Few-Shot Emotion Recognition using intelligent voice assistants and wearables

Learning from few samples of Speech and Physiological Signals

Master Thesis

Mihir Kapadia

Delft University of Technology

TUDelft

Few-Shot Emotion Recognition using intelligent voice assistants and wearables

Learning from few samples of Speech and Physiological Signals

by

Mihir Kapadia

Student Number	5096278	
Project Duration	January, 2021 - March, 2022	
Defence Date:	March 29, 2022	
Supervisors:	Prof. dr. ir. Alle-Jan van der Veen	TU Delft, Supervisor
	Prof. dr. Pablo Cesar	TU Delft, CWI Supervisor
	Dr. Abdallah El Ali	CWI Supervisor

Cover Image: https://www.intelligentliving.co/ai-can-preserve-persons-voice-few-hours-recordings/

TU Delft

Preface

Sometimes, aspirations require a lot of patience and courage to come to life. Our dreams are only as real as the effort we put in to bring them to life. And sometimes, normal things need years and years of waiting to become true. Exactly 10 years ago in 2012, as I prepared to start my bachelor's, I dreamed of pursuing a Master's program in a research university where scientific exploration was the centre of study. This dream has ever since directed my life and career strongly enough for me to make decisions I would not make otherwise. This thesis is the culmination of that small dream of a rather insignificant individual, into reality. While it has been a long and tiring journey to reach this point, it has surely been one of the most fulfilling experiences of my life. This Master's program has taught me things that beyond the courseware. It has also challenged me in newer ways that have made me a better engineer.

I would like to thank all the professors at TU Delft and everyone at CWI to guide me throughout the numerous courses and projects I have done. My special thanks go to Abdallah El Ali for his guidance and tips on writing and forming concrete research ideas. His help and patience have been extremely valuable to me especially in a very difficult year in quarantine. I would also like to thank Prof. Pablo Cesar for accepting me as a thesis student and allowing me to work on the thesis despite poor performances in the initial months of the thesis. Most importantly I would like to thank him for not withdrawing this thesis project from me. Finally, I would like to thank Prof. Alle-Jan for providing me with the support and guidance that helped me through difficult times. Your help and support have been invaluable to my mental health and my ability to work.

This thesis and the master's would not have been possible without the constant, unconditional support and love of my parents. They have spent sleepless nights with me just to be there for me far from home when I had been lonely here. I could not live or sustain the pandemic if it was not for the calls with my parents every day. However late I may be, I have completed this master's only because of their constant motivations. I am forever indebted to them for their unflinching support during difficult times.

Finally, I would like to thank all my friends from the bottom of my heart to motivate me and cheering me up throughout these 2 and a half years in Delft. The journey through this master's has only been possible due to the support of my friends. I am eternally grateful to Brenda, Christian, Felix for providing a roof over my head and being the kindest friends ever. I cannot thank all of them enough - Stefanie, Marta, Gaia, Arda, Nisarg and all my classmates for their continuous check-ins during quarantine. I also thank them for dealing with me through the good, bad and ugly days. Finally, I thank my friends and colleagues in India - Tapasya, Akash, Milay, Rutuja, Sagar, Rajshree and my mentors Swati, Anuj and Esha, who have constantly motivated me to pursue my dreams.

To chasing our dreams!

Mihir Kapadia
Delft, March 2022

Summary

Emotion Recognition is one of the vastly studied areas of affective computing. Attempts have been made to design emotion recognition systems for everyday settings. The ubiquitous nature of Intelligent voice assistants (IVAs) in households, make them a great anchor for the introduction of emotion recognition technology to consumers. The existing systems lack such pipelines and rely on dictionary-based architectures in their design. Further, these systems lack conversational properties and are merely an extension of information retrieval engines.

In this setting, we propose to introduce and develop emotion recognition pipelines that are suited to the interactions, common with these IVAs. To augment the existing emotion recognition pipelines which rely on audio information, we look at physiological information derived from wearables. The design motivations for this combination of data streams stem from the existing examples of such combination of devices such as the Apple Air watch, which combines the capabilities of a wearable sensor with an embedded IVA - Siri.

In this thesis, we try to develop a few shot emotion recognition systems to meet the challenges posed by the devices and user behaviours. Unobtrusive and ubiquitous sensing is utilized to augment the scarcity of samples owing to the short and limited user interaction common with IVAs. This motivates the design choice of few-shot learning algorithms, which rely on a fractional amount of data as compared to common machine learning and deep learning algorithms. Our proposed model uses multimodal embeddings with a Siamese Network to achieve the task of emotion recognition from a few samples. Physiological signals of blood volume pulse (BVP) and electrodermal activity (EDA) are used as additional input embeddings to two audio embeddings arising from the speech samples. We employ the state-of-the-art training schedules for Siamese Networks, which use a very limited amount of training on support datasets via sample pair comparisons.

The proposed model is applied on two datasets that denote two unique experimental settings - the K-EmoCon dataset and RECOLA dataset. We demonstrate an improvement in the state-of-the-art accuracy with the K-EmoCon dataset with accuracies of 63.97% and 66.91% on arousal and valence dimensions respectively. Further, on the RECOLA dataset, the model performs moderately well with 53.81% and 53.87% respectively for arousal and valence dimensions. In addition to this, we present a study of the effects of variation of available support set for training from the dataset. We make some salient observations for these experiments across individual participants and also identify how the label distributions affect the performance of the model. Further, we investigate the impact of real-world noise samples from the DEMAND dataset on the two datasets. We observe that the proposed model is robust and performs sustainingly well even in the presence of imputed noise.

The performance of the proposed model presents new opportunities in the domain of emotion recognition with few-shot learning techniques. This work demonstrates the ability of the Multimodal Siamese Network to predict emotion dimensions even with a limited amount of data. Although satisfactory, the performance does not sustain across the two data sets. We discuss the issues of the proposed model and the potential sources of its moderate performances and motivate the future work for opportunities for improvement. This work contributes to the inquiry of few-shot emotion recognition in everyday household scenarios.

Contents

Nomenclature

Abbreviations

Abbreviation	Definition
IVA	Intelligent Voice Assistants
EDA	Electrodermal Activity
BVP	Blood Volume Pulse
PPG	Photoplethysmography
FSL	Few Shot Learning
DNN	Deep Neural Network
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
LPC	Linear Prediction Coefficients
MFCC	Mel-Frequency Cepstral Coefficient
LPCC	Linear Predictor Cepstral Coefficient
GMM	Gaussian Mixture Models
HMM	Hidden Markov models
ANS	Autonomous Nervous System
CNS	Central Nervous System
SNS	Somatic Nervous System
SCL	Skin Conductance Level
SCR	Skin Conductance Response
MAML	Model Agnostic Meta-Learning

List of Figures

List of Tables

Part I - Introduction

1

Introduction

Affective computing aims at the development of systems and devices that can recognize, interpret, and simulate human emotions. Rosalind Picard [71] states the ability to simulate and adapt emotional intelligence in machines, as the motivations of affective computing. One of the many ways of understanding human emotions is by capturing expressions through facial expressions and voice. Besides such direct emotional attributes, indirect attributes of emotion include body gestures and physiology [89] [109] [86]. Over the last decade, affective computing has found applications in many areas from education, healthcare, virtual and social experiences amongst many. Based on applications, there could be different kinds of agents possible, for example, a virtual chatbot application, speech-based interactive voice assistants, conversational and expressive autonomous robots, to even virtual avatars in a video game. These applications need different abilities expression and perception of human emotions. Their complexity depends on the degree to which they perceive or express or perform both actions. One such application is intelligent voice assistant systems (IVAs).

Intelligent Voice Assistants (IVA) are software-based interaction agents operating through voice commands. These are available on variety of devices from application-specific devices to desktops to home assistants to smartphones. Apple's Siri, Amazon's Alexa, Google's Assistant, Microsoft's Cortana are some of the popular commercial voice assistants currently available [40]. Thus, voice interfaces have become ubiquitous in households and present a great opportunity in understanding the everyday interactions of users. Research on voice interfaces describes the complex relationships between the user(s) and the voice assistant systems [56].This is attributed to the multiple-user interactions, simultaneous commands, use of conversational language [10] [73]. These interactions are short utterances and consist of directed questions. Household applications of IVAs also pose a challenge of combating background noise from the environment to clearly identify the command and decode it.

This dynamic environment of household provides the setup for the problem of emotion recognition in everyday contexts. Here, speech is the primary source of information for inference of emotions from the user. Since IVAs are considerably available across devices, we pose a problem of the integration of physiological information from wearable devices with speech captured by IVAs. Some wearable smartwatches from Apple, Samsung and Fitbit to name a few, already capture physiological information from the user such as heart rate, electrodermal activity and blood volume pulse as a part of the health and fitness applications[84][15]. This persistent availability of physiological information helps in forming multimodal emotion

Figure 1.1: Commercial Intelligent Voice Assistants [24]

recognition systems [109]. The fusion of speech and physiology for emotion recognition has not been examined in conversational contexts [108] in literature.

In this thesis, we propose the use of multimodal information to help augment the traditional speech emotion recognition methods. When designing multimodal emotion recognition systems with IVAs, we aim our focus on two sets of devices – household voice assistant systems (for example Amazon's Alexa) and a smartwatch (for example Apple's Series 7 Smartwatch). Speech and physiological signals are collected from these devices and integrated at a central server (on the smartphone) for emotion recognition. This collective information would thus allow multimodal emotion recognition of the user's interactions in that environment.

In the remaining of the chapter, a brief overview of the workings of an IVA is presented, followed by the current commercial use-cases. Following, a short literature study on the user's interactions with IVAs forms the basis of the motivations of the research. Finally, the research questions and contributions are proposed which address some of the challenges in the current methods used for emotion recognition.

1.1. Intelligent Voice Interfaces

Voice Assistants operate by responding to a specific keyword as a trigger for interaction. Upon receiving the keyword, the IVA sends the voice command of the user to a central server for interpretation [40]. The speech recognition engine sends back the appropriate information to the specific device to perform the task. Many such IVAs are designed to have social interactions with their users including telling jokes or stories. In user research, there are examples of several task-specific IVAs designed for domains like automotive assistance, intelligent tutoring and health monitoring and assistance. These devices have a limited ability of perception as opposed to their commercial counterparts.

The need for an emotionally aware and cognitive voice assistant can be attributed to several applications found in the literature. Healthcare is one area, where conversational agents have been effective in personal assistance for several activities – from self-management such as harmonization of food, exercise and medication to assisting elderly with technology use [76]. Several characteristics requirements of such agents are – natural interaction, personal and situational context, empathy, emotion, multimodal interaction and adaptive responses [54] [47] [96]. These have been shown to affect the level of trust and frequency of interaction of the device [18]. Literature on affect-sensitive educational technologies has moved from traditional passive computer-based training to more intelligent systems which can capture the cognitive states of students while learning [26]. Improvements that have helped students' learning have been attributed to focusing on a natural language dialogue and discourse processing. These systems focus on the detection of basic affect descriptors such as boredom, confusion, frustration, and anxiety as opposed to comprehensive 2-dimensional affect detection [5].

In the seminal paper [22], the authors described possible future use-cases and design considerations of IVAs from the perspective of household use. Here the authors categorize IVAs by their degree of assistance and agency. The state-of-the-art IVAs have achieved a certain degree of 'assistance' in that, they can perform directive tasks. Yet, they lack 'agency' - they cannot yet act on behalf of users or function outside the scope of a command [108]. Although intelligent and adaptable, they are poor in their usability for complex tasks and general inferences [11] [56]. This is because IVAs are incapable of inferring context and user intentions and emotions. Further, current IVAs are far from personalization for a specific user. Personalization can aid in understanding user contexts and emotions from the user's voice and build a cognitive model of the user. Emotion intelligence in IVAs can bridge this gap of 'agency' in IVAs, together with bringing personalization to the inferences of the user.

Before proposing the solution for emotional intelligence in IVA, we examine the nature of the use of IVA and smartwatches. Various factors such as duration of use, interaction settings and environment act as the primary source of constraints for the proposed system. Literature on the everyday use of IVAs shows nuances of interactions of the users with IVAs. A conversationalist approach to IVA design has increased its usability and user experience in commercial devices [110]. They can hold dialogue-like interactions and attempt to provide a sense of companionship. User research of voice interfaces describes the complex relationships between the user(s) and the voice assistant systems. In F. Bentley et al.[11], a study on the use of Google Home is presented by studying the logs of the interactions from 88 households. The use of the IVA is restricted to a few basic domains – namely 'music', 'information seeking', 'automation' and 'small talk'. Most sessions of interaction have a small number of commands, with very few words per command (4.1 on average). Only 10% of the sessions involved more than 10 words per command. This resounds the need of systems which work with very little amount of data. The authors in [73] study IVA interactions at home using Amazon Echo, in several households. Principal observations show the seamless embedding of the IVA within activities as a resource of action. The commands and interactions with Echo are also highly colloquial. This provides the need to test proposed systems on speech samples which exhibit spontaneous interactions between individuals, rather than acted speech. A recurring theme in most of the brown-field user studies is the prevalence of background chatter and noise, particularly in a household setting. While analyzing speech, the IVAs need inherent noise removal algorithms before processing the commands for execution. Therefore, proposed systems must account for robustness against characteristic noises in the speech signal.

Smartwatches distinctively differ from IVAs in the sense that they encounter application micro usage. Visuri et al.[101] summarize their user studies with the following observations. Approximately 64.87% interactions last for less than or equal to 5 seconds, referring to 'peek' sessions, while 35.13% of interactions are longer than 5 seconds referring to 'interaction' sessions; in addition, approximately 82.31% interactions are user-initiated. Further, the mean length of interaction duration is 7.94 seconds for user-initiated sessions. The main interaction categories involve productivity and communication tasks. In McMillan et al.[59], the authors present a similar study on patterns of smartwatch use with a focus on social context, current activity and location. Firstly, when comparing social contexts, domestic use and socializing account for 6% and 15% of the usage. Secondly, comparing the location of use, home use dominates with 44% interactions of use. With a mean usage for the interaction of 6.9 seconds, duration is found like the previous works. The interaction duration derived in the studies motivate the need for analyzing suitable sample segment lengths when designing emotion recognition system.

The results obtained form the user research of IVAs and smartwatches provides directions for implementation of emotion recognition systems using these devices. These factors also contribute to the constraints of the proposed system. In the next section, we present the motivations for the research direction derived from the settings discussed so far.

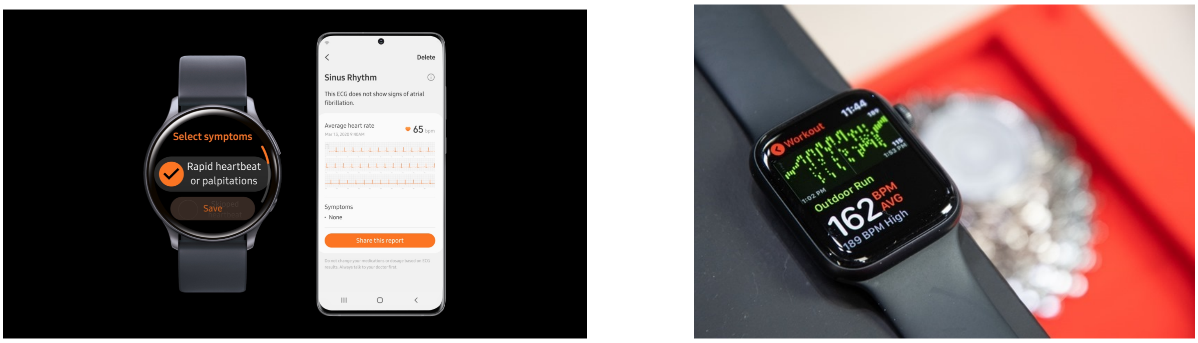
1.2. Motivation

The usage scenarios described in the previous section act as motivations for the proposed system. Several factors are accounted for. Firstly, we restrict the problem to the application of IVAs in a multimodal fashion in a household setting. This means a smartwatch and an IVA device (smartphone or a commercial device such as Amazon's Alexa) act as input sources for physiological and speech signals, respectively. Secondly, the interactions assessed in the problem are natural conversations between the IVA and the user or between multiple users of IVAs. The nature of interactions is restricted to a specified length of 5 seconds or lesser. This is derived from the results of user studies of interaction and command lengths for IVAs and smartwatches discussed previously. The proposed system must operate on very few samples of data with each sample being of this length. Thirdly, the household environment is characterized by noises from the surroundings, which must be corrected by the proposed system.

Having provided an overview of the problem, we aim to solve this using an emotion recognition system that is suitable subject to the factors described above. One of the immediate ways of introducing emotional intelligence is through speech emotion recognition. Additional user information may be derived from physiology captured by wearable devices such as a smartwatch. Evidence shows multimodal emotion recognition systems outperform their unimodal counterparts [25] [69] [74]. In this context, a combination of speech and physiological signals can improve emotion recognition over their unimodal models. Thus, developed multimodal emotion recognition system can aid in deducing user emotions and providing personalization for the user.

For the task of speech emotion recognition, literature provides exhaustive list of methods. Recent developments in large-scale deep neural networks have led to the development of several deep learning techniques which explore specific attributes of speech [75] [57] [79]. On the other hand, emotion recognition from physiological signals is motivated by literature from physiology. Physiological signals such as electrocardiography (ECG), electroencephalography (EEG), electrodermal activity (EDA), among others have been identified as biomarkers for emotional responses [53]. In the current work, we focus on physiological signals readily available with commercial smartwatches. Therefore, we select PPG and EDA signals, which are the most common set of physiological signals found in smartwatches.

The key motivation of the proposed work is that of inference of emotions from a few samples of data. Literature on Multimodal emotion recognition focuses on audio-visual and textual mediums [88] . The focus of such studies has been on emotion classification from the audio-visual medium. Many studies focus on integrating data from video and audio modalities with



(a) Samsung Smartwatch with BVP tracking [30]

(b) Apple Watch with BPM Measurement [4]

Figure 1.2: Commercial Smartwatch with physiological measurement

physiological data collected with medical-grade equipment [89]. These methods need a large volume of data to perform emotion classification. However, very few methods explore to learn from a few samples of multimodal data. In contrast recent developments in deep learning have proposed multiple solutions to this problem. One such paradigm is that of Few-shot learning (FSL). Few-shot learning refers to algorithms which aim at learning from a small set of supervised data samples. Here, the unknown set of samples is considerably larger as compared to the known set of samples.

While few shot learning is widely explored in literature for numerous problems, very little exploration has been done with these methods for emotion recognition problem [105] [19]. Particularly, the wide benefits of FSL problem in multimodal settings have not been explored sufficiently. Literature of multimodal emotion recognition from a few samples is scarce and focuses on facial emotion recognition [21] [46] [52] [50]. Existing work also lacks exploration of noisy speech when dealing with a few samples of data [12]. In addition, the existing body of work lacks studies on the effect of segment length on the performance of a few shot emotion recognition system. These research gaps are explored in our work.

The following summarizes the problem constraints for the proposed work –

- short spontaneous and infrequent utterances of speech
- sample duration of less than or equal to 5 seconds
- very small amount of data used for learning
- use of wearable sensors to facilitate in-the-wild data capture
- noisy speech samples

Summarizing the details of the system requirements, we propose a few-shot multimodal emotion recognition system. We use Siamese Networks [45] as the backbone for the multimodal emotion recognition pipeline. Siamese Networks are powerful metric learning algorithms which require few samples of data to obtain a satisfactory generalization on classification tasks. Multimodality is introduced in the backbone using an independent embedding structure per modality. Embeddings are optimized for the modality they are created for, thereby preserving the characteristics within the modality, while also providing classification. We use mel-frequency spectrograms as one of the embedding inputs owing to their powerful discriminating mapping and robustness against noise. Further, we propose to use eGeMAPS [33] feature set to complement the frequency domain features of mel-frequency spectrograms with time-domain features. Finally, we propose use of deep Gated Recurrent Unit (GRU) embedding for physiological signals to embed time-recurring EDA and PPG signals. Siamese Networks work on the basis of discrimination between similar and dissimilar pairs and hence, the loss function for this architecture is different. Here, contrastive loss is used as a cost function for optimization. The performance of the proposed system is therefore quantified mainly by loss rather than accuracy or squared error metrics. This is because the loss gives a direct measure of the performance of the system.

The effectiveness of the proposed system is tested on real-world dyadic interactions. This is achieved by testing on numerous participants from two multimodal emotion recognition datasets which posit two settings described in the previous section. We use K-EmoCon [70], a dataset with dyadic debate setting; and RECOLA [81], another dataset with spontaneous conversation setting. To simulate the real-world setting, speech signals from these datasets are overlaid with four classes of real world noise samples namely - *living*, *kitchen*, *office* and *hallway* (names corresponding to the setting in which they are captured) from from the DEMAND dataset [97], to simulate the natural household setting. This thesis provides empirical findings of the proposed architecture for different segment lengths as well as different amounts of samples used in training the model. This study contributes to identifying the effect of length of input sample, number of samples and background noise on performance of the system. In the next section, we present the research questions for this thesis, followed by research contributions.

1.3. Research Questions

We present a problem of few-shot multimodal emotion recognition. In the proposed setting speech and physiological signals are obtained from commercial microphones and Empatica E4, respectively. These signals are fed to the multimodal emotion recognition set-up to learn from a few samples of data. To simulate this setup, we use signals from 2 datasets and perform an Oracle-based learning approach. Emotion prediction by a few-shot learning pipeline is trained and validated per participant on 2 datasets proposed in the previous section. Our goal is to generalize emotion recognition using a few-shot learning models. With the motivation presented in the previous section, the thesis intends to answer the following research question :

How can we use few samples of EDA, PPG and speech signals derived from interactions with an IVA, to perform emotion classification ?

This question is investigated using the following sub-questions :

1. How to integrate information from speech and physiological signals - PPG and EDA, to perform emotion recognition from a few samples of data?
2. How many data samples are needed to achieve the state of the art results using the integrated speech and physiological signals?
3. What is the effect of segment length of samples on the performance of the multimodal emotion recognition algorithm?
4. How does noise from real-world settings affect the performance of the multimodal emotion recognition algorithm?

1.4. Research Contributions

The following are the expected contributions of the thesis work:

1. Build a multimodal few-shot learning model for emotion recognition with speech, PPG and EDA signals.
2. Provide empirical findings of performance of the system against 6 different sizes of training set.
3. Provide empirical findings of performance of the system against 2 different audio lengths based on the annotation lengths of the two datasets (1s and 400ms).
4. Provide empirical findings of performance of the system against 4 different background noises.
5. Compare the performance of the proposed system with state-of-the-art results.

1.5. Outline

In this chapter we introduced the problem of few shot emotion recognition for the application of IVAs. The specificity of usage of IVAs and smartwatches provide the basis for the motivations of the research described in 1.1. Thereafter, motivations for the proposed work is presented in 1.2 leading to the research questions presented in section 1.3. Finally, the research contributions are presented in 1.4.

The remaining of this thesis is organized as follows. In Part II, an in-depth literature review of theories of emotion, and provide state of the art methods for emotion recognition and deep learning techniques are popular. In Chapter 2 a background on the problem of emotion recognition is established, detailing the theories of emotion and verbal as well as non-verbal emotion expressions. In Chapter 3, emotion recognition methods are presented with a focus on speech, physiological signals and combined modalities. This includes the state-of-the-art methods available currently. In Chapter 4, we describe some powerful few-shot learning techniques, which have found recent use in data-starving deep learning problems. This concludes the literature review.

In Part III, the proposed methodologies and setups are discussed. In chapter 5, the selection of datasets is presented together with pre-processing requirements. A detailed account of the proposed architecture is provided in chapter 6. Here, the model together with suggested embedding from the different modalities is discussed. Finally, the description of fusion and loss criteria for the deep-learning model developed so far is provided.

Part IV discusses the experimental setup and results of the proposed methodology. This is divided into two tasks according to the two datasets on which the tasks are performed.

Finally in Part V, a conclusion with some discussion points is provided.

Part II - Literature Review and Background

2 Background

This chapter provides the required background for establishing the emotion recognition problem and provides the preliminary definitions for the deep learning concepts introduced in the later chapters. Firstly, section 2.1 gives a short literature review on theories of emotions. Two sets of theories are discussed and motivation for the selection of the desirable theory is presented. Next, in section 2.2 common topics from deep learning are introduced. Starting with the perceptron in sub-section 2.1, a basic deep learning network is explained. Thereafter, various useful activation layers are discussed in sub-section 2.2.2. Next in sub-section 2.2.3, the concept of convolutional neural networks is introduced with its component layers. Finally in sub-section 2.2.4, the concept of recurrent neural networks is presented. Overall, the section explains the basic building blocks useful for the architectures discussed in the thesis.

In the next section 2.3, the concept of few-shot learning is introduced and explained together with its mathematical formulation in sub-section 2.3.1. The concept of distance learning methods that form the basis of the proposed architecture is presented in sub-section 2.3.3.

2.1. Theories on Emotions

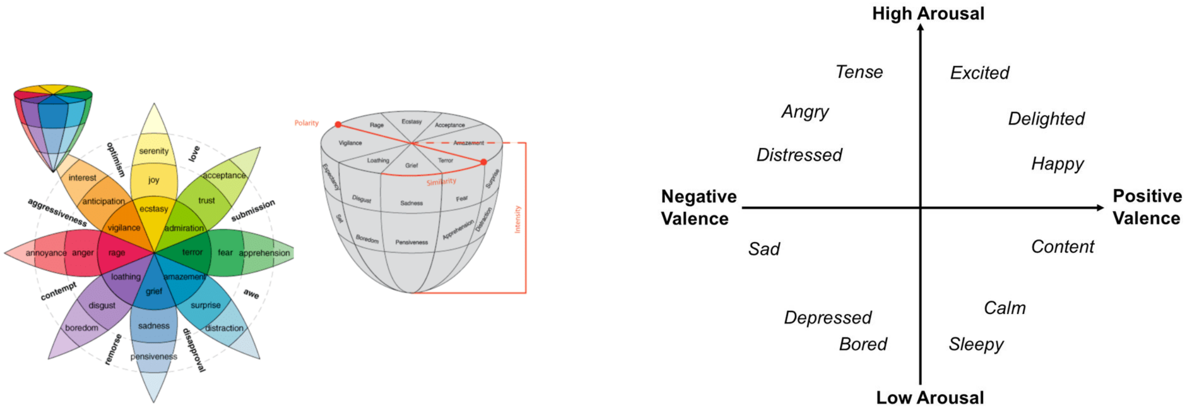
Emotions are a result of internal and external psychological reactions to events. For automated emotion classification, the definition of emotion can be narrowed down to the following – "*a response of the organism to a particular stimulus (person, situation or event). Usually it is an intense, short-duration experience and the person is typically well aware of it*". The description of emotions requires some kind of a model to measure. Literature from psychology broadly categorizes the available emotion models into two categories [69] [1] – Discrete emotion theory and Dimensional emotion theory. These are presented briefly in the next section.

Discrete Emotion Theory

Discrete Emotion theory [23] puts emotions in discrete categorizations. It asserts the universal and biological invariance of emotions in all humans. Several models fall within this theory. One of the most popular models is *Plutchik's wheel of emotions*. Plutchik's Wheel Model [72] maps a set of basic emotions on a wheel and provides an added dimension of intensity. The basic emotions of this model are – joy, trust, fear, surprise, sadness, anger, disgust and anticipation.

Dimensional Emotion Theory

Dimensional Emotion Theory [35]formulates emotions on one or more dimensions in space. This is usually a smaller number of latent dimensions in a continuous space. This provision allows for similar emotions to exist in space nearby, while also providing a measure for differentiation and intensity. One of the most common dimensional models is *Russel's Circumplex model*. Russel's Circumplex Model [82] uses 2 dimensions to describe an emotion – arousal and valence. Valence refers to the nature of the emotion – either positive or negative, and arousal refers to the intensity of the said emotion – either calm or excited.



(a) Plutchik's Wheel of Emotions [72]

(b) Mehrabian's PAD Model [60]

Figure 2.1: Russel's Circumplex model of emotions [82]

Choice of Emotion Models

To select an emotion model, we compare the two theories discussed. Discrete emotion theory has several shortcomings. Firstly, it uses words as descriptions of emotional experiences. This may result in vague translations in case of complex emotions or with different languages and cultures. Secondly, there is a possibility of overlapping emotion categories captured with the facial expressions, speech and physiology of individuals. Finally, categorical labels to emotions can be very idiosyncratic. On the other hand, the dimensional emotion theory has several issues too. For instance, it is not intuitive. In some cases, there is also the ambiguity of axes. However, continuous emotion spaces restrict the task of automated emotion recognition within a defined search space and make the problem tractable. We, therefore, select Russel's Circumplex Model of emotions for our experiments and analysis.

2.2. Deep Neural Networks

2.2.1. The Perceptron

A perceptron is a basic unit of neural networks. A basic perceptron is shown in Fig. 2.2. Here x_i represents the inputs, $w_{i,j}$ represents the weights for the corresponding inputs i. The weighted inputs are added together to get a network input shown as net_j. This is transformed through a nonlinear activation function with a threshold θ_j. The result is the output o_j of the perceptron. Mathematically, the above can be described by the following equation 2.1 :

$$\begin{gathered} net_j = \sum_{i=0}^{n} x_i w_i \\ o_j = \phi(net_j) = \left\{ \begin{array}{ll} 0 & \text{if } net_j \leq \theta_j \\ 1 & \text{if } net_j \geq \theta_j \end{array} \right. \end{gathered} \tag{2.1}$$

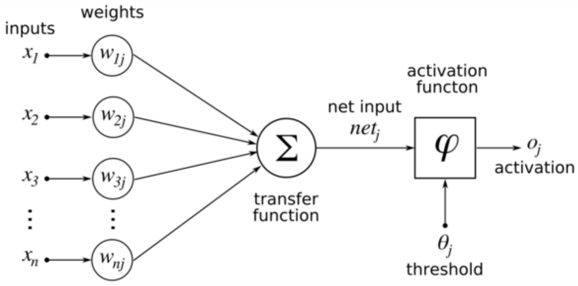


Figure 2.2: Perceptron

Deep Neural Networks (DNNs) (also known as *feedforward neural networks*) are multilayer perceptrons with a large number of hidden layer units. Multilayer Perceptrons (MLP) are networks consisting of an input layer, a hidden layer and an output layer. These can be thought of as composed of multiple perceptrons (or neurons) cascaded end to end. A schematic diagram of a multilayer perceptron is shown in Fig. 2.3. Theoretically, this structure can represent any nonlinear function with appropriate architecture and weights. Training a multilayer perceptron refers to finding appropriate weight values for various connections between the layers.

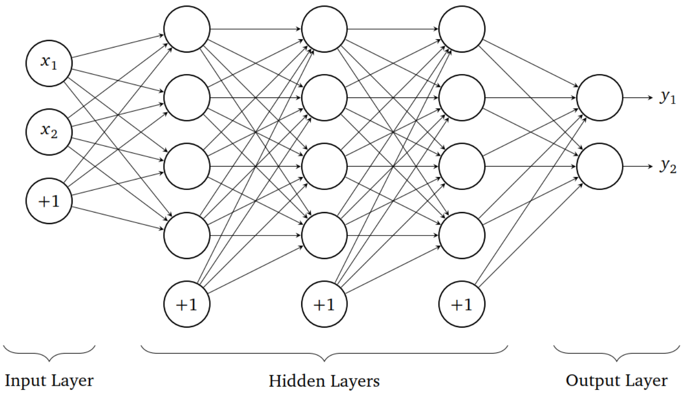


Figure 2.3: Deep Neural Network (Multilayer Perceptron)

2.2.2. Activation Layers

Activation Layers are non-linear functions that map the sum of the weighted inputs to produce the required output. Sigmoid, ReLU, tanh are commonly used activation functions used in neural network architectures which are shown in Fig. 2.4.

ReLU

ReLU stands for Rectified Linear Unit. ReLU functions are easily optimizable, and have properties of linear activation, keeping gradients large when the inputs are positive. For any input $x < 0$, the activation is 0 otherwise the output is x. It is commonly used across most neural network architectures owing to its properties and less susceptibility to vanishing gradients.

Sigmoid

Sigmoid activation transforms any real-valued input to a $[0, 1]$. This activation is useful in generating probability distributions in network outputs. For a single neuron, a sigmoid activation is a transformation of the weighted linear combination of inputs of the neuron to a probability between $[0, 1]$.

Tanh

The tanh activation function is the hyperbolic tangent. It is similar to the sigmoid activation layer with the difference in output mapping to $[-1, 1]$. Tanh activation is useful when designing recurrent neural networks.

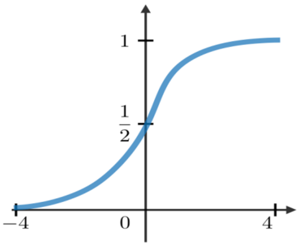
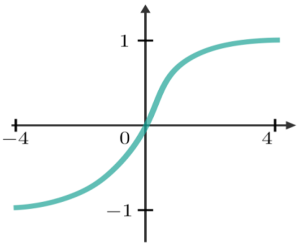
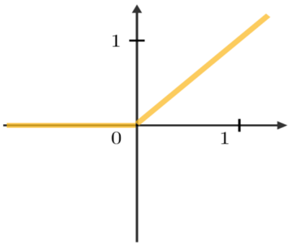
Sigmoid	**Tanh**	**RELU**
$g(z) = \dfrac{1}{1+e^{-z}}$	$g(z) = \dfrac{e^z - e^{-z}}{e^z + e^{-z}}$	$g(z) = \max(0,z)$
		

Figure 2.4: Activation Layers

2.2.3. Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of DNNs that operate on images as inputs. These can operate with 2-dimensional inputs (typically images). CNNs are characterized by specific architectural layers. These include the *convolutional layer* which generates the convolved feature maps; the *pooling layer* which generates averaged outputs from the generated feature maps and the *fully connected layer*, which acts as the learning layer for the pooled feature maps. These are discussed in the following sections.

Convolutional Layer

The convolutional layer maps the input layer through convolution across its dimensions to extract local features. The convolution operation involves a kernel (also known as a filter) of a defined dimension). The kernels act as learnable units with a typically small width and height. The kernel acts as a medium to learn the spatial position of features in a 2-dimensional grid (or image). The kernel parses through the image at each point generating a sparse convolution map. A sample operation is shown in Fig. 2.5. The desirable shift is provided through the stride parameter which measures the number of units shifted per convolution operation. To prevent loss of information due to convolution at edges, padding of values is applied. The density of features to be learned may be specified through the number of kernels since the maps thus, generated will be stacked depth-wise. The activation layer follows the convolution layer to generate feature maps. Therefore, a convolutional layer can be characterized by :

- kernel size : the size of feature maps to be generated
- number of kernels : depth of feature maps
- stride : shift per convolution operation in a specific direction
- padding : addition of values around input edges to maintain the size of output after convolution.

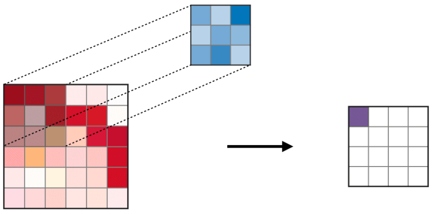
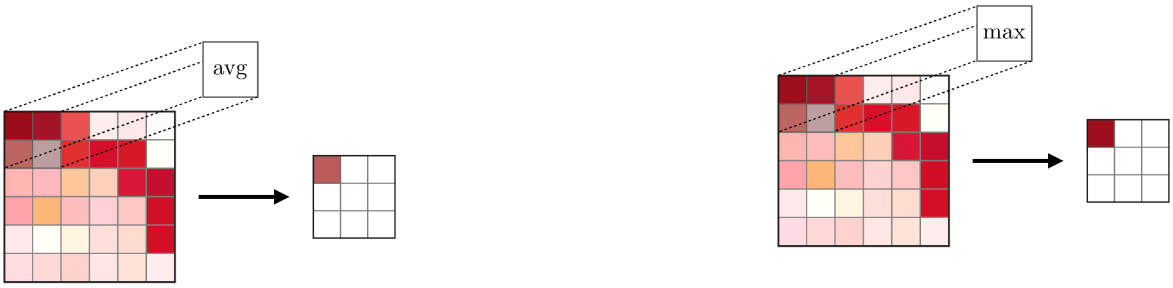


Figure 2.5: Convolutional Layer

Pooling Layer

The pooling layer generates down-sampled feature representation through the reduction of spatial dimensions of the forthcoming layers. As described in Fig. 2.6 this is done either by retaining maximum-value (in case of max-pooling) or averaging (in case of average pooling) of input kernel values. It is characterized by :

- stride: similar to convolution operation, it decides the shift per pooling operation.
- kernel size: the desirable size of the succeeding layer



(a) Average Pooling

(b) Max Pooling

Figure 2.6: Types of Pooling

Fully connected Layer

A fully connected layer is a composite multilayer perceptron with all connections between the neurons activated. Generally, this layer is preceded by a flattening layer (as in Fig. 2.5), which converts the down-sampled 2-dimensional feature maps generated by the max pool layer to a single long vector. This vector can then be used for the activation of class probabilities to predict the output.

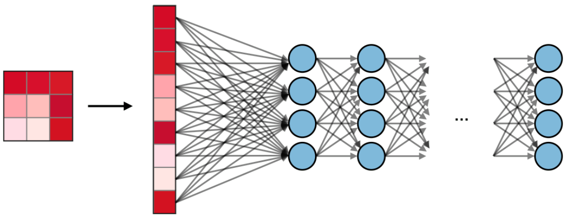


Figure 2.7: Fully Connected Layer

2.2.4. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of neural networks which are used in processing of sequential data. These are characterized by their ability to allow previous outputs to be utilized inputs in successive time stamps. For each time stamp t, the activation $a^{<t>}$ and the output $y^{<t>}$ are expressed as :

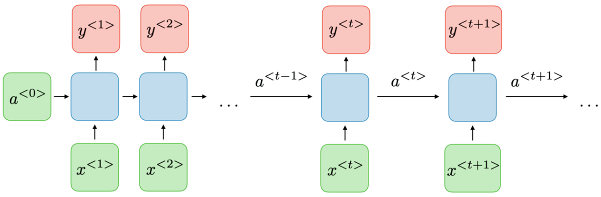


Figure 2.8: Recurrent Neural Network

$$\begin{aligned} a^{<t>} &= g_1(W_{aa}a^{<t>} + W_{ax}x^{<t>} + b_a) \\ y^{<t>} &= g_2(W_{ya}a^{<t>} + b_y) \end{aligned} \tag{2.2}$$

Where $W_{aa}, W_{aa}, W_{aa}, W_{aa}, b_a, b_y$ are coefficients that are shared temporally and g_1, g_2 are activation functions. A schematic of a typical RNN unit is shown in Fig. 2.8. RNNs have the advantage of processing time series inputs of any length while accounting for historical information. Generally, RNNs are created using a gating structure that represents the operations shown in the equation 2.3, where W, U, b are coefficients specific to the gate and σ is the sigmoid operation. These include update, relevance, forget and output gates.

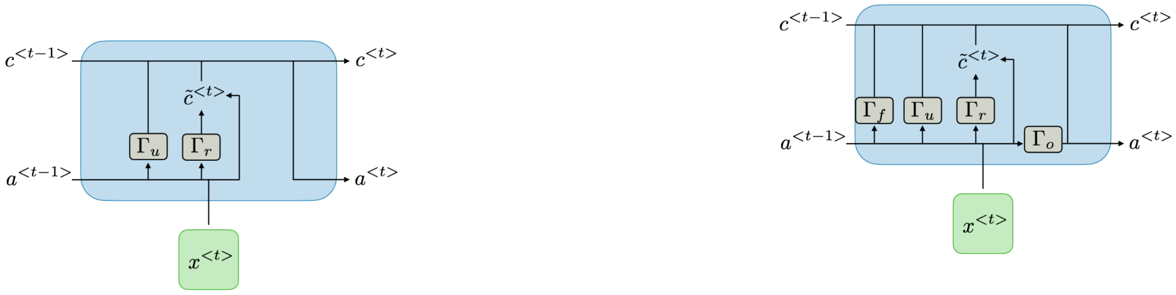
$$\Gamma = \sigma(Wx^{<t>} + Ua^{<t-1>} + b) \tag{2.3}$$

GRU

Gated Recurrent Units (GRUs) are specialized RNN units that do not contain the forget and output gates. This simplifies their structure and computation time, as there are lesser number of weights to be propagated per optimization operation.

LSTM

Long Short-Term Memory units (LSTMs) are RNNs that have all the gates of the traditional RNN unit. They are the more generalized version of GRUs.



(a) GRU unit

(b) LSTM Unit

Figure 2.9: RNN Types

2.3. Few Shot Learning

Few Shot Learning (FSL) is a class of techniques involving learning from a limited amount of samples of supervised data. In addition to limitations of supervised data, it is also useful in cases where there is a lack of annotated data due to extensive costs of annotation; or scarcity of data from rare classes. Few shot learning reduces the burden of data annotation and makes learning inexpensive. It also provides a robust framework to deal with rare classes. These benefits make Few Shot Learning a powerful technique in many applications today. In this thesis, the data constraints require learning emotions from very few samples. Hence, FSL is useful in designing the framework.

2.3.1. Mathematical Formulation

Consider a classification task T for which we have a data set $D = \{D_s, D_q\}$, spanning in a hypothesis space H with a training set $D_s = \{(x_{support}, y_{support})\}_{i=1}^{I}$ where I is small, and a testing set $D_q = \{x_{\text{query}}\}$. Let (x, y) denote an input-output pair from the dataset D and If $\hat{h}$ represents the optimal hypothesis from x to y. A Few Shot Learning problem is an optimization problem of finding an approximate function $h^* \in \mathcal{H}$ of the optimal hypothesis $\hat{h}$ spanning the dataset D. This is achieved by parameterizing the hypothesis as $h(\cdot;\theta)$ where θ denotes all the parameters used by h. An optimization space schematic for the above described few shot learning problem is shown in Fig. 2.10. The optimization cost of FSL problem is measured by a loss function $\ell(\hat{y}, y)$ defined over the prediction $\hat{y} = h(x;\theta)$ and the observed output y. The performance metric of the task T is denoted by P through the extracted knowledge E. Here, the FSL optimization algorithm is encapsulated by the knowledge E, which contains a small number of supervised information samples of T.

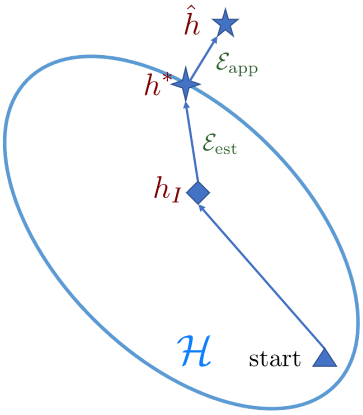


Figure 2.10: Illustration of the Few Shot Learning Problem (adapted from [105])

In literature, FSL problems are defined with two parameters. These are the number of classes N and the number of samples per class available for training K. We formally define a N-way-K-shot learning problem as one in which D_s contains $I = K \times N$ examples from N classes each with K examples. A few shot classification problem learns classifiers given only a few labelled samples of each class, and a few shot regression problem estimates regression function from data, given only a few input-output examples pairs from that function. If the number of samples in the training scheme with E is only one, then the problem becomes a one-shot learning problem. Alternatively, if there are no supervised training pairs in E, the problem is termed a zero-shot learning (ZSL) problem. With zero-shot learning problems, additional information about tasks is needed from auxiliary channels. In this thesis, we strictly consider the size of the training space E to be greater than zero.

2.3.2. Methods

Literature on few-shot learning has varied taxonomy for different approaches of few-shot learning. While there are no formal definitions to classify a method as a few shot learning model, several reviews try to address this. In Wei-Yu Chen et al.[19], the authors classify existing methods into four categories - initialization based, metric learning based, hallucination based and domain adaption based methods. In Y. Wang et al.[105], the authors take a fundamental approach to differentiate methods by the aspect of problem augmented – Data, Model and Algorithm. Yet another categorization is presented in [41]. Without referring explicitly to any specific classification, we can summarize literature into four groups given in Fig. 2.11. In this thesis, the goal is to obtain a novel emotion classifier using speech and physiology.

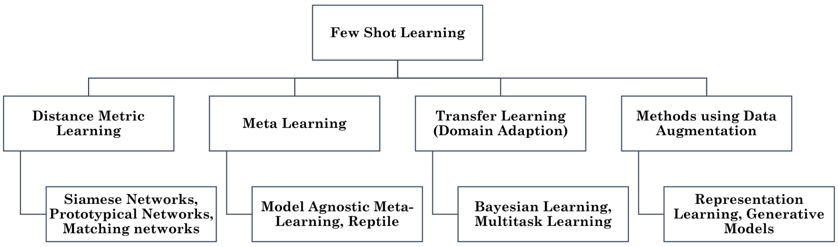

Few Shot Learning
Distance Metric Learning
Meta Learning
Transfer Learning (Domain Adaption)
Methods using Data Augmentation
Siamese Networks, Prototypical Networks, Matching networks
Model Agnostic Meta-Learning, Reptile
Bayesian Learning, Multitask Learning
Representation Learning, Generative Models

Figure 2.11: Snapshot of Few Shot Learning Methods (adapted from [19, 105, 41])

Of the several class of few shot learning techniques shown in Fig. 2.11, we refrain from using techniques which use data augmentation techniques in order to increase data samples. This is because low sampling rates of the physiological signals from the wearables result in disproportionate effects of augmentation techniques. In addition, the specificity of application requires realistic elicited speech corpus with a specific context. This narrows the datasets available for analysis. Due to limited corpus' which fulfil the criteria, transfer learning architectures are not suitable. Therefore, we constrain the survey (and use) on the methods where the learning is achieved through the comparison of data samples. Such methods are called Distance Metric Learning methods. Additionally, the focus of comparison is the task of binary arousal and valence discrimination. In the next section, a short introduction of popular Metric Learning methods is provided.

2.3.3. Distance Metric Learning Methods

Distance Metric learning methods refer to the class of algorithms which use an abstract embedding space as a region of measuring distances between samples. Here, instead of prediction of labels, the goal of the algorithm becomes that of prediction of this distance. Various encoding schemes may be used to achieve a coherent embedding space. Optimization improves the ability of embeddings to provide sufficient distance between different classes. Considering the earlier defined task setting, we search for embeddings for each sample $x_{support} \in \mathcal{X} \subseteq \mathbb{R}^d$ onto a lower-dimensional $z_i \in \mathcal{Z} \subseteq \mathbb{R}^m$, such that similar samples are in proximity while dissimilar samples are far apart. A simple overview of the method is shown in Fig. 2.12. Mathematically, the distance metric learning problem can be composed of the following [105] :

1. an function f for test sample $x_{query} \in D_q$ to $\mathcal{Z}$,
2. an function g for training sample $x_{support} \in D_s$ to $\mathcal{Z}$,
3. a similarity function $s(\cdot, \cdot)$ which measures the similarity between $f(x_{query})$ and $g(x_{support})$ in $\mathcal{Z}$.

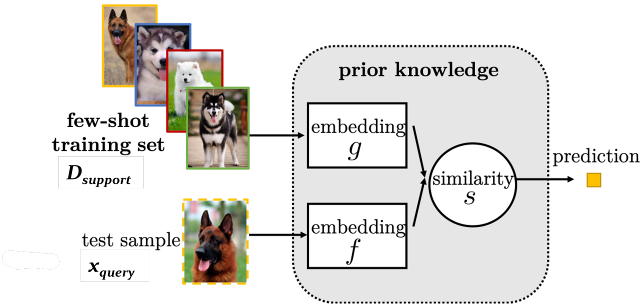


Figure 2.12: Schematic of Distance Metric Learning Methods

The choice of mapping function is varied and motivated by the task in question. A few powerful distance metrics for comparison, have been identified – ℓ_1 distance (Manhattan Distance), ℓ_2 distance (Euclidean Distance), cosine similarity shown in Fig. 2.13. These are defined below:

$$d_{l_1}[x_{s,nk}, x_{q,nk}] = \sum_{n=1}^{N}\sum_{k=1}^{K} ||f(x_{s,nk}) - g(x_{q,nk})|| \tag{2.4}$$

$$d_{l_2}[x_{s,nk}, x_{q,nk}] = \sum_{n=1}^{N}\sum_{k=1}^{K} \sqrt{||f(x_{s,nk}) - g(x_{q,nk})||^2} \tag{2.5}$$

$$d_{cosine}[x_{s,nk}, x_{q,nk}] = \frac{f[x_{s,nk}]^T g[x_{q,nk}]}{||f[x_{s,nk}]|| \cdot ||g[x_{q,nk}]||} \tag{2.6}$$

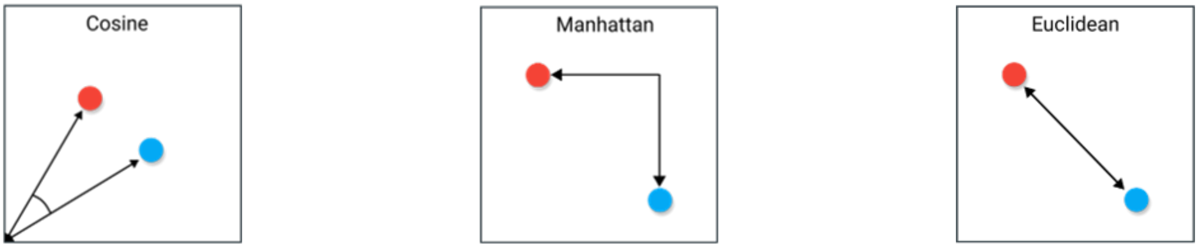


Figure 2.13: Distance Metrics

2.4. Conclusion

In this chapter, the background concepts behind this thesis is introduced. Firstly, various theories for description of emotions for affective computing tasks were introduced. The parameters that lead to the choice of the dimensional emotion model are discussed. Next, a brief introduction to deep learning concepts and architectures is provided. An overview of various activation functions used in this thesis is also given. This includes the typical structures of popular architectures such as the Convolutional Neural Networks and Recurrent Neural Networks is presented. Finally, the paradigm of few-shot learning is presented. The mathematical formulation of the $N \times K$ few-shot learning is given within an abstract data space. Distance learning methods are found to be useful methods for this thesis. This class of methods is explained briefly. This concludes the background discussion.

3

Emotion Recognition

The emotion recognition problem is one of the most active fields of research in affective computing. It is a problem of inferring emotions given a sample of data. This data can either be unimodal or multimodal. It can be defined as a machine (or deep) learning problem where the inputs are the features from the modalities and the labels of emotions are outputs. Emotion Recognition research has traditionally focused on single modalities. Early research focused on facial expressions using images; thereafter broadening to audio-visual modality. In the past decade, efforts to collect and study emotion in different modalities resulted in a plethora of multimodal datasets. Such datasets include audio-visual, textual, physiological, gesture, human activity information. This literature forms the background for this chapter.

In this chapter, we present an overview of emotion recognition with IVAs using speech and physiological modalities. In 3.1, an overview of speech emotion recognition literature is provided. Firstly, we present the nuances of emotional expression in a speech in the section 3.1.1, while also exploring elicitation methods. Then in section 3.1.2, audio features are discussed in depth to provide the basis for recognition method pipelines. Finally, in section 3.1.3, classification methods for speech features are presented. This concludes the survey on speech emotion recognition.

In section 3.2, an overview of physiological emotion recognition literature is provided. Here, section 3.2.1 explains the biological background of emotions with physiology. Thereafter, section 3.2.2 presents common physiological features for the selected signals of PPG and EDA. Finally, in section 3.2.3, classification methods for physiological features are discussed. This concludes the section on physiological emotion recognition.

Finally, in section 3.3, few shot emotion recognition literature is examined. This concludes this chapter.

3.1. Emotion Recognition with Speech

Speech emotion recognition systems are widely studied in the literature. A typical pipeline involves a series of steps as shown in Fig. 3.1. Firstly, audio samples are normalized and cleaned for anomalies. Raw speech is also cleared for noise. This also involves the segmentation of audio signals in frames of a typical length of 20-30ms. Thereafter, feature extraction techniques are used to obtain meaningful representations of audio signals. These features may comprise either numerical or image-based features (in the case of spectrograms). In some cases, for a large number of features, some feature selection is also implemented using suitable algorithms. Finally, classification algorithms are used on the selected set of features to learn a transformation function between the speech features and annotation labels. For the case of emotion recognition, these labels are arousal and valence. In this section, these steps are surveyed from literature and important features and classification techniques are pointed out. Firstly, in section 3.1.1, emotional expressions from the speech are discussed, presenting important speech attributes to be considered while conducting an emotion recognition task. Next, popular audio features are identified from literature in 3.1.2. Finally in section 3.1.3, classification methods for speech emotion recognition are provided.

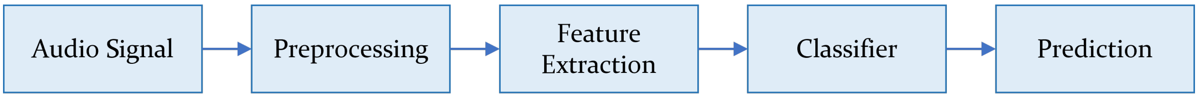


Figure 3.1: Typical Speech Emotion Recognition Pipeline

3.1.1. Emotional Expressions in Speech

Communication by speech is characterized by verbal and nonverbal means. Verbal communication is linguistic; it communicates through both the meaning of the words and the way they are said [7]. Language forms the basis of verbal communication. Nonverbal communication is para-linguistic; it modifies verbal communication to convey the emotions of the speaker. Vocal and non-vocal qualities such as prosody, stress, intonation add the para-linguistic part of communication.

A study by Knapp and Hall [43] shows para-language as described by two distinct sets of characteristics. These are voice qualities (such as pitch, rhythm, tempo, articulation and resonance) and vocalizations (such as laughing, crying, sighing, belching, etc.). In another study, Trager [98] presents classifications for vocalizations. These are - vocal characterizers comprising of non-language sounds; voice qualifiers of language such as pitch, intensity, extent; and vocal segregators including fillers, silent pauses and hesitation.

Broadly, these non-verbal vocalizations correlate with different kinds of emotions. Prosody features like pitch, rhythm, stress, and loudness are rich sources of affective state [3]. These features are suprasegmental - their interpretation extends over segments or phonemes. Several studies show the use of nonsensical utterances to add to the emotional context besides the lexical and semantic effects [8] [85]. Natural conversations are filled with these characteristic speech events which convey emotions subliminally.

So far, we identified several characteristic features which can act as emotional markers in speech. These form the general set of features present in all kinds of speech. For this thesis, we consider the specific case of the exhibition of emotion in conversations or natural speech, therefore we first identify the possible elicitation methods suitable for our task. This forms the basis of selecting datasets as will be described in Section 4.1.1.

Emotion Elicitation methods with Speech

Emotion elicitation is an important aspect of the emotion recognition task. It describes the environment and methodology used to elicit and trigger emotions for an emotion recognition dataset. This is why it is also a differentiating factor in performing emotion recognition experiments. Three kinds of speech corpora are described in literature [29] [48]. These are explained briefly –

- **Natural Speech**: This is the real speech of an individual with spontaneous and natural emotions. It is usually collected from real-life environments and the speech is neither acted nor enforced. Emotional expression in such speech is usually mild and underlying. Such a corpus usually has an imbalance in elicited emotion classes/space. Moreover, its annotation is subjective.
- **Simulated or Acted Speech**: This kind of speech is professionally acted to express emotions by deliberation. It is usually collected from trained actors and theatre artists. Variations in the expression of emotions are performed to obtain the complete range and space of emotions. While this speech may provide a more pronounced display of arousal and valence, it does not resemble the true emotions conveyed. The annotation of such speech is external.
- **Elicited Speech**: This speech is neither natural nor simulated. Here emotions are induced using stimuli/context. Such interactions exhibit true emotions of the person if such a trigger would naturally occur to the person. Speakers can be made to hold an emotional conversation or provoked to provide a charged response. This kind of elicited speech requires both self and external annotation to differentiate between felt and observed emotion labels.

Inducing natural utterances in elicited speech is important to keep the emotions exhibited, as true as possible. This may be achieved using a ***Wizard-of-Oz*** scenario [9], where an external mediator induces interactions that can generate emotional responses between a group of individuals whose speech is to be examined. Alternatively, there could be other means to induce emotions, such as completing a task interactively with a computer, playing video games with instructions, debating on a specific topic or holding conversations on a topic. In the present problem, the goal is to perform emotion recognition on speech collected with IVAs. Therefore, elicited corpora where conversations would naturally induce emotions would closely resemble the nature of emotions that happen in interactions with IVAs.

Up to this point, we have considered qualitative features of speech that act as emotion markers. Next, we examine these quantitatively and identify a concrete set of features for our analysis of the few-shot emotion recognition task. In the next section, audio features are discussed and commonly available feature sets are presented. Motivation for the selection of features is also built around the specific task of few-shot learning, as features must exhibit information from short segments of speech.

3.1.2. Audio Features

Speech (or Audio signal) is a time-series signal which represents the pressure changes in the air that produce a sound. A typical audio sample is shown in Fig. 3.2 where the y-axis represents the amplitude of the waveform and the x-axis represents time. An audio signal can be represented as a complex signal made of multiple sinusoidal components. Audio is not a wide-sense stationary signal. However, when considered in small segments – referred to as frames, typically 20-30ms, it can be considered as wide-sense stationery. Typically, speech features are quantified in frames or as a whole. This gives rise to local features which describe frame-level changes and global features which are statistics across a set of frames or an utterance. Most of the analysis of audio signals today is performed in the frequency domain due to its superiority in describing and summarizing time-series signals efficiently.

The commercial success of intelligent voice assistants has spawned the field of conversