, but also the annotation details for the created data are unmatchable by the human annotators [72, 128]. Only a little research has taken place using these synthetic datasets. Nonetheless, it is on the rise, with many workshops[1] and competitions[2] held to use synthetic data and analyse it in real-world applications.

[1] ECCV 2022: 4th Workshop on Affective Behavior Analysis in-the-wild (ABAW) focusing on "Learning from Synthetic Data and Multi-Task Learning".
CVPR 2023: 5th Workshop on ABAW focusing on "Valence-Arousal (VA) Estimation".

[2] ECCV 2022: 4th Competition on ABAW focusing on "Learning from Synthetic Data".
CVPR 2023: 5th Competition on ABAW focusing on "Valence-Arousal (VA) Estimation".

# Chapter 3

# Background Information

Before diving deep into the architecture, working and analysis of the built models, this section provides background information about the emotional aspects, including the discrete emotions and the perceived dimensions of these emotions. The focus then shifts to the RL methods and their related research work. This information is beyond any specific domain and emphasises past developments in the landscape of ML.

## 3.1 Emotions

Emotions play a vital role in day-to-day conversations to express one's feelings. Not limited to expression, emotions convey much about an individual's mental health and have been used to monitor disorders like depression in the medical field [57]. Emotions affect health and well-being and can profoundly impact thought and action [98]. Human beings also have this unique ability to express and infer emotions from the combination of text, speech and facial emotions [70]. Even though a combination of all can provide the best inference of the emotion, every modality is still essential as it can express different aspects.

There are two ways in which emotion perception transpires, either discretely or continuously. The following sections focus on these two elements.

### 3.1.1 Discrete Emotions

The most common way of perceiving an emotion, as well as expressing the emotion specifically through conversations, is discrete. Discrete or categorical emotions portray what dimensional emotions cannot do accurately, i.e. the underlying emotional response [50]. These emotions are universal across cultures compared to dimensional emotions (SubSection-3.1.2), which vary in expression and perception [96]. There are seven primary discrete emotions: neutral, happiness, sadness, surprise, anger, fear, and disgust. This theory of discrete emotions has recently been part of our daily lives on all social media platforms, with" likes" and" dislikes" as

common discrete sentiments.

### 3.1.2 Dimensional Emotions

The second way of the two possible ways of perceiving emotions is continuous. It is a much more complex way to sense emotions than the discrete way of perception, as this can vary culturally. Even within a culture, it can vary depending on how an individual perceives emotions. As it is continuous, it reflects the state of the emotion in terms of the degrees or intensity of being a positive or negative reflection of emotion. This dimensional model builds around three mutually orthogonal emotion dimensions: valence, arousal, and dominance [50].

#### Valence

Valence reflects pleasant to unpleasant responses or perceptions of stimuli compared to the rest of the environment [104]. It envisions a continuous range within the contours of extremeness, for example, extremely unhappy to extreme happiness. It can be summarised as happy-unhappy or pleased-annoyed, defining a person's level of pleasure. Figure-1.1 partially reflects this continuous range of emotions.

#### Arousal

Compared to valence, arousal reflects the energy or intensity of the pleasantness or unpleasantness reflected through the valence dimension of the emotion. It conceives a continuous range within the contours of activity, for example, stimulated-relaxed or awake-sleepy. Figure-1.1 partially reflects this continuous range of emotions.

#### Dominance

Dominance, on the other hand, relates to feelings of control. It devises a continuous range within the contours of dominance and submissiveness, for example, influencing-influenced.

This continuity makes dimensional emotion recognition challenging because it is harder to locate the emotions in the three dimensions of valence, arousal and dominance than to predict the category, as human emotions lack temporal boundaries [24].

## 3.2 Learning Methodologies

One thing common among all the ML approaches is that every approach focuses on the same end goal: optimal and generalisable feature extraction and data representation. While conventionally, these features, patterns and representations were hand-crafted using feature engineering. As this feature extraction process could not help find discriminative information, newer implementations soon replaced these weaker algorithms.

RL refers to learning, capturing and extracting more abstract and valuable concepts that can improve performance on a range of downstream tasks given the input data. It aims to find out the posterior probability distribution through the underlying factors of the data. It helps encode high-dimensional data into a lower-dimension representation that is representative of the data itself. It is often confused with dimensionality reduction. However, it is different as RL not only reduces the dimensions but also learns a mapping that can generalise on the unseen data, which does not apply to all dimensionality reduction techniques. Due to the core capabilities of this learning methodology, it uses several principles to ensure good representations [66], as below:

- Representations are distributed and are expressive irrespective of the configuration, unlike one-hot-encoding.
- Representations are abstract and invariant to local changes.
- Representations represent disentangled factors captured from the data.

Recently, RL has established itself in nearly all domains, from speech recognition to object detection, in language through word embeddings and classification, with SOTA performance and real-world applications like driverless cars, agriculture, and smartphones. All these applications either use modality-specific, joint, or coordinated representations. These representations can be a result of using different methods like:

- ***Supervised Learning***: Extract important features making target predictions from the input data.
- ***Unsupervised Learning***: Extracting the features without the labels by clustering or principal component analysis or GANs.
- ***Reinforcement Learning***: Extracting features by maximising the reward on the learnable task.
- ***Self-Supervised Learning***: Similar to UL, extracting features without labels using some pretext task like predicting the missing word in a sentence or a patch of the image.

Considering the scope of this thesis, the following sections will only focus on SSL.

### 3.2.1 Self-Supervised Learning (SSL)

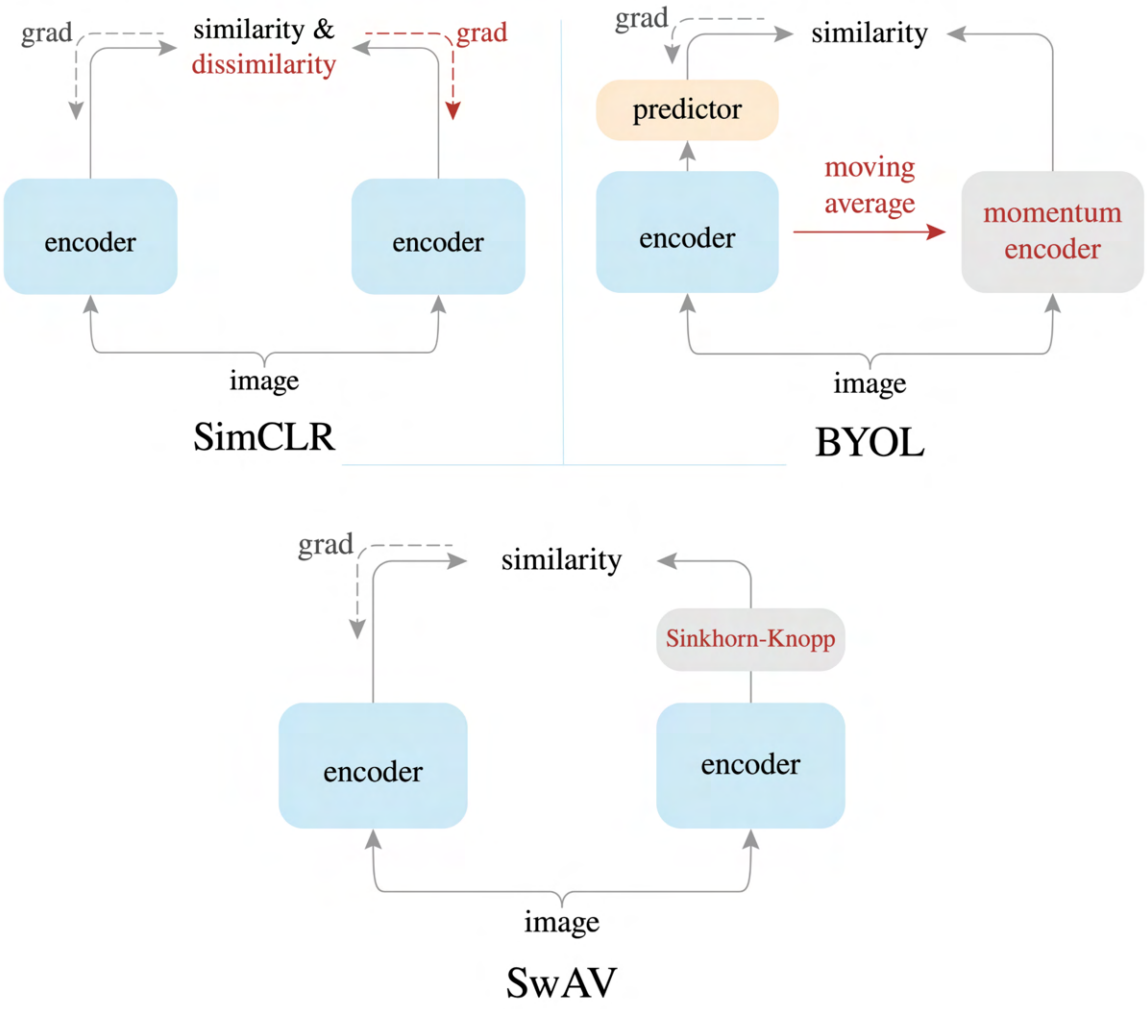


Figure 3.1: Siamese Network based architecture comparison [20].

SSL is a technique to learn the hidden patterns using underlying data but also considering this underlying data as the label itself [88], for example, (1) autoencoders, where the input reconstruction takes place, or (2) the generative models, where we compare the data sample generated with the input samples or (3) predictive coding where the technique focuses on predicting parts of the input including patches of images or predicting rotated images in a visual domain or predicting a word or text in the language domain. Bidirectional Encoder Representations from Transformers (BERT) [28] is a prime example of models based on predictive coding.

All these techniques have a pre-text task on which the learning is accomplished compared to other UL methods. This helps build a richer and more general representation of the underlying patterns in the data. The pretext task of autoencoders includes the reconstruction of the input data from a compressed representation

learned by the encoder. The encoder maps the input data to the latent space which is then decoded back by the decoder to form the output. While the autoencoders try to reconstruct from the input sample itself, generative models try to generate the data samples from random noise and match them with the distribution of the input samples.

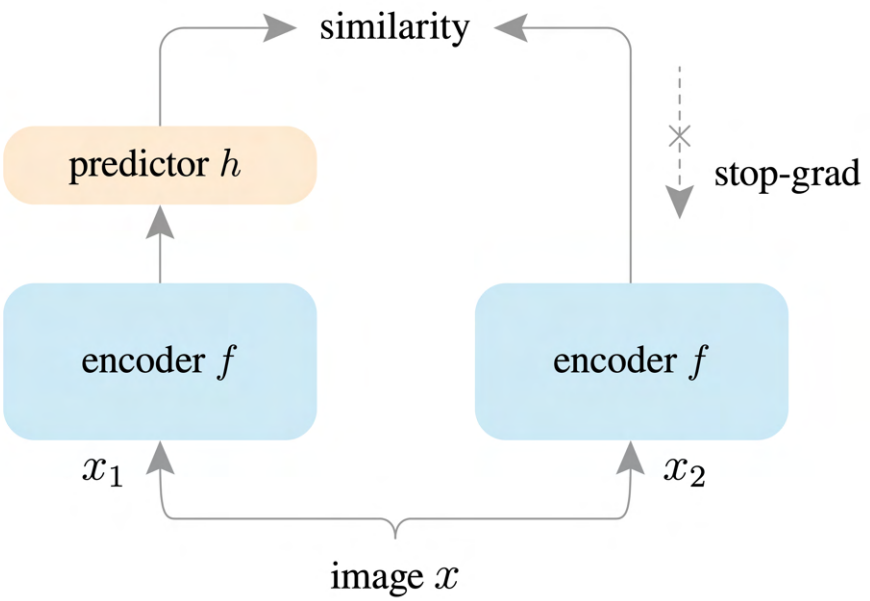


Figure 3.2: SimSiam network with stop-gradient technique in one of the sub-networks [20].

### Contrastive Learning (CL)

Apart from autoencoders and generative models, there is another technique to learn rich representations. It is a sub-branch of SSL called Contrastive Learning (CL) [39, 84, 19, 45, 127]. CL is a technique where the pretext is to learn the representations by distinguishing between similar and dissimilar samples. It differs from clustering as clustering involves grouping similar data points together by maximising the similarity within a cluster and the distances between different clusters. However, in CL, the data points are distinguished based on similarity and dissimilarity by learning the representations and mapping them in the feature space.

CL networks are a form of Siamese networks [14] consisting of two or more identical branches or sub-networks to feed two or more different inputs for learning and distinguishing representations. Initially, these branches share the weights, and the branches are identical. However, this leads to a collapsing problem. It is a problem where the network collapses all the inputs to the same feature or embedding. There are several possible ways to prevent the models from collapsing,

which include (1) adding random transformations to the training samples through data augmentation [127, 52], (2) hyperparameter tuning by tweaking the learning rate, batch size or the temperature scaling [56] or (3) by making architectural changes. It has proven to be the most successful approach in evading the collision problem. Several methods and networks have emerged, including some SOTA models [17, 37, 16] with figure-3.1 representing three of these architectures.

### Non-Contrastive Learning (NCL)

Except for Bootstrap Your Own Latent (BYOL) [37] and Momentum Contrast (MoCo) [45], which use a clustering-based approach through the online network and a momentum encoder, respectively, which provide the negative samples, most of the other methods use direct negative samples along with the anchor input to learn the representations. However, another technique to learn rich representations without using negative samples is Non-Contrastive Learning (NCL). Simple Contrastive Learning (SimSiam) (figure-3.2) is one of the network architectures that learn representations without using any negative samples. This network is not a pure non-contrastive network due to the use of contrastive loss, but despite using no opposing pairs, this method has achieved SOTA on various benchmarks.

SimSiam is very similar to all the previously mentioned contrastive learning architectures. It is BYOL without the momentum encoder, SimCLR without opposing pairs and SwAV without online clustering. Thus related to every architecture by removing one of its core components [20].

# Chapter 4

# Dimensional Emotion Recognition

To recognise and analyse the dimensional emotions, we chose two separate networks for evaluation based on the modality; speech and visual modality, respectively. We passed both networks through two stages: (a) we pre-trained the networks for feature extraction using a large unannotated dataset, and (b) we fine-tuned the networks using the annotated dataset. We introduced an intermediate stage for the speech network where we pre-trained the network in the first stage using the NCL method. In contrast, we used subsequent network pre-training for the intermediate stage using CL methodology and negative samples. We also introduced a similar intermediate stage for the visual network. After pre-training the visual network through CL, we fine-tuned the network on the downstream task of recognising facial landmarks before passing it to the final stage for fine-tuning on the downstream task of dimensional emotion prediction. The following section discusses both networks in detail.

## 4.1 Speech Model - NCL Architecture

In order to independently execute feature extraction and learn robust representations for the analysis and prediction of the dimensional emotions from the input waveform $X_s$, a transformer-based encoder $e_s(\cdot)$ architecture is designed following the stop-gradient[1] technique used in the SimSiam model. This baseline model uses a combination of CNNs and transformers, mainly Convolution-augmented Transformer (Conformer) [38]. It is because CNNs can effectively capture local features and patterns. At the same time, the transformers are efficient in understanding the global patterns and dependencies, and having a combination of both has been successful in recent studies [17]. Figure-4.1 gives the overview of the entire architecture, showcasing the critical components of the speech network, which includes (a) the speech encoder, comprising of the convolutional stack and, subsequently,

[1]**Stop-Gradient**: It is a technique to stop the flow of gradients through a specific part of the network. Instead of using architectures similar to momentum encoder [37], simply using this technique helps in preventing model collapse and assists in extracting salient features without the use of negative samples.

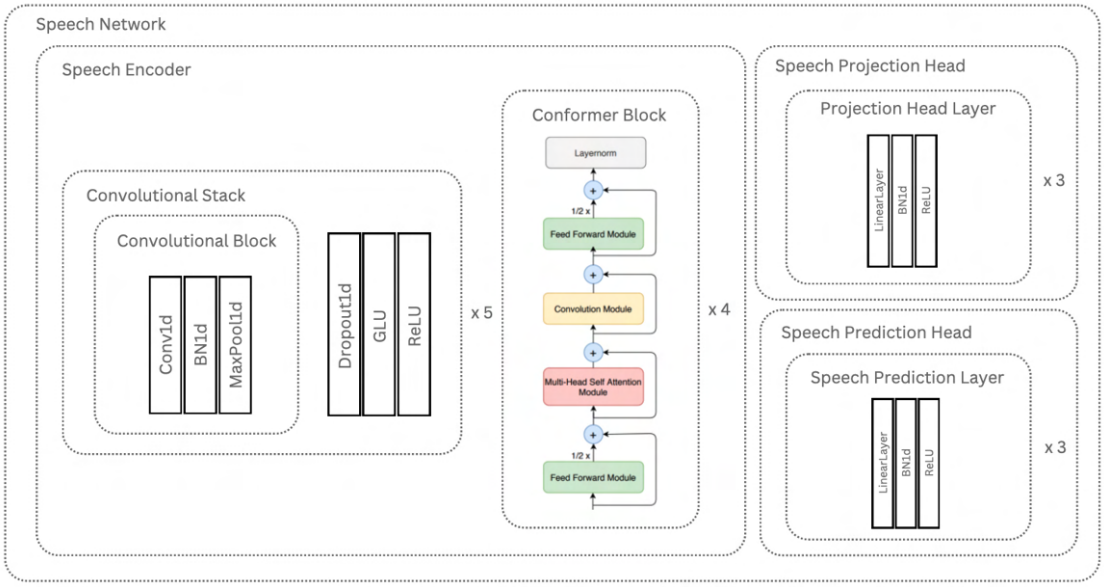


Figure 4.1: Architectural overview of the network used for representation learning in the speech domain. The architecture consists of three main components, namely: (a) *Speech Encoder*, (b) *Speech Projection Head* and (c) *Speech Prediction Head*. The speech encoder is used to capture the representations, while both the speech projection head and speech prediction head are important components that help in optimising the representations learnt through the network.

the convolutional block along with multi-head self-attention through Conformer, (b) the speech projection head, which projects the representations to the latent embedding space and, (c) the speech prediction head, which is a bottleneck structure, behaves like an auto-encoder and helps retain all vital information.

Before jumping onto the other architectures used during the intermediate stage and for fine-tuning, the following sections detail the information on these critical components.

### 4.1.1 Attention-based Speech Encoder

The first and foremost component, and the most important, is the attention-based speech encoder. Depending on the model training stage (as mentioned in Chapter-1 or refer to figure-1.2), the encoder network $e_s(\cdot)$ processes the input waveform[2] $X_s$ through two different layers within the encoder network. One includes the convolution stack, while the other is a layer of conformers, as can be seen in figure-4.1.

[2]**Input Waveform**: Based on the training-stage, input waveform can an anchor waveform $X_s$, an augmented version of the anchor waveform $X_{s_a}$ (Figure-4.2) or the positive $X_{s_p}$ and negative $X_{s_n}$ samples.

#### Convolution Stack

The convolution stack consists of multiple convolution blocks containing the 1D convolution layer, followed by a 1D Batch Normalisation (BN) layer and a 1D max-pooling layer. Since the prediction of emotional dimensions involves using speech utterances $U_i^n$ within a conversation $C$, where $C = \{U_1, U_2, \ldots, U_n\}$, $\forall$ $U_i \subset C$ and $i \in 0 \ldots n$, a local utterance level embedding is desired through the convolution block. The embedding produced through the convolution block is then passed through a 1D dropout layer, Gated Linear Units (GLU) followed by a Rectified Linear Unit (ReLU), a combination of which allows only relevant information being passed through the network and further introduces non-linearity to the network which allows learning more complexities and robust representations.

The stacking of multiple convolution stacks together improves the performance of the network [138] significantly. It allows the network to learn abstract features at every layer, from high-level features in the initial layers to lower-level features in the deeper layers.

#### Conformers

While certain parts of an utterance $U_i^n$ contain more salient features than others, the convolution stack does not capture this. Therefore, we introduce a layer of Conformer having a multi-head self-attention mechanism to learn the global context [38]. Similar to the stack of multiple convolution blocks, multiple conformer heads provide robustness and generalisation to the model as every conformer block forms a separation between the multi-head self-attention and the feed-forward modules with a convolution module. It proves to be highly beneficial to many recognition applications [116]. The input to the conformer layer is the batch-enabled, convolved and encoded waveform features from the convolution stack, which is further transformed and encoded contextually through the conformer modules. For this study, a conformer with four layers, four heads and using a 512D output.

### 4.1.2 Speech Projection Head

Similar to multiple other models [17, 20] based on the Siamese networks [14], where the non-linear projection heads are not only beneficial but outperform networks where no projection or linear projection occurs, the MLP projection head $f_s(\cdot)$ (Figure-4.2) is introduced to maintain the learnt information. This projection layer produces better and more robust representations, which benefit the learning process [17]. The input to this projection head is a 512D contextual representation of the input waveforms. It forms a 2048D output in the latent space through its 3-layer MLP with BN part of every layer followed by ReLU to incorporate non-linearity except for the last layer.

### 4.1.3 Speech Prediction Head

Finally, as no negative samples are available and the network asymmetry relies on the stop-gradient method during the NCL training, the model can be less robust, fail or, in some cases, collapse due to similar representations across the branches of the network. Similarly, the findings reveal that the model finds it challenging to find robustness during the CL training. Therefore, a prediction head $g_s(\cdot)$ or a bottleneck structure (figure-4.2) is introduced to counter these gaps, which works as an auto-encoder and helps create the desired robust representations.

Like the speech projection head, the prediction head is also a 3-layer MLP, receiving a 2048D feature embedding from the projection head, which is passed through the 1024D middle layer to retain the relevant information while providing a 2048D output.

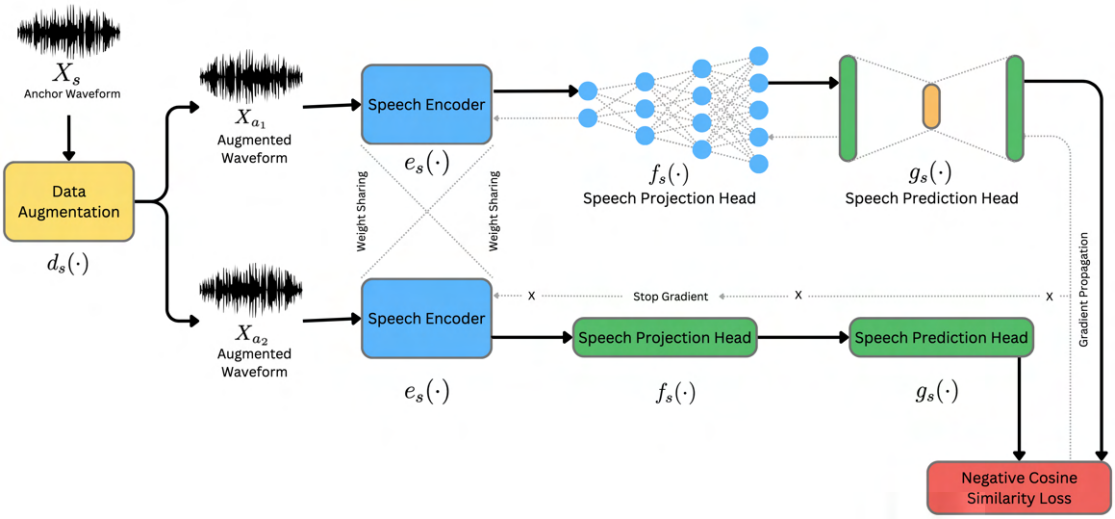


Figure 4.2: NCL speech model training process along with the architectural overview. As shown, an input waveform $X_s$ is passed through a stochastic data augmentation module $d_s(\cdot)$ forming a pair of augmented waveforms, passed onto the respective network branch for feature learning.

Thus, in the first stage of the training involving NCL, the network receives two differently augmented waveforms, $X_{a_1}$ and $X_{a_2}$ waveforms through the speech data augmentation module $d_s(X)$ (Chapter-5, Sub-section-5.2.2), where $X$ is the anchor waveform passed through the augmentation module to produce the $X_{a_1}$, and $X_{a_2}$ augmented waveforms respectively. As can be seen from figure-4.2, both these augmented waveforms are passed onto the two branches of the network, where the speech encoder $e_s(X_{a_1})$ and $e_s(X_{a_2})$ encodes them before projecting them through the feature space $f_s(e_s(X_{a_1})$ and $f_s(e_s(X_{a_2})$ and finally auto-encode them through the bottleneck structure of the prediction head as $g_s(f_s(e_s(X_{a_1})$ and $g_s(f_s(e_s(X_{a_2})$ respectively. Finally, both these representations are evaluated using the negative cosine similarity loss (Chapter-5, Section-5.3) before propagating the gradient

through one branch while using stop-gradient in the other branch to create the required asymmetry. While the method is straightforward, it is important to note that both branches of the network share the weights of the encoders, which is an essential step for optimum learning.

### 4.1.4 Speech Model - (NCL + CL) Architecture

The architecture discussed is unfit for all because the training involves multiple steps, including the NCL, CL and model fine-tuning. As the learning methodology differs, specific changes are required to the base architecture to enable subsequent training using other techniques.

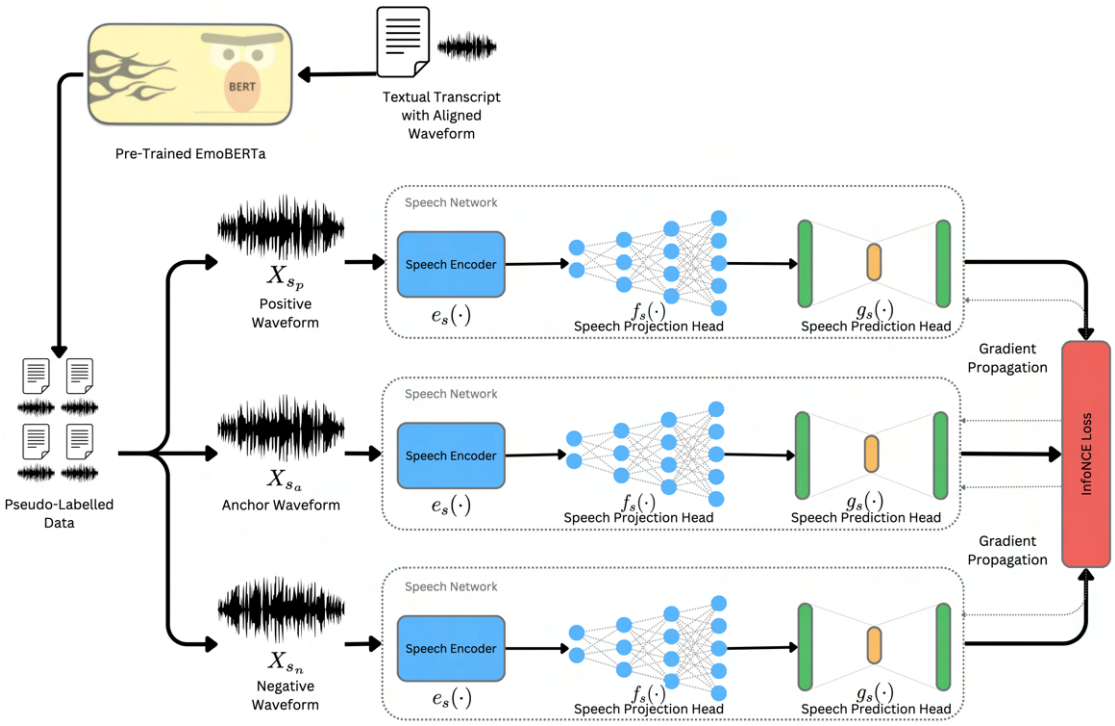


Figure 4.3: CL training process along with the architectural overview. The process involves pre-processing the data using the EmoBERTa to create the required pseudo-labels for creating negative samples. These negative samples along with the anchor and the positive samples are processed respectively for improved representations.

Figure-4.3 represents the architectural variation suited for CL. As can be seen in the figure, core components (speech encoder, speech projection head and speech prediction head) are kept intact from the NCL pre-training while performing subsequent training; however, removing the stop-gradient component. Furthermore, as the training now involves three sets of inputs, including the anchor waveform,

positive (augmented) waveform and the negative (augmented and belongs to a different distribution) waveform, sampled using the pseudo-labels, the negative cosine similarity loss is replaced with the InfoNCE loss for gradient propagation.

### 4.1.5 Speech Network Fine-Tuning Architecture

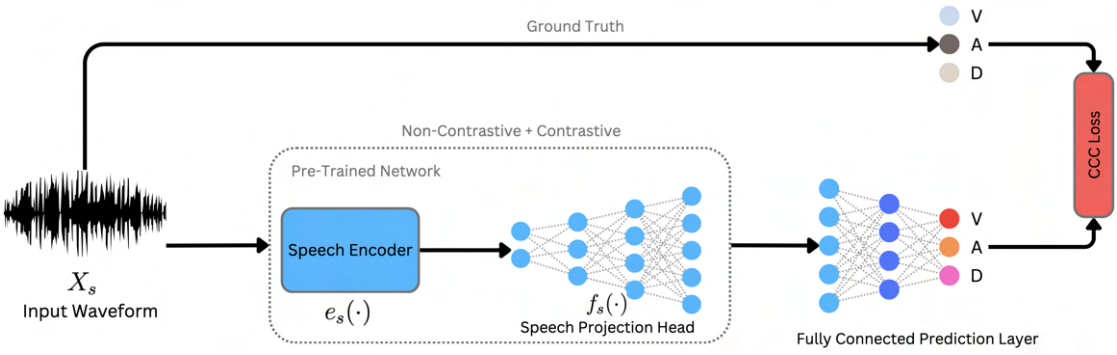


Figure 4.4: Fine-tuning process along with the architectural overview for the speech network using the CCC loss function.

The final changes to the architecture are during the third and final stage of speech model training involves fine-tuning the network for the downstream task of dimensional emotion prediction. Figure-4.4 shows the architectural overview of the model while fine-tuning. For this stage, the input waveform $X_s$ is solely used without passing them through the data augmentation $d_s(\cdot)$ module. The waveform is an input for the non-contrastive and contrastively pre-trained network (NCL + CL). During this phase, the speech prediction head, also referred to as the bottleneck structure, is removed. Since the task of this stage is to use the learnt representations and fine-tune them, therefore, the prediction head is not suitable as the prediction head helped retain the information learning during the training cycle and helps optimise the encoder representations through the loss, which is not essential in this phase due to already learnt representations. Instead, a new 4-layer fully connected network helps predict the required VAD emotions.

As the final predictions are the continuous values, hence CCC loss function replaces the current loss function (Chapter-5, Section-5.3) which is suitable to evaluate the distribution-related similarities and is also used to compare different models because of a uniform range

## 4.2 Visual Model

For executing the feature extraction and learning robust representations for the analysis and prediction of the dimensional emotions from the visual input (input image $X_v$), CL based Mutual Contrastive Learning (MCL)[3] [132] method for learning visual representations. This method helps learn the representations through contrastive learning and mutually following transfer learning, where a cohort of networks transfers the knowledge for optimised representations. Figure-4.5 represents the slightly modified variant of the original MCL architecture. The following sections describe the details of the modification.

### 4.2.1 Mutual Contrastive Learning (MCL) Based Architecture

The core idea behind the MCL approach is to not only perform the training of the single network in an isolated manner for feature learning in a contrastive manner but also to mutually share the knowledge among different sub-network for better feature representations. Thus, through this approach, the authors involve two separate learning schemas, namely Vanilla Contrastive Learning (VCL) and Interactive Contrastive Learning (ICL). While the VCL approach is no different from the general CL approach, where a single network learns from the similarity and dissimilarity of the positive and negative samples compared to the input anchor, whereas in the ICL approach, the contrastive knowledge maximises the feature learning of the other sub-network [132]. In other words, a single network learns from not only it is training but also the training of the other network using the contrastive knowledge attained by the other sub-network.

As shown in figure-4.5, there are two sub-networks having, in total, four branches benefitting from the VCL approach. While this is effective, the branches or the two sub-networks also participate in the contrastive knowledge of the other network through negative sample queues. It is a way of doing online KD, where each sub-network reacts as a teacher to the other sub-network reacting as a student and vice-versa while starting from a different initial condition.

Another essential aspect is that the anchor image $X_v$, through the visual data augmentation module $d_v(X_v)$ creates three different sets of augmentations ($X_{v_1}$, $X_{v_2}$ and $X'_{v_2}$). It differs from the original approach followed by the authors, where only two sets of augmented inputs are created and passed on to the sub-networks. Creating different augmentations for the different sub-networks helps one learn better feature representations as more contrastive knowledge is available compared to similar inputs. Moreover, it creates another level of asymmetry among the networks, which drives collaborative learning and truly takes advantage of the online KD.

[3]**MCL**: Code is available at https://github.com/winycg/MCL.

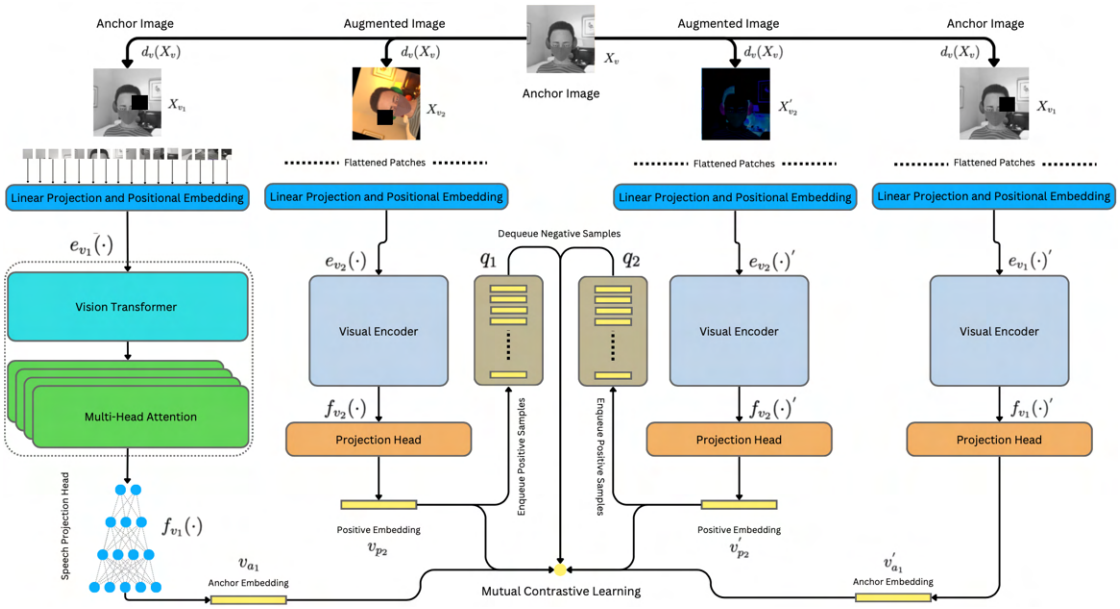


Figure 4.5: The figure exhibits the MCL training architecture as well as the process involved to learn visual representations.

### Vanilla Contrastive Learning (VCL)

VCL approach followed in this model is no different from the one detailed in Chapter-3, Section-3.2.1, where the method tries to bring the positive sample and the anchor closer in the feature embedding space while taking the negative samples and the anchor or negative samples and the positive samples far apart. Like the speech network, we use InfoNCE loss to evaluate similar and dissimilar distributions.

During the VCL phase, the sub-networks receive two different inputs. Taking the case of one of the sub-networks shown in figure-4.5, the sub-network receives $X_v$ as the slightly augmented anchor image and $X_{v_2}$ as the heavily augmented positive sample. Each anchor input image $X_v$ has one positive sample $X_{v_1}$ and $K$ negative samples where $K \geqslant 2$. We convert these input images into a group of flattened patches forming a linear projection with positional embedding and processing them via the sub-network encoder $e_{v_1}(\cdot)$ and $e_{v_2}(\cdot)$, respectively. Each encoder formulation contains the vision transformer followed by the multi-head attention module. The created pipeline passes on the features encoded through the encoder to the visual projection head, a similar MLP network used in the speech network for projecting the encoded features to the feature embedding space before being evaluated using the InfoNCE-based visual contrastive loss.

$$Loss_{VCL} = -log(\frac{exp(\frac{sim(X_v, X_{v_1})}{\tau})}{\Sigma_1^K exp(\frac{sim(X_v, X_{v_K})}{\tau})}) \tag{4.1}$$

where $X_v$ is the anchor input, $X_{v_1}$ is the positive sample and $X_{v_K}$ are the $K$ negative samples chosen for the training. Since, there can be multiple sub-networks where each sub-network follows the isolated VCL approach, therefore the combined visual contrastive loss can be formulated as

$$Loss_{VCLcomb} = \Sigma_1^M Loss_{VCL}(M) \tag{4.2}$$

where $M$ refers to the total number of sub-networks involved for collaborative learning.

**Interactive Contrastive Learning (ICL)**

Compared to VCL, the ICL approach facilitates cross-network collaborative learning, enabling the sub-networks to learn from peer knowledge. In contrast to VCL, where the learning of the contrastive distributions occurs within the embedding space of the network, ICL facilitates the contrastive distribution learning from the embedding space of the peer networks using the gathered contrastive knowledge through the respective sub-network queues $q_1$ and $q_2$, respectively in this two sub-network setup, providing the sub-network specific negative samples. Compared to the VCL, the loss function used here is the Kullback–Leibler Divergence (KL) loss function for comparing the cross-network distributions for better feature learning using the encoder-derived probability distributions.

$$Loss_{ICL} = KL(p_1 \| p_2) \tag{4.3}$$

where $p_1$ and $p_2$ are the derived probability distributions of the two sub-networks, leading to a combined loss as

$$Loss_{ICLcomb} = \Sigma_1^M Loss_{ICL}(M) \tag{4.4}$$

where $M$ refers to the total number of sub-networks involved for collaborative learning.

### 4.2.2 Visual Network Intermediate Training Architecture

Similar to the previously discussed speech network, where architectural modifications were essential for different stages of training, the visual network requires similar modifications for the intermediate training and final fine-tuning. Figure-4.6 showcases the architectural variation required for the intermediate training step.

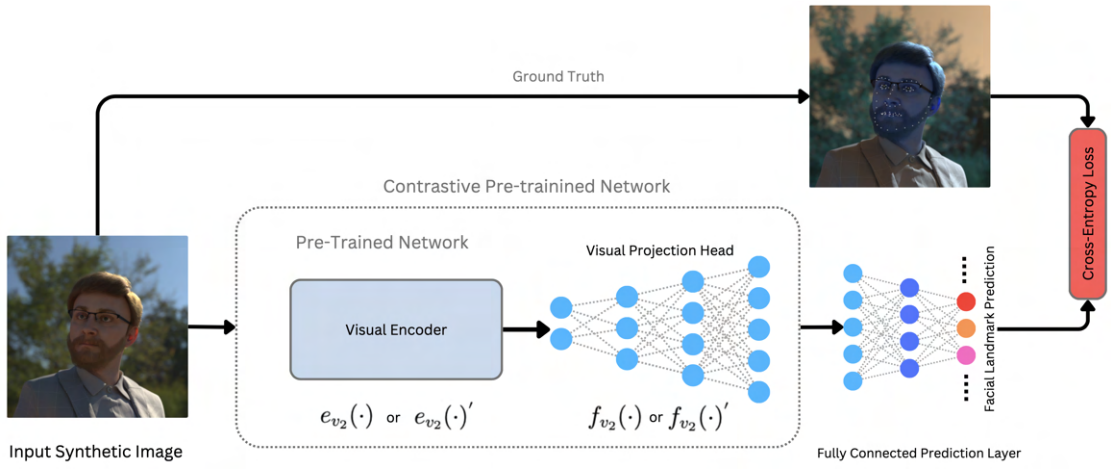


Figure 4.6: Overview of the facial landmark recognition process and architecture using the pre-trained encoder network and cross-entropy loss.

As can be seen from the figure, the core components, including the visual encoder (visual transformer and the multi-head attention module) and the projection head, remain intact while changing the surrounding components. Since this training phase aims to adapt the already learnt representations further to optimise the facial landmarks, a fully-connected prediction network is added to predict the facial landmarks. There is only one branch or sub-network compared to the multiple branches and sub-networks used in the pre-training approach.

For this stage, the input is the synthetic image, passed on to the pre-trained encoder for feature representation and finally to the fully connected layer for landmark predictions. In contrast to the speech network's intermediate phase, the input passes through the data augmentation module $d_v(\cdot)$ as it enhances the learnt embeddings to recognise the facial landmarks. Finally, as the prediction involves 70 facial landmarks, the cross-entropy loss function (Chapter-5, Section-5.3) is used for this classification.

### 4.2.3 Visual Network Fine-Tuning Architecture

The final changes to the architecture are during the third and final stage of the visual model training. The final task is to fine-tune the network on the downstream task of recognising the dimensional emotions. Figure-4.7 represents the final architecture to fine-tune the visual model.

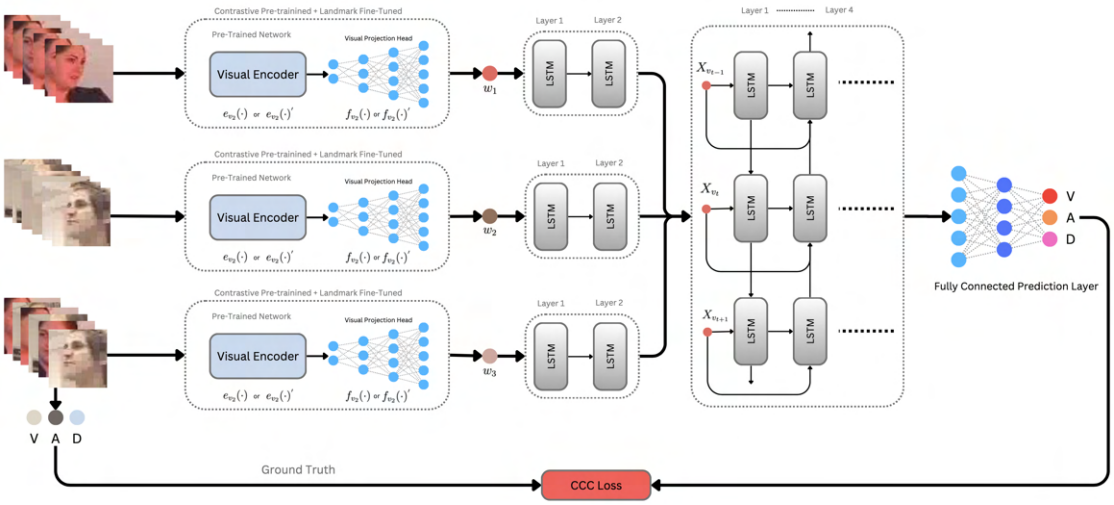


Figure 4.7: Architectural overview of the visual model used to fine-tune the pre-trained encoder embeddings for the downstream task. The network incorporates multiple layers of uni-directional and bi-directional Long Short-Term Memory (LSTM)'s to capture sequential dependencies.

The sequence of images $X_{vt}$ where $t$ refers to the time steps belonging to the frames of the video conversations acts as the input for the pre-trained and optimised encoder network. The network considers the speaker-aware conversational and emotional variations through the inclusion of two different branches respectively for each speaker and the third branch to capture the speaker influence jointly.

While the pre-trained and optimised encoder network forms the feature embeddings, the embeddings assist in capturing the sequential dependencies and variations through the multi-layer unidirectional LSTM based network. While capturing the speaker-aware contextual embeddings from different image frames across time steps is essential, gathering the contextual speaker dependencies is also of utmost importance. Therefore, a third branch achieves this, where the encoder takes the image frames from both the speakers across time steps. An important aspect is that the gated weights control these networks. As in multiple utterances, only one speaker is active while the other speaker is dormant; therefore, in this scenario, more weightage must be given to the network of the active speaker, keeping the information from the other speaker to the minimum and is controlled through these learnable weighted gates.

Even though the smaller LSTM network can capture some context, it cannot capture both past and future context. Hence, to enable the network to capture the entire context of an utterance and unify the speaker aware and joint dependencies, a larger four-layer bi-directional LSTM network is applied for enhanced contextual representations. A fully-connected multi-layer network follows this to help predict the potential VAD continuous distribution. Similar to the fine-tuned speech network architecture, the loss evaluation is through the use of the CCC loss function (Chapter-5, Section-5.3), which is suitable to evaluate the distribution-related similarities and is also used to compare different models because of a uniform range.

# Chapter 5

# Experimentation and Results

In this section, we lay out overall details regarding the steps involved for each training phase and evaluate the effectiveness of all the proposed systems for DER, at the same time comparing them with the SOTA models in the same domain.

## 5.1 Datasets

For the pre-training, incremental training and fine-tuning of the respective AV models, multiple datasets are considered for both the speech modality as well as the visual modality, with a common evaluation dataset to compare the results. For the speech modality, two publicly available speech datasets are used, including TED-LIUM [108, 47] and the IEMOCAP [15] dataset. A detailed description of the datasets is in the upcoming sections.

### 5.1.1 Speech Datasets

#### TED-LIUM Corpus

The first of the many datasets used for this study is the TED-LIUM dataset from OpenSLR. The pre-training of the speech model using NCL is accomplished using this TED-LIUM Corpus[1]. It is an in-domain audio dataset collected from the TED Talks done in English-language. The dataset is available with the transcriptions, with the first two versions of the corpora (TED-LIUM release-1 and TED-LIUM release-2) including 118 hours of audio and 207 hours of audio, respectively. The newer release (TED-LIUM release-3) provides 452 hours of audio [47] in addition to the previous two releases, with all audios sampled at a 16kHz sampling rate. Table-5.1 shows the comparison between the three releases.

[1]**TED-LIUM Corpus** is available with multiple releases (1-3) through OpenSLR-Resources. For the purpose of this study, a combination of all three released versions are used.

Table 5.1: TED-LIUM Corpus characteristics comparison across all three releases (v1 vs v2 vs v3 release).

| Characteristic | TED-LIUM Release Version | | |
|---|---|---|---|
| | v1 | v2 | v3 |
| Total Duration | 118 Hours | 207 Hours | 452 Hours |
| Male Hours | 81 Hours | 141 Hours | 316 Hours |
| Female Hours | 36 Hours | 66 Hours | 134 Hours |
| Total Speakers | 666 | 1242 | 2028 |

### 5.1.2 IEMOCAP Dataset

The second dataset used for the purpose of fine-tuning and evaluation is the IEMOCAP dataset, a widely used dataset for SER. It is a dataset comprising scripted and improvised conversations with dialogues and utterances in the English language, exhibiting a strong correlation with discrete as well as dimensional emotions. It contains 12 hours of audio and visual data, interlocutor's speech, face capture and text transcriptions. The dataset consists of 157 videos of recorded conversations between one male and a female actor, including a total of 10 unique speakers. These conversations remain in a logical sequence forming a total of 10039 utterances, with each audio utterance sampled at 16kHz of the sampling rate.

At least three annotators have analysed each utterance in a conversation before annotating the dataset into emotion labels and deciding the continuous dimensional ratings. Discrete emotion labels include happiness, sadness, anger, fear, joy, frustration, surprise and neutral. The data is collected and distributed into five different sessions having different speakers (both male and female), with each session comprising the recordings of dialogues between the two speakers. Within these sessions, the dimensional attributes of the emotion are provided through valence, arousal and dominance, with the scores for all of them falling under the range of +1 to +5. Ground-truth labels for each of the conversations and utterances are determined by a majority vote, with nearly 25% of the utterances having no majority label. Hence, in order to counter that, the model evaluation is done using the 4-Way or the 6-Way classification due to the availability of the majority label.

Similar to the speech modality, for training the model using the visual modality, two different datasets are utilised. FaceSynthetics [128], a synthetically generated dataset along with the already mentioned IEMOCAP dataset, are used for the training and evaluation. A detailed description of the FaceSynthetics datasets is in the upcoming section.

### 5.1.3 FaceSynthetics Dataset

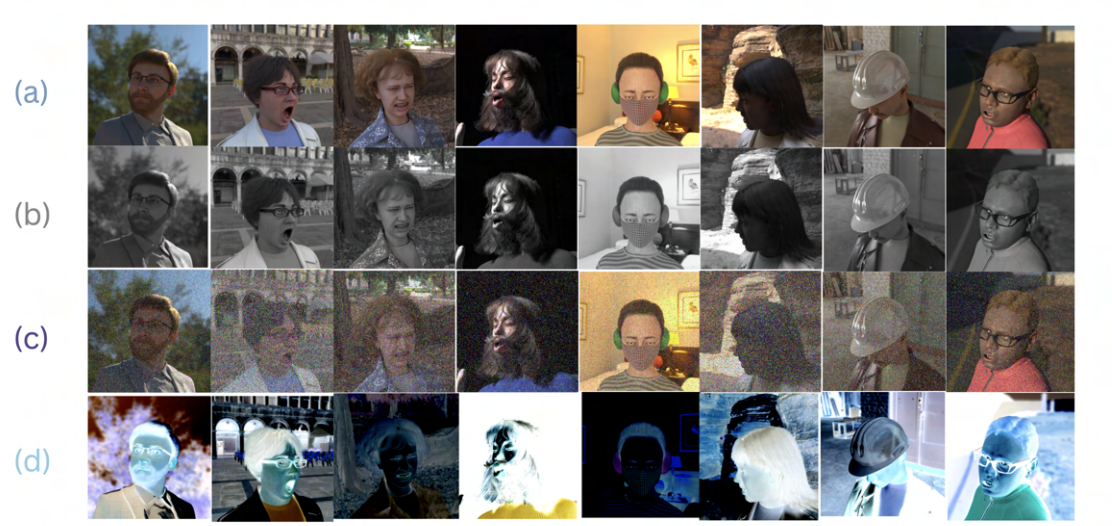


Figure 5.1: Sample images from FaceSynthetics Dataset [128]. From top to bottom, the image shows different types of augmentation (a) Original Image, (b) Grayscale Image, (c) Noisy Image and (d) Brightness and Contrast formatted Image.

FaceSynthetics[2] dataset is a collection of face-related images created synthetically using 3D facial modelling and CV for application through tasks in the wild. The dataset helps deal with the fairness and ethical concerns related to the models and the data in the field of CV. It provides an alternative approach to collecting the data manually, which is not only expensive and time-consuming but also prone to annotation noise due to human error. Instead, synthesised data provides a way to create data with rich annotations with variations across any domain and metric. The dataset consists of 100,000+ high-resolution (512 x 512) images of synthetically rendered faces with diverse ethnicities, backgrounds, clothing, expressions and occlusion, providing a diverse and challenging dataset.

[2]**FaceSynthetics Dataset**: https://github.com/microsoft/FaceSynthetics

## 5.2 Data Preparation

Despite all these being finely curated datasets, they cannot be readily used for our study. Therefore, before jumping on to the training details, this section provides details information about the process followed for preparing the datasets for training.

For our task of learning speech representations using NCL, we use all the released versions of the TED-LIUM corpus. However, the dataset cannot be used directly. As multiple talks from TED-LIUM release-1 and TED-LIUM release-2 overlap[3] with TED-LIUM release-3 having different transcripts, the dataset is first cleaned up to create a unique set of talks with aligned transcripts. The cleanup activity the total time from 777 hours to 664 hours of clean audio, which can be used for pre-training the speech model.

### 5.2.1 Speech Data Pre-processing

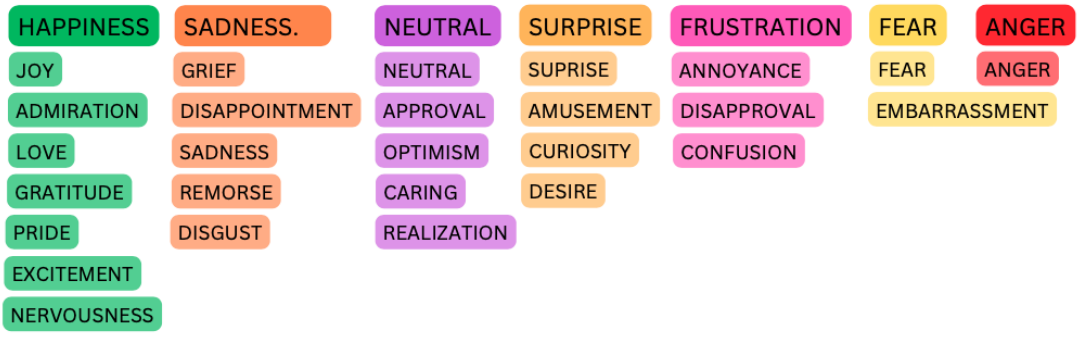


Figure 5.2: Mapping of 28 distinct emotions of GoEmotions dataset to 7 distinct emotions of IEMOCAP dataset.

While the cleanup activity ends up with 664 hours of clean audio, in order to efficiently pre-train the model, some pre-processing is still required. As the approach involves pre-training the model initially with the NCL and subsequently with the CL approach, firstly pseudo-labelling[4] is performed on the dataset using the Speaker-Aware Emotion Recognition in Conversation with RoBERTa (EmoBERTa) [61] as shown in figure-4.3.

EmoBERTa[5] is a model pre-trained to solve the Emotion Recognition in Conversation (ERC) task using textual transcripts. Therefore, the aligned transcripts

[3]Information about overlapping audios with different transcript files can be found here https://www.openslr.org/51/.

[4]**Pseudo-Labels or Pseudo-Labelling** is the task of creating the labels for the unlabeled data using a pre-trained network or sometimes the teacher network to enhance the performance of the underlying model [100].

[5]**EmoBERTa**: https://github.com/tae898/erc

available with the audio are used to create pseudo-labels for the audio, which in turn help curate the negative samples required for the CL task. The model pre-training is done using the A Dataset for Fine-Grained Emotion Classification (GoEmotions)[6] dataset [26].

This dataset contains 28 distinct emotions, which means that the pseudo-labels created using the EmoBERTa model comprises of these 28 emotions. However, as the The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset used for training and evaluation contains only seven distinct emotions, these 28 emotions are to be mapped to these 7 IEMOCAP emotions. This is done by creating a custom logic based on the research under the umbrella of Social Cognitive (SC) and Neuroscience [109, 64], where the emotions like joy, admiration, and love are mapped to a single category of happiness. The complete mapping of emotions is represented in Figure-5.2.

### 5.2.2 Speech Data Augmentation

Data Augmentation is a mechanism which helps attain better generalisation and model performance, especially when using CL or NCL techniques. In order to achieve the same through the NCL approach followed in the first iteration of the model, the data is passed through a stochastic speech data augment module $d_s(\cdot)$, which comprises a combination of augmentations available through A Time-domain Data Augmentation library (WavAugment)[7] [58] and PyTorch's torchaudio[8] library. There are multiple augmentations available, including pitch modification, additive noise and reverberation, including the chain effect, which applies the combination of augmentations over the same waveform. According to Kharitonov et al., combinations of augmentations (also known as the chain effect) result in better and more generalised representations compared to single augmentations. However, in order to create differences across the two branches of the model, we apply single and multiple augmentations[9] randomly in order to achieve consistent performance.

While the process so far involved the pre-processing and augmentation of the speech data, a similar path has to be followed for the visual data aswell. The following section gives the detailed information about the same. While the process so far involved the pre-processing and augmentation of the speech data, a similar path has to be followed for the visual data as well. The following section gives detailed information about the same.

[6] **GoEmotions**: Data and tutorial available at https://github.com/google-research/google-research/tree/master/goemotions.
Decsription available at: GoEmotions Dataset

[7] **WavAugment**: https://github.com/facebookresearch/WavAugment

[8] **Pytorch's torchaudio augmentations**: Audio Data Augmentation

[9] A total of 10 different augmentations are applied including PitchShift, Noise, Reverberation, HighLossPass Filtering and Time Masking with random probabilities along with the augmentations available through the PyTorch's torchaudio library.

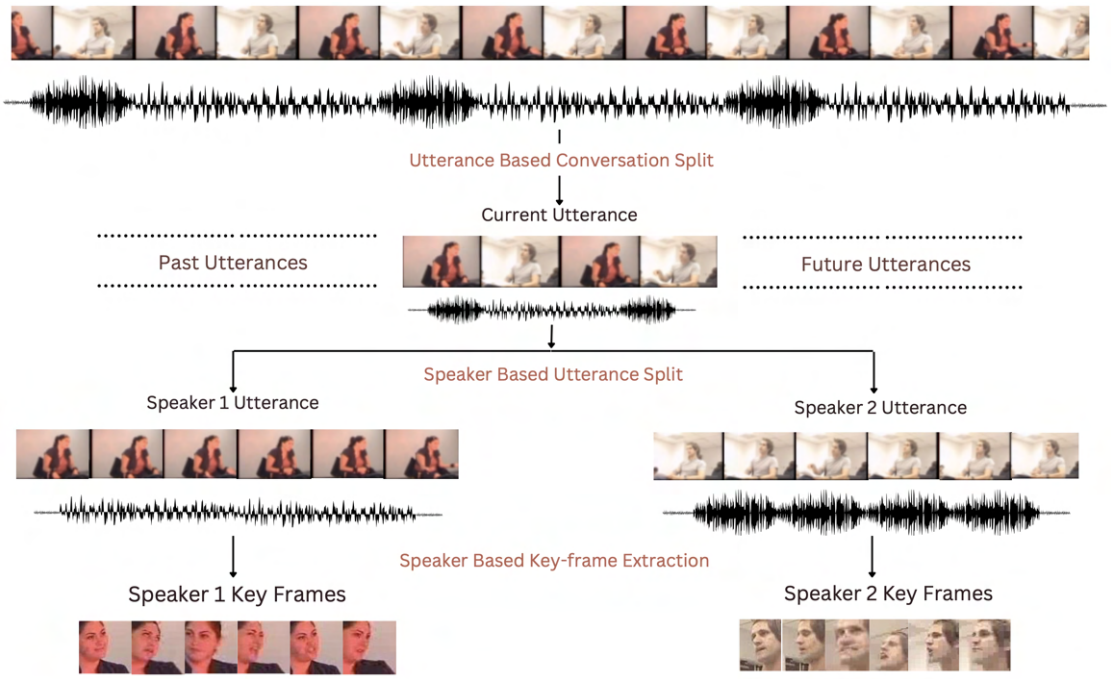


Figure 5.3: IEMOCAP dataset pre-processing steps. Starting from the available conversations, each conversation is split into unique utterances. These utterances are further split based on the speaker to be further processed to extract the speaker-based keyframes.

### 5.2.3 Visual Data Augmentation

Figure-5.1 provides an overview of the dataset and a few selected augmentations (for a complete set of augmentations, refer to Chapter-7). As the synthetic data is nearly perfect, Wood et al. suggest applying multiple augmentations during training which tend to bring better performance and generalisation over the course of model training [128]. Hence, a set of appearance augmentations, including colour shifts, brightness and contrast along with full augmentation[10], varying both appearance and geometry, including rotation, is applied to the dataset through the data augmentation $d_v(\cdot)$ module.

Finally, the dataset used for model fine-tuning and evaluation is prepared before use. The details for the same are mentioned in the upcoming section.

### 5.2.4 Audio and Visual (AV) Data Pre-processing

The first and foremost step while preparing the IEMOCAP dataset involves extracting the audio and visual information (waveforms and videos separately), keeping aside the textual information. However, the textual information could not be

[10]**Full Augmentation** includes rotation and warping along with appearance augmentation. However, warping is not considered in this study due to inconsistencies across images.

completely discarded as the speaker information, utterances and conversation sequences, along with the labels, are also part of the textual transcripts.

Therefore, using the available information through the transcripts, both the audio and the videos are segmented, forming a set of 10,000+ unique utterances across all interactive sessions available through the dataset. For the speech model, this data processing is adequate. However, for the visual model, further processing is required to completely utilise the data available. Therefore, the videos are passed through a split module, where each video is split into two separate videos containing only a single speaker. This is done using the OpenCV[11] library. As the visual model works on the images and not raw videos, hence frames are extracted from the split videos. And not all frames but only keyframes are extracted to avoid redundant frames, which do not add much to the training data. Finally, these frames are further processed to extract the facial information and crop the entire frame using the Haar feature-based cascade classifiers[12] [122, 71] which uses the concept of integral images for computing different features with reduced computation and time. Figure-5.3 showcases this entire process.

### 5.2.5 Audio and Visual (AV) Data Post-processing

There is one more step which is followed during the training or evaluation phase. This phase involves post-processing the data by transforming the valence, arousal and dominance values. As IEMOCAP dataset provides the annotations in the range of +1 to +5, this is not suitable for training or evaluating the model. Hence, this continuous range is changed to a range of -1 to +1 for better comparison. This range also correctly reflects the emotions both negatively and positively across emotional dimensions.

[11] **OpenCV** is a CV library providing tools for real-time analysis and optimisation. More information available at: https://opencv.org.

[12] **OpenCV library for Haar feature-based cascade classifiers**: Cascade Classifiers.

## 5.3 Loss Functions

One of the key components during the training of any model, and especially during representation learning, is the choice of the loss function. There are many different types of loss functions; however, considering the scope of the study and research for this thesis, including the NCL, CL and the dimensional emotions, the following loss functions are evaluated.

### 5.3.1 Negative Cosine Similarity Loss

The Negative Cosine Similarity loss is a loss function which computes the cosine angle between the predicted values and true values to determine the similarity and the dissimilarity between the two. Since it follows the cosine signal, a higher value occurs when the predictions are dissimilar, whereas lower values refer to similar predictions. Therefore, the negative cosine similarity loss tries to maximise the cosine similarity for better performance (Equation-5.1).

$$Loss_{Cosine} = \frac{y_i \cdot x_i}{|y_i||x_i|} \tag{5.1}$$

Where $y_i$ represents the ground truth distribution, and $x_i$ is the predicted distribution for the $i^th$ sample.

### 5.3.2 Cross-Entropy Loss

Cross-Entropy loss is a log-based loss function which measures the difference between the predicted probabilities of the model and the actual probability distribution of the target labels (Equation-5.2).

$$Loss_{CE} = -\sum_{c=1}^{M} y_i \log(x_i) \tag{5.2}$$

Where $y_i$ represents the ground truth distribution, and $x_i$ is the predicted distribution for the $i^th$ sample.

This loss function penalises the model for predicting wrong values with high confidence.

### 5.3.3 Mean Absolute Error (MAE) Loss

The Mean Absolute Error (MAE) loss function is a measure used to calculate the average absolute difference between the predicted values and the true values. Considering the dimensional emotions, it is used to measure the difference between the predicted VAD distribution and the ground truth distribution (Equation-5.3).

$$Loss_{MAE} = \frac{\Sigma_{i=1}^{n}|(y_i - x_i)|}{n} \tag{5.3}$$

Where n is the number of samples, $y_i$ represents the ground truth distribution, and $x_i$ is the predicted distribution for the $i^th$ sample.

The absolute value is used to ensure that the errors are positive and to eliminate any negative errors that may be introduced by the model. As it can be seen from the equation above, the MAE is the non-weighted loss function where all errors have equal weight regardless the absolute value of the error.

### 5.3.4 Mean Squared Error (MSE) Loss

The MSE, similar to MAE, is a loss function used to measure the difference between the predicted distribution and ground truth distribution. Compared to the absolute difference using MAE, it calculates the average of the squared difference (Equation-5.4).

$$Loss_{MSE} = \frac{\Sigma_{i=1}^{n}(y_i - x_i)^2}{n} \tag{5.4}$$

Where n is the number of samples, $y_i$ represents the ground truth distribution, and $x_i$ is the predicted distribution for the $i^th$ sample.

Being a quadratic loss function, it puts greater weight on larger errors by penalizing them heavily compared to smaller errors. This is also a disadvantage as it is sensitive to outliers and is prone to extreme values.

### 5.3.5 Concordance Correlation Coefficient (CCC) Loss

The CCC loss [65] function does not measure the difference but instead the agreement between the two distributions. It takes into account the correlation and deviation of the predicted distribution from the ground truth distribution for evaluation (Equation-5.5).

$$Loss_{CCC} = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_y - \mu_x)^2} \tag{5.5}$$

where $\rho_{xy}$ is the Pearson correlation coefficient between the predicted and the ground truth prediction, $\sigma_x$ and $\sigma_y$ are the standard deviations of the predicted and ground truth distribution, with $\mu_y$ and $\mu_x$ representing the mean values.

The CCC loss function measures within the range of -1 to +1, with +1 indicating perfect agreement compared to -1 indicating maximum disagreement. Therefore, in order to maximise the agreement [3], the loss can be modified and can be defined as:

$$CCC_{max} = 1 - CCC \tag{5.6}$$

The advantage with CCC loss is that as it considers both correlation and deviation, it is robust against the outliers and because of the consistent range, it makes different models more comparable regarding the performance.

Since in the evaluation, all three dimensions are predicted and evaluated; hence, we modify the loss function to fit into this objective. Therefore, for all the attributes of valence, arousal and dominance, we define the loss function with attribute specific weights and for evaluation these weights are set in a balanced manner.

$$CCC_{max} = 1 - [(w_v * CCC_v)] - [(w_a * CCC_a)] - [(w_d * CCC_d)] \tag{5.7}$$

where $w_v, w_a, w_d$ are the respective weights for the each attribute and $CCC_w, CCC_a, CCC_d$ are the respective loss values computed for each attribute. For initial evaluation, the weights are set to be equal with $w_v = w_a = w_d = 1/3$.

### 5.3.6 Triplet Loss

While the above-mentioned loss functions work efficiently for a regression problem, however considering the CL these loss functions do not fit in considering there is no ground truth value to compare to. Hence, there are other loss functions to focus on and one of them being the triplet loss [30].

It is a kind of loss function to measure the similarity among different samples and map the samples in the feature space in such a way that similar samples are close together while keeping the dissimilar samples far apart. It uses three different samples: an anchor sample against which the other two samples, positive (belonging to the same class or distribution) and negative samples (belonging to a different class and distribution) are compared. It is represented as:

$$Loss_{Triplet} = max(0, d(a,p) - d(a,n) + margin) \tag{5.8}$$

Where $d(a, X)$ is the function to measure the distance between either the anchor and the positive samples or the distance between the anchor and the negative samples in the feature space, and the margin is a hyperparameter that controls the minimum distance between the positive and negative samples. The triplet loss is efficient as it penalises a smaller distance between the anchor and negative samples or a greater distance between the anchor and the positive samples. Another advantage is that this measure is effective in the case of imbalanced datasets.

### 5.3.7 InfoNCE Loss

Similar to the triplet loss, InfoNCE loss function is also used while training the model using in the CL technique and is used to capture the structure and learn the underlying representations. However, compared to the former, InfoNCE loss measure the agreement or the similarities between the distributions of the anchor compared to the samples and mathematically equates as:

$$Loss_{InfoNCE} = -log(\frac{exp(\frac{sim(i,j)}{\tau})}{\Sigma_k exp(\frac{sim(i,k)}{\tau})}) \tag{5.9}$$

where sim(i,j) is the similarity between the $i^th$ and $j^th$ examples in the input data, $\tau$ is a temperature hyperparameter controlling the softness of the probability distribution, and $\Sigma_k exp(\frac{sim(i,k)}{\tau})$ is a normalisation term ensuring probability sums up to 1.

As portrayed through the Equation-5.9, it is different from triplet loss as it does not require triplets and can be used with multiple forms of data or datasets.

## 5.4 Training Setup

The experiments focus on six different setups, one for each training stage (three stages including pre-training, intermediate-training and fine-tuning) per domain (speech and visual).

### 5.4.1 Model Hyperparameters

All our models have a specific role to play in the entire pipeline of detecting the dimensional emotions. Hence, the hyperparameter details for each of the networks are laid out in the following sections. However, few of the hyperparameters used are consistent throughout all the models unless specified otherwise, which includes the learning rate ($lr$), which is set to 0.05 and is configured using cosine annealing technique to improve convergence. The optimiser used is Stochastive Gradient Descent (SGD), for which the momentum is set to 0.9 with weight decay ($wd$) of $10^{-4}$.

#### Non-Contrastive Learning (NCL) Based Speech Architecture

For the training of this model, we select a batch size of 128 which is in line with the findings by the authors in the SimSiam model [20], where the performance remains same or decreases after increasing the batch size above 512. For the purpose of evaluation, gradient accumulator ($ga = 2$) is applied taking the total batch size to 256. Along with that, the final embedding size $D = 2048$, which is achieved using the projection head and recreated using the prediction bottleneck structure where the hidden layer's dimension size is set to 1024 unlike the original model, where this has a dimension of 512.

For the multi-head attention, a 4-layer, 4-head Conformer is used which receives the embeddings from a 5-layer convolution stack with an embedding size of 512. In total, the model is trained for a total of 50 epochs.

#### Contrastive Learning (CL) Based Speech Architecture

For this phase of the training, the core structure and the hyperparameters are kept unchanged with the exception of batch size, which due to the size of the model and complexity of the problem is set to 16. For efficiency, it is increased to 64 with the use of gradient accumulator $ga = 4$. Since, the batch size is of 16, a total of $K = 14$ negative samples are used per batch, however, this number increases to $K = 62$ with the use of gradient accumulator. Similar to the NCL approach, the model is trained for 50 epochs.

#### Fine-Tuning Based Speech Architecture

The only change applied to this model is the addition of the multiple dense layers, producing a final embedding of $D = 3$, corresponding to valence, arousal and

dominance for the evaluation against the ground truth. The training is done with a batch size of 64, without any gradient accumulator for a total of 200 epochs.

Table 5.2: Hyperparameter comparison across the three stages of the Speech Model Training.

| Attributes | NCL-Based Speech Model | CL-Based Speech Model | Fine-Tuned Speech Model |
|---|---|---|---|
| Batch Size | 128 | 16 | 64 |
| Gradient Accumulator | Yes | Yes | No |
| Batch Size with Gradient Accumulator | 256 | 64 | N/A |
| Learning Rate | 0.05 | 0.05 | 0.05 |
| Momentum | 0.9 | 0.9 | 0.9 |
| Weight Decay | $1.0e4$ | $1.0e4$ | $1.0e4$ |
| Warm Start | No | No | No |
| Epochs | 50 | 50 | 200 |
| Frozen Encoder | No | No | Yes |

### Mutual Contrastive Learning (MCL) Based Visual Architecture

For this model, the hyperparameter settings followed in SimCLR [17] are used, where $\tau = 0.1$. However, the embedding size is changed from 128 to 512 for consistency across models, which is achieved using the projection head.. In contrast to the SimCLR approach, the total batch size[13] used is of 16. In order to increase the batch size, similar to that of the speech model, gradient accumulator ($ga = 4$) is used to help enhance the representations. Similar to the speech contrastive network, with the batch size of 16, a total of $K = 14$ negative samples are used per batch, however, this number increases to $K = 62$ with the use of gradient accumulator.

The learning rate of 0.1 is used and is configured using the cosine annealing technique, which kicks in after the period of 20 epochs, as the first 20 epochs are used as a warm start within the total of 100 epochs.

[13]Due to hardware limitations, only a batch size of 16 was possible for efficient learning

#### Optimisation Based Intermediate Visual Architecture

This model uses a total batch size of 64, and is trained for a total of 100 epochs. As the network follows the SL approach, the pre-trained encoder is followed by a fully-connected layer, leading to a final output of $D = 70$, where each prediction leads to a facial landmark point.

#### Fine-Tuning Based Visual Architecture

Finally, the model used for fine-tuning remains similar to the intermediate network, but the fully-connected layer is replaced by a smaller fully-connected layer, to predict 3 outcomes of valence, arousal and dominance with the output of $D = 3$. The network is trained for a total of 200 epochs.

Table 5.3: Hyperparameter comparison across the three stages of the Visual Model Training.

| Attributes | MCL-Based Visual Model | Intermediate Visual Model | Fine-Tuned Visual Model |
|---|---|---|---|
| Batch Size | 16 | 64 | 64 |
| Gradient Accumulator | Yes | No | No |
| Batch Size with Gradient Accumulator | 64 | N/A | N/A |
| Learning Rate | 0.1 | 0.05 | 0.05 |
| Momentum | 0.9 | 0.9 | 0.9 |
| Weight Decay | $1.0e4$ | $1.0e4$ | $1.0e4$ |
| Warm Start | Yes | No | No |
| Epochs | 100 | 100 | 200 |
| Frozen Encoder | No | No | Yes |

### 5.4.2 Evaluation Metric

As the scope of the thesis is limited to the continuous prediction of the VAD dimensional attributes of the emotion, CCC scores are used as an evaluation metric for each of the attributes of valence, arousal and dominance both for the speech and the visual domain. For fair evaluation, the metric is used using the best-performing epoch on the validation set for all the models. On top of that, direct comparison with the SOTA is another metric used to evaluate the model performance.

Furthermore, for appropriate comparison, all the results are computed using the standard Leave-One-Session-Out (LOSO) 5-fold session cross-validation and the average of all the results is reported.

## 5.5 Results

### 5.5.1 Contrastive over Non-Contrastively learnt Representations

Table-5.4 and Table-5.5 present the performance of the models curated in this thesis in terms of CCC scores for valence, arousal and dominance in both speech and visual domain respectively based on two different emotion classifier: (a) 6-Way emotion classifier which includes "neutral", "anger", "sadness", "happiness", "excitement" and "surprise" as the discrete emotions whereas (b) 4-Way emotion classifier includes just "neutral", "anger", "sadness" and "happiness" as the discrete emotions.

Table 5.4: Comparison of the two different approaches followed for learning speech representations. The analysis is based on the 4-way and 6-way emotion classifiers. The arrow indicator (⇑) represents better performance with a higher score.

| Speech Model | $CCC_{Valence}$ ⇑ | $CCC_{Arousal}$ ⇑ | $CCC_{Dominance}$ ⇑ |
|---|---|---|---|
| **NCL (4-Way)** | **0.563** | **0.636** | **0.511** |
| NCL (6-Way) | 0.561 | 0.625 | 0.503 |
| **NCL + CL (4-Way)** | **0.601** | **0.686** | **0.612** |
| NCL + CL (6-Way) | 0.596 | 0.646 | 0.603 |

Table 5.5: Visual Model performance comparison on 4-way and 6-way emotion classifier. The arrow indicator (⇑) represents better performance with a higher score.

| Visual Model | $CCC_{Valence}$ ⇑ | $CCC_{Arousal}$ ⇑ | $CCC_{Dominance}$ ⇑ |
|---|---|---|---|
| **CL (4-Way)** | 0.6 | **0.5877** | 0.513 |
| CL (6-Way) | 0.6 | 0.574 | **0.515** |

As can be seen from both the tables, the models in both domains perform significantly better using the 4-way emotion classifier compared to the 6-way emotion classifier. This shows that the model is able to generalise well on the data when there is a majority label available (as explained in Section-5.1.2) compared to when they are not. Especially with the models in the speech domain, there is a noticeable improvement across all three attributes of the dimensional emotions compared to the visual model, where the improvement is only seen for arousal

while it remains the same for the other two attributes.

Another important aspect that can be noticed is the performance improvement caused by the intermediate pre-training performed using CL on the already pre-trained encoder using NCL. The NCL + CL based pre-trained encoder outperforms the NCL based encoder. It can be hypothesised that despite learning good representations without the use of negative samples, the representations can be further improved by providing contrastive knowledge (negative samples) to the network. Thus, producing better and more generalised representations which facilitates and helps improve the performance on the downstream task, in this case, dimensional emotion prediction.

Table 5.6: Comparison of ***Our*** model with the SOTA models in the visual or joint domain based on the CCC scores on the IEMOCAP Dataset. The arrow indicator (⇑) represents better performance with a higher score.

| Visual Model | $CCC_{Valence}$ ⇑ | $CCC_{Arousal}$ ⇑ | $CCC_{Dominance}$ ⇑ |
|---|---|---|---|
| Joint Temporal Aware Model** [83] | 0.5883 | 0.6689 | - |
| Cross-Attention Fusion Model** [102] | 0.670 | 0.590 | - |
| User Emotion Recognition** [76] | 0.586 | 0.171 | - |
| $VGG - Video$* Base Model [137] | 0.486 | 0.549 | 0.212 |
| Real-Time Emotion Detection* [135] | 0.389 | 0.39 | 0.403 |
| DialigueRNN* [1] | 0.37 | 0.6 | 0.37 |
| MIMAMO Net* [2] | 0.529 | 0.377 | - |
| Multi-CNN* [63] | 0.535 | 0.365 | - |
| Personalised Affective Memory* [12] | 0.60 | 0.34 | - |
| ***Ours Visual Model (CL)*** | ***0.6*** | ***0.5877*** | ***0.513*** |

This is also visible through the 2D prediction space for both the 4-way (Figure-5.4) and 6-way (Figure-5.5) emotion classifier. Both of these figures portray the suitability of the representations across discrete emotions, being able to separate the properties of each discrete emotion from the other. For example, the network

is able to distinguish between sadness and anger as well as happiness from neutral emotion.

### 5.5.2 Comparison with SOTA

Table-5.7 and Table-5.6 showcases the comparison of our approach with the previous SOTA models using the CCC scores individually across all the dimensional attributes. As it can be seen from the tables, the NCL + CL speech model approach outperforms all the other previous SOTA models evaluated on the IEMOCAP dataset with an improvement of 34% on valence, 3.4% on arousal and 18% on dominance except preCPC model [69], which is trained using the LibriSpeech [97] using CPC approach. While there is no visible evidence for the difference in the performance, however, it can be due to the hardware capabilities available[14] which restrict the batch size compared to the multi-GPU training implemented with higher batch size for CPC.

For the visual model, however, it is really difficult to evaluate the model as there are not many models available which predict the dimensional emotion attributes using the CCC score using IEMOCAP as the evaluation dataset. Hence, we compare the results with the models being evaluated on IEMOCAP. Along with them, we also report the results of the models using different datasets (represented using $^*$) solely for reporting purposes. Among these models, some of the approaches involve joint predictions using either the text or speech along with the visual features (represented using $^{**}$). Furthermore, none of the models using the synthetic data applies the learning for this downstream task. Hence, for the evaluation, the visual model is compared with the SOTA joint learning models where the speech modality assists the visual modality for prediction. It is important to note that all these approaches involve real-world images or faces compared to the synthetic data used in our approach.

As seen from Table-5.6, CL based visual model is comparable to all the SOTA models using the dual modality for learning representations. This not only supports the hypothesis that synthetic data help learn generalisable representations but also the approach used in this study helps in learning efficient feature embeddings for emotion prediction in the continuous domain. It is also worth noticing that the performance of the visual model is on par with the NCL approach used for the speech model.

[14] All the experiments are performed using 2 NVIDIA GeForce RTX 2080 Ti GPUs, having 11GB of available memory

Table 5.7: Comparison of ***Our*** models with the SOTA models in the speech domain based on the CCC scores on the IEMOCAP Dataset. The arrow indicator (⇑) represents better performance with a higher score.

| Speech Model | $CCC_{Valence}$ ⇑ | $CCC_{Arousal}$ ⇑ | $CCC_{Dominance}$ ⇑ |
|---|---|---|---|
| **preCPC [69] SOTA** | **0.752** | **0.752** | **0.691** |
| 14-Layer CNN [120] | 0.259 | 0.431 | 0.272 |
| Wav2Vec-2.0 (Base) [11] | 0.409 | 0.602 | 0.488 |
| Hubert (Base) [51] | 0.413 | 0.630 | 0.487 |
| Wav2Vec-2.0 (Large) [11] | 0.384 | 0.654 | 0.478 |
| Hubert (Large) [51] | 0.425 | 0.639 | 0.465 |
| Wav2Vec-2.0 (Large and Robust) [51] | 0.448 | **0.663** | **0.518** |
| Wav2Vec-2.0 (Large and Vox) [124] | 0.412 | 0.658 | 0.496 |
| Wav2Vec-2.0 (Large and Xls-R) [8] | 0.399 | 0.657 | 0.496 |
| ***Ours Speech Model (NCL)*** | ***0.563*** | 0.636 | 0.511 |
| ***Ours Speech Model (NCL + CL)*** | ***0.601*** | ***0.686*** | ***0.612*** |

## 5.6 Ablation Study

### 5.6.1 Impact of Batch Size

As visible from figure-5.6, a larger batch size benefits the overall model performance. With the smaller batch size, the models stop learning and never converge, i.e overfitting whereas, through larger batch size, the weight updates are not as frequent as in the case of lower batch size. Hence, optimised performance.

A similar trend is seen in the case of CL as well, which benefits from a larger batch size as more negative samples are available compared to having a lower batch size.

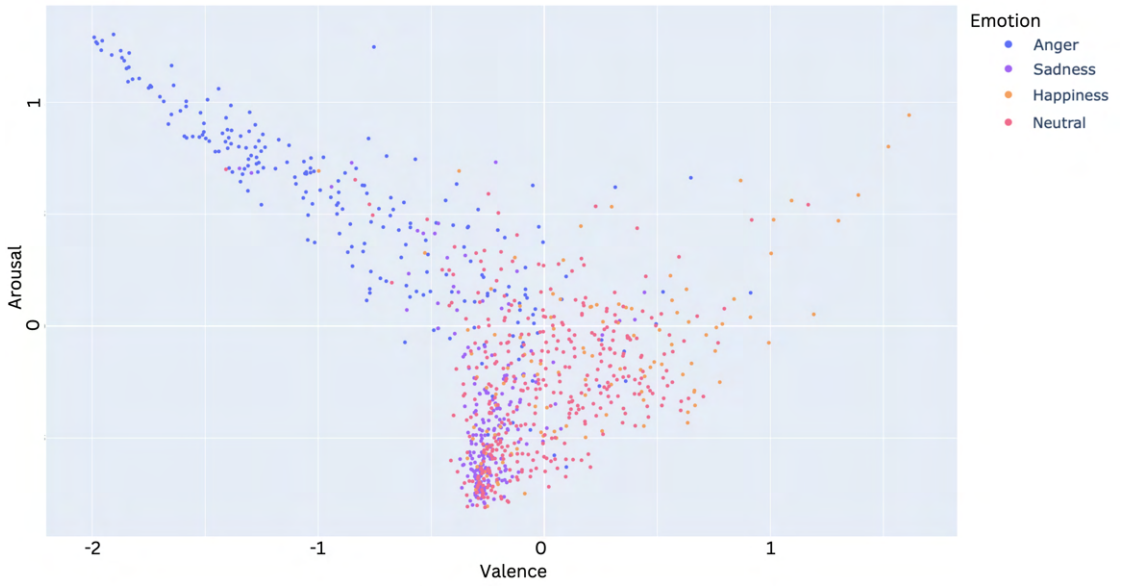


Figure 5.4: 2D attribute space for the dimensional emotions using the 4-way classifier.

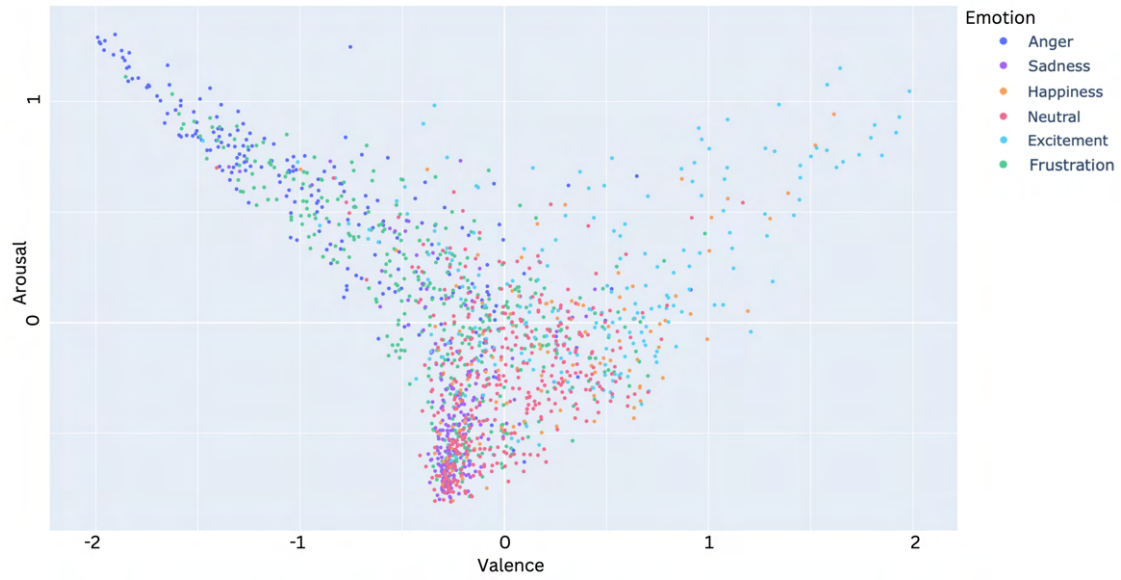


Figure 5.5: 2D attribute space for the dimensional emotions using the 6-way classifier.

### 5.6.2 Impact of Gradient Accumulator

Gradient Accumulator works in a way that reduces the frequency of weight updates by accumulating the gradient over multiple steps. This proves to be beneficial in the case of NCL as the learning approach benefits from less frequent updates. Compared to NCL, CL seem to have less to no performance improvement with gradient accumulator. The hypothesis behind this could be the number of negative

samples which remain the same. As only the gradient is accumulated which helps improve the training time; however the number of negative samples seen by the model remains the same. Hence, the benefit of the training process is limited.

### 5.6.3 Impact of equally weighted CCC Loss

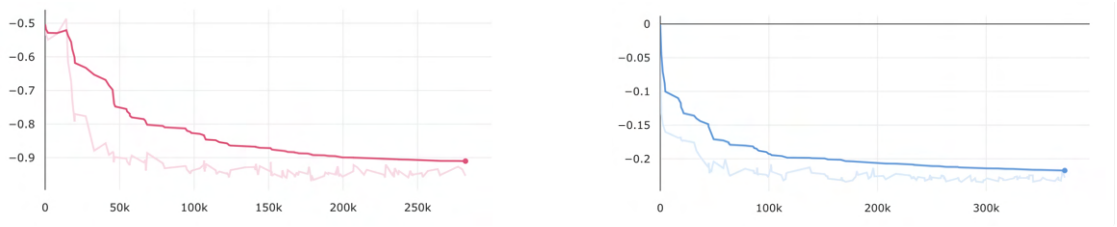


Figure 5.6: (Left) NCL training loss plot with batch size = 128. (Right) NCL training loss plot with batch size = 32.

As explained using equation-5.7, initially the model performance is evaluated using the equal weights of 1/3. However, the loss has a different effect on different models. While using different weights of the loss for speech model benefits when penalising arousal and domain heavily, compared to valence (Table-6.1), the effect is negligible for the visual model. This is further discussed in Chapter-6 under Sub-section-6.1.1.

# Chapter 6

# Discussion

As mentioned in Section-1.3, the focus of the thesis revolved around four research questions. Taking into consideration the experimentation, results and analysis, the upcoming sections will bring every research question into focus and delve into the findings of this study.

## 6.1 Dimensional Emotion Representation

***Research Question 01***: Is the emotional dimension of "arousal" better represented through speech modality, whereas the visual modality correlates highly with the "valence" dimension of emotion?

In order to investigate and do the analysis for this question, we design the scope of this study in such a way that the two modalities of speech and vision are kept isolated. We curate two models for each modality, pre-train them and finally, fine-tune them for the best possible results.

### 6.1.1 Training Results

As shown in Table-5.4 and detailed in Section-5.5, the speech models including both the Speech-NCL model as well as Speech-CL model, perform significantly better on the arousal dimension of the emotions compared to the valence as well as the dominance dimension. While the speech modality can capture the distribution of the valence dimension across the discrete emotions to some extent, the correlation is significantly higher for the arousal dimension.

We notice a similar correlation while performing the ablation using the weighted loss function in Section-5.6. The performance of the speech models significantly improved when the loss was heavily penalised for the arousal domain compared to the penalty applied for the valence domain, where there were no significant gains in the final prediction. However, the model's performance degrades instead, suggesting more correlation with the arousal dimension for the speech modality.

Table-6.1 shows the quantitative results

Table 6.1: Effect for penalising the dimension of the emotion equally as well as heavily within the loss function.

| Penalised Dimension | $CCC_{Valence}$ | $CCC_{Arousal}$ |
|---|---|---|
| Equal | **0.563** | 0.636 |
| Arousal | 0.443 | **0.654** |
| Valence | 0.513 | 0.586 |

While the results for the speech model are easy to interpret, it is slightly more complex in the case of the visual modality. As shown in Table-5.5, the model outperforms in the valence domain of emotion compared to the arousal dimension. However, the difference is insignificant. It could be due to the variety of expressions available, including many tangible expressions through the synthetic data. It is usually not the case with real-world data where the expressions are more subtle.

## 6.2 Emotional Dimension Correlation

***Research Question 02***: Do emotional dimensions correlate with facial attributes like facial landmarks?

It is an essential aspect as humans tend to portray and perceive emotions through facial expressions, which affect facial attributes. In order to do our analysis on this question, an intermediate training step for the visual modality involves fine-tuning the encoder representations using the facial landmarks available through the FaceSynthetics dataset. Table-6.2 shows the impact of this step, which clearly shows the positive impact of the fine-tuning approach, helping gain significant performance on all dimensions.

Table 6.2: Effect for penalising the dimension of the emotion equally as well as heavily within the loss function.

| Training Phase | $CCC_{Valence}$ | $CCC_{Arousal}$ | $CCC_{Dominance}$ |
|---|---|---|---|
| Contrastive Pre-Training | 0.349 | 0.401 | 0.243 |
| Encoder Fine-Tuning | **0.6** | **0.5877** | **0.513** |

## 6.3 Synthetic Data - The Solution?

***Research Question 03***: Can synthetic data help learn universal and generalisable facial features and bridge the gap with real-world data?

Since the scope of studying the emotional dimensions is in the visual domain, synthetic data is a potential solution for further analysis in the field. The performance achieved solely through the pre-training was nominal. However, further fine-tuning of the model using the synthetically available landmarks brings the model performance on par with the previously available SOTA models.

## 6.4 Iterative Pre-Training

***Research Question 04***: Can iterative pre-training of the models using unsupervised learning improve the performance on the downstream tasks?

Our hypothesis about the performance optimisation using iterative pre-training proves to be correct as the iteratively trained model outperforms the model pre-trained once by 11% on mean performance across all the dimensions of the emotion. It is a significant improvement showing better representations than the ones learnt using only the NCL for the speech model.

# Chapter 7

# Conclusion

In this work, we presented multiple approaches and architectures for predicting the dimensions of the emotions using the Non-Contrastive Learning (NCL) and Contrastive Learning (CL) approach for the speech modality, whereas using CL and intermediate encoder fine-tuning for the visual modality to predict the emotion dimensions. The effects of the training and the impact of the training approach, along with the choice of datasets, are carefully analysed. Combining the findings improves the models considerably over previous methods using Self-Supervised Learning (SSL) or transfer learning. The model training and the results discussed previously are through the use of the publicly available standard datasets of IEMOCAP, TED-LIUM and FaceSynthetics. This thesis demonstrates a correlation between speech and arousal and between the visual aspect and valence dimension of emotion.

Furthermore, we discuss the positive effects of iterative pre-training using UL, improving the performance on the dimensions of the emotion significantly (11% on the mean prediction across all three dimensions of valence, arousal and dominance). The approaches followed led to better model performance than all previous SOTA and were on par with the current SOTA models in the speech domain. Whereas in the visual domain, the model outperforms all SOTA models, and performance is on par with the fusion models using more than one modality.

An additional contribution is the highlighted correlation between the facial landmarks and the emotion dimensions. While the pre-training provided nominal results on the downstream task, the facial landmark-based fine-tuning propelled the performance. Moreover, this is only possible due to the availability of 70 facial landmark points annotated synthetically and precisely, which is not possible with a human annotation which is error-prone. It further leads to the hypothesis that synthetic data can help bridge the gap with real-world datasets, as curating these datasets is cost-effective with the data generation done using computer graphics and without human annotations. It is also less time-consuming as the data creation task replaces data collection. Another aspect is the effectiveness of the dataset as the data is diverse, with variations in ethnicity, background, lighting effects and

expressions, which is difficult to achieve on real-world datasets.

Therefore, the proposed approach and architecture can be used further in future applications for learning effective speech and visual representations. Studying the effect of these representations in a fusion approach using collaborative learning will be interesting. Thus, it is kept as a future work and as an aspiration to further expand the capabilities of speaker detection and identification, speech-to-text, speech to visual expressions, along with other applications that could solve the real-world problems

# Bibliography


[1] Dialoguernn: An attentive rnn for emotion detection in conversations. 33:6818–6825, 2019.

[2] Mimamo net: Integrating micro- and macro-motion for video emotion recognition. 34:2621–2628, 2020.

[3] Evaluation of error-and correlation-based loss functions for multitask learning dimensional speech emotion recognition. In *Journal of Physics: Conference Series*, volume 1896, page 012004. IOP Publishing, 2021.

[4] Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121, 2020.

[5] Yasmina Al Khalil, Sina Amirrajab, Cristian Lorenz, Jürgen Weese, Josien Pluim, and Marcel Breeuwer. On the usability of synthetic data for improving the robustness of deep learning-based segmentation of cardiac magnetic resonance images. *Medical Image Analysis*, 84:102688, 2023.

[6] Rami Al-Rfou, Dokook Choe, Noah Constant, Mandy Guo, and Llion Jones. Character-level language modeling with deeper self-attention, 2018.

[7] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou, and Ioannis Giannoukos. Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011. *Artif. Intell. Rev.*, 43(2):155–177, 2015.

[8] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale, 2021.

[9] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[10] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations, 2019.

[11] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460, 2020.

[12] Pablo Barros, German Parisi, and Stefan Wermter. A personalized affective memory model for improving emotion recognition. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 485–494. PMLR, 2019.

[13] Pascal Belin, Shirley Fecteau, and Catherine Bédard. Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3):129–135, 2004.

[14] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In J. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann, 1993.

[15] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[16] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

[18] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22243–22255. Curran Associates, Inc., 2020.

[19] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[20] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021.

[21] Yuedong Chen, Xu Yang, Tat-Jen Cham, and Jianfei Cai. Towards unbiased visual emotion recognition via causal intervention. MM '22, page 60–69, New York, NY, USA, 2022. Association for Computing Machinery.

[22] Jaejin Cho, Jesús Villalba, Laureano Moro-Velazquez, and Najim Dehak. Non-contrastive self-supervised learning for utterance-level information extraction from speech. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1284–1295, 2022.

[23] Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4652–4661, 2022.

[24] Sevegni Odilon Clement Allognon, Alceu de S. Britto, and Alessandro L. Koerich. Continuous emotion recognition via deep convolutional autoencoder and support vector regressor. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.

[25] Dongyang Dai, Zhiyong Wu, Runnan Li, Xixin Wu, Jia Jia, and Helen Meng. Learning discriminative features from spectrograms using center loss for speech emotion recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7405–7409, 2019.

[26] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions, 2020.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[29] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[30] Xingping Dong and Jianbing Shen. Triplet loss in siamese network for object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

[32] Paul Ekman and Wallace V Friesen. Head and body cues in the judgment of emotion: A reformulation. *Perceptual and motor skills*, 24(3):711–724, 1967.

[33] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M. Hospedales. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, 2022.

[34] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation, 2019.

[35] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Representation learning for speech emotion recognition. 2016.

[36] M Gopichand, K Rajeswari, and E Deepthi. Human–machine interface for wheelchair control using semg signals. In *Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 2*, pages 395–406. Springer, 2023.

[37] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020.

[38] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

[39] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006.

[40] Bin Han and Hans D. Schotten. Multi-sensory hmi for human-centric industrial digital twins: A 6g vision of future industry. In *2022 IEEE Symposium on Computers and Communications (ISCC)*, pages 1–7, 2022.

[41] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing XU, and Yunhe Wang. Transformer in transformer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 15908–15919. Curran Associates, Inc., 2021.

[42] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy, and Kong-Pang Pun. An efficient mfcc extraction method in speech recognition. In *2006 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 4 pp.–, 2006.

[43] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access, 2018.

[44] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020.

[45] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[46] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[47] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. Ted-lium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In Alexey Karpov, Oliver Jokisch, and Rodmonga Potapova, editors, *Speech and Computer*, pages 198–208, Cham, 2018. Springer International Publishing.

[48] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[49] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization, 2018.

[50] M. Horvat, A. Stojanović, and Ž. Kovačević. An overview of common emotion models in computer systems. In *2022 45th Jubilee International Convention*

*on Information, Communication and Electronic Technology (MIPRO)*, pages 1008–1013, 2022.

[51] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[52] Jongheon Jeong and Jinwoo Shin. Training gans with stronger augmentations via contrastive discriminator, 2021.

[53] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery.

[54] Frédéric Joassin, Mauro Pesenti, Pierre Maurage, Emilie Verreckt, Raymond Bruyer, and Salvatore Campanella. Cross-modal interactions between human faces and voices involved in person recognition. *Cortex*, 47(3):367–376, 2011.

[55] M. KAMACHI. Putting the face to the voice : matching identity across modality. *Current Biology*, 13(19):1709–1714, 2003.

[56] Bulat Khaertdinov, Stylianos Asteriadis, and Esam Ghaleb. Dynamic temperature scaling in contrastive self-supervised learning for sensor-based human activity recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(4):498–507, 2022.

[57] Md Ayshik Rahman Khan, Marat Rostov, Jessica Sharmin Rahman, Khandaker Asif Ahmed, and Md Zakir Hossain. Assessing the applicability of machine learning models for robotic emotion monitoring: A survey. *Applied Sciences*, 13(1), 2023.

[58] Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. Data augmenting contrastive learning of speech representations in the time domain, 2020.

[59] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, Alexandre Kaspar, Mohamed Elgharib, and Wojciech Matusik. On learning associations of faces and voices, 2018.

[60] Jaebok Kim, Khiet P. Truong, Gwenn Englebienne, and Vanessa Evers. Learning spectro-temporal features with 3d cnns for speech emotion recognition. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 383–388, 2017.

[61] Taewoon Kim and Piek Vossen. Emoberta: Speaker-aware emotion recognition in conversation with roberta, 2021.

[62] You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung, Yoohwan Kwon, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Look who's talking: Active speaker detection in the wild, 2021.

[63] Dimitrios Kollias and Stefanos Zafeiriou. Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset. *IEEE Transactions on Affective Computing*, 12(3):595–606, 2021.

[64] Philip A. Kragel, Marianne C. Reddan, Kevin S. LaBar, and Tor D. Wager. Emotion schemas are embedded in the human visual system. *Science Advances*, 5(7):eaaw4358, 2019.

[65] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

[66] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934, 2020.

[67] Sangmin Lee, Hyung-Il Kim, and Yong Man Ro. Weakly paired associative learning for sound and image representations via bimodal associative memory. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10524–10533, 2022.

[68] Yuanyuan Lei and Houwei Cao. Audio-visual emotion recognition with preference learning based on intended and multi-modal perceived labels. *IEEE Transactions on Affective Computing*, pages 1–16, 2023.

[69] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang. Contrastive unsupervised learning for speech emotion recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6329–6333, 2021.

[70] Ming Li, Yusheng Su, Hsiu-Yuan Huang, Jiali Cheng, Xin Hu, Xinmiao Zhang, Huadong Wang, Yujia Qin, Xiaozhi Wang, Zhiyuan Liu, and Dan Zhang. Human emotion knowledge representation emerges in large language model and supports discrete emotion inference, 2023.

[71] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings. International Conference on Image Processing*, volume 1, pages I–I, 2002.

[72] Feng Liu, Luan Tran, and Xiaoming Liu. 3d face modeling from diverse raw scan data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[73] Jiaxing Liu, Yaodong Song, Longbiao Wang, Jianwu Dang, and Ruiguo Yu. Time-frequency representation learning with graph convolutional network for dialogue-level speech emotion recognition. In *Interspeech*, pages 4523–4527, 2021.

[74] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[75] Vasista Sai Lodagala, Sreyan Ghosh, and S. Umesh. Ccc-wav2vec 2.0: Clustering aided cross contrastive self-supervised learning of speech representations. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1–8, 2023.

[76] Fei Lu, Long Zhang, and Guohui Tian. User emotion recognition method based on facial expression and speech signal fusion. In *2021 IEEE 16th Conference on Industrial Electronics and Applications (ICIEA)*, pages 1121–1126, 2021.

[77] Florian Lux, Ching-Yi Chen, and Ngoc Thang Vu. Combining contrastive and non-contrastive losses for fine-tuning pretrained models in speech analysis. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 876–883, 2023.

[78] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617, 2021.

[79] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria. Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowledge-Based Systems*, 161:124–133, 2018.

[80] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.

[81] Hao Meng, Tianhao Yan, Fei Yuan, and Hongwei Wei. Speech emotion recognition from 3d log-mel spectrograms with deep learning network. *IEEE Access*, 7:125868–125881, 2019.

[82] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Yuan Cheng, Meng Wang, Chuanhe Liu, and Qin Jin. Multi-modal emotion estimation for in-the-wild videos, 2022.

[83] Liyu Meng, Yuchen Liu, Xiaolong Liu, Zhaopei Huang, Wenqiang Jiang, Tenggan Zhang, Chuanhe Liu, and Qin Jin. Valence and arousal estimation based on multimodal temporal-aware features for videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2345–2352, 2022.

[84] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[85] Vikramjit Mitra, Vasudha Kowtha, Hsiang-Yun Sherry Chien, Erdrin Azemi, and Carlos Avendano. Pre-trained model representations and their robustness against noise for speech emotion analysis, 2023.

[86] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1359–1367, 2020.

[87] Yujie Mo, Liang Peng, Jie Xu, Xiaoshuang Shi, and Xiaofeng Zhu. Simple unsupervised graph representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7797–7805, 2022.

[88] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, 2022.

[89] Mustaqeem, Muhammad Sajjad, and Soonil Kwon. Clustering-based speech emotion recognition by incorporating learned features and deep bilstm. *IEEE Access*, 8:79861–79875, 2020.

[90] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Learnable pins: Cross-modal embeddings for person identity. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[91] Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, and Alessandro Calefati. Deep latent space learning for cross-modal mapping of audio and visual signals. In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–7, 2019.

[92] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84. Springer, 2016.

[93] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[94] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

[95] Alan V. Oppenheim. Speech spectrograms using the fast fourier transform. *IEEE Spectrum*, 7(8):57–62, 1970.

[96] Andrew Ortony. Are all "basic emotions" emotions? a problem for the (basic) emotions construct. *Perspectives on psychological science*, 17(1):41–61, 2022.

[97] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[98] Reinhard Pekrun, Herbert W Marsh, Andrew J Elliot, Kristina Stockinger, Raymond P Perry, Elisabeth Vogl, Thomas Goetz, Wijnand AP Van Tilburg, Oliver Lüdtke, and Walter P Vispoel. A three-dimensional taxonomy of achievement emotions. *Journal of Personality and Social Psychology*, 124(1):145, 2023.

[99] Liang Peng, Yujie Mo, Jie Xu, Jialie Shen, Xiaoshuang Shi, Xiaoxiao Li, Heng Tao Shen, and Xiaofeng Zhu. Grlc: Graph representation learning with constraints. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2023.

[100] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V. Le. Meta pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11557–11568, 2021.

[101] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Alexander Gelbukh, and Amir Hussain. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018.

[102] R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L. Koerich, Simon Bacon, Patrick Cardinal, and Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2486–2495, 2022.

[103] R Gnana Praveen, Wheidima Carneiro de Melo, Nasib Ullah, Haseeb Aslam, Osama Zeeshan, Théo Denorme, Marco Pedersoli, Alessandro L. Koerich, Simon Bacon, Patrick Cardinal, and Eric Granger. A joint cross-attention model for audio-visual fusion in dimensional emotion recognition. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2485–2494, 2022.

[104] R. Gnana Praveen, Eric Granger, and Patrick Cardinal. Cross attentional audio-visual fusion for dimensional emotion recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021.

[105] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.

[106] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

[107] Morgane Rivière and Emmanuel Dupoux. Towards unsupervised learning of speech features in the wild. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 156–163, 2021.

[108] Anthony Rousseau, Paul Deléglise, and Yannick Estève. TED-LIUM: an automatic speech recognition dedicated corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 125–129, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[109] Heini Saarimäki, Lara Farzaneh Ejtehadian, Enrico Glerean, Iiro P Jääskeläinen, Patrik Vuilleumier, Mikko Sams, and Lauri Nummenmaa. Distributed affective space represents multiple emotion categories across the human brain. *Social Cognitive and Affective Neuroscience*, 13(5):471–482, 2018.

[110] Samik Sadhu, Di He, Che-Wei Huang, Sri Harish Mallidi, Minhua Wu, Ariya Rastrow, Andreas Stolcke, Jasha Droppo, and Roland Maas. Wav2vec-c: A self-supervised model for speech representation learning, 2021.

[111] Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584, 2015.

[112] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition, 2019.

[113] Björn W. Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM*, 61(5):90–99, 2018.

[114] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Gerhard Rigoll, and Andreas Wendemuth. Acoustic emotion recognition: A benchmark comparison of performances. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 552–557, 2009.

[115] Vincenzo Scotti, Federico Galati, Licia Sbattella, and Roberto Tedesco. Combining deep and unsupervised features for multilingual speech emotion recognition. page 114–128, Berlin, Heidelberg, 2021. Springer-Verlag.

[116] Joel Shor, Aren Jansen, Wei Han, Daniel Park, and Yu Zhang. Universal paralinguistic speech representations using self-supervised conformers. In

*ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3169–3173, 2022.

[117] Anuroop Sriram, Michael Auli, and Alexei Baevski. Wav2vec-aug: Improved self-supervised training with limited data, 2022.

[118] Vladan Stojnic and Vladimir Risojevic. Self-supervised learning of remote sensing scene representations using contrastive multiview coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1182–1191, 2021.

[119] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.

[120] Andreas Triantafyllopoulos, Uwe Reichel, Shuo Liu, Stephan Huber, Florian Eyben, and Björn W. Schuller. Multistage linguistic conditioning of convolutional layers for speech emotion recognition. *Frontiers in Computer Science*, 5, 2023.

[121] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[122] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[123] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W. Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap, 2022.

[124] Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, 2021.

[125] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023.

[126] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020.

[127] Xiao Wang and Guo-Jun Qi. Contrastive learning with stronger augmentations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2022.

[128] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: Face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3681–3691, 2021.

[129] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J. Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3661–3671, 2021.

[130] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018.

[131] Wei Xia, Chunlei Zhang, Chao Weng, Meng Yu, and Dong Yu. Self-supervised text-independent speaker verification using prototypical momentum contrastive learning. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6723–6727, 2021.

[132] Chuanguang Yang, Zhulin An, Linhang Cai, and Yongjun Xu. Mutual contrastive learning for visual representation learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):3045–3053, 2022.

[133] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[134] Sung-Lin Yeh, Yun-Shao Lin, and Chi-Chun Lee. An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689, 2019.

[135] Dong Zhang, Weisheng Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Modeling both intra- and inter-modal influence for real-time emotion detection in conversations. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 503–511, New York, NY, USA, 2020. Association for Computing Machinery.

[136] Haoran Zhang, Yuexian Zou, and Helin Wang. Contrastive self-supervised learning for text-independent speaker verification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6713–6717, 2021.

[137] Jinming Zhao, Ruichen Li, Shizhe Chen, and Qin Jin. Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions. AVEC'18, New York, NY, USA, 2018. Association for Computing Machinery.

[138] Liang Zheng, Yali Zhao, Shengjin Wang, Jingdong Wang, and Qi Tian. Good practice in cnn feature transfer, 2016.

# Appendix

## FaceSynthetics Dataset

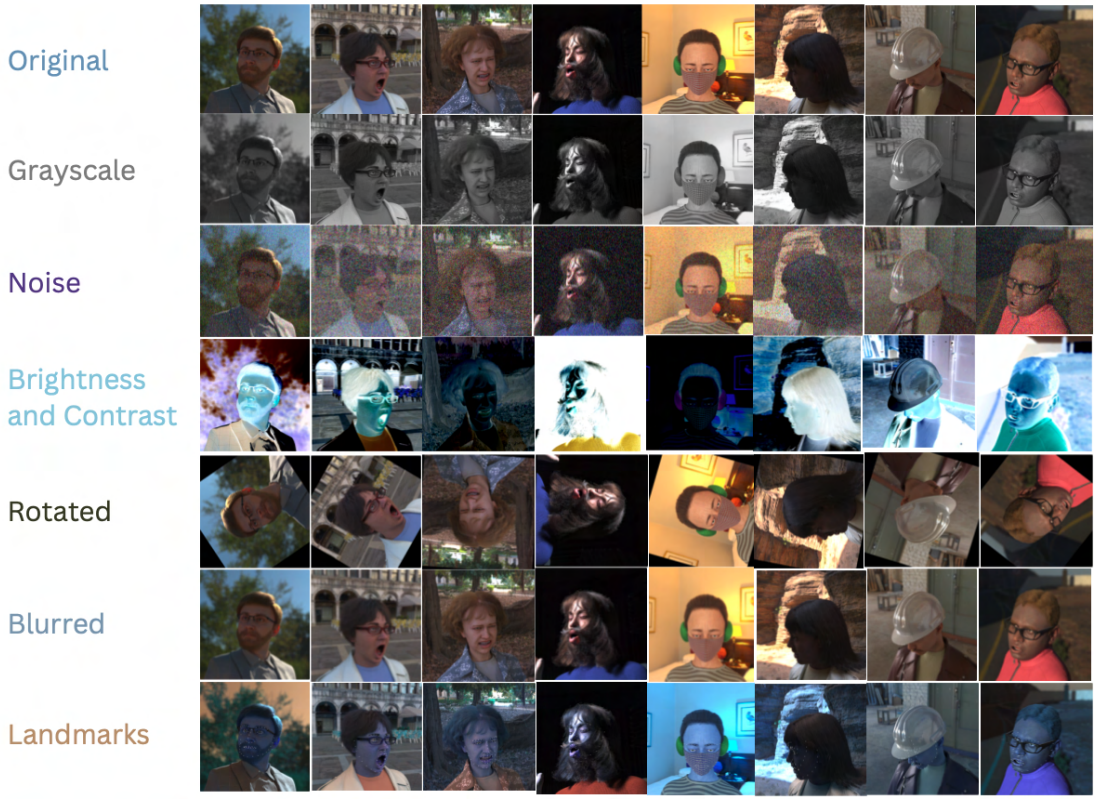


Figure 7.1: Sample images from FaceSynthetics Dataset [128]. From top to bottom, the image shows 6 different types of augmentation including the original image and the image with facial landmark points.

Figure 7.2: Per-pixel semantic class segmentation mapping to the original synthesized image.

Apart from the augmentations already mentioned, facial segmentation images are also used as part of the visual model training as can be seen in Figure-7.2[1].

[1]Figure-7.2 is directly taken from the source: https://github.com/microsoft/FaceSynthetics.

# Erklärung der Urheberschaft

Ich versichere an Eides statt, dass ich die Master thesis im Studiengang Informatik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.
Ort, Datum Unterschrift

**Hamburg, 26.03.2023**

# Erklärung zur Veröffentlichung

Ich erkläre mein Einverständnis mit der Einstellung dieser Master thesis in den Bestand der Bibliothek. Ort, Datum Unterschrift

**Hamburg, 26.03.2023**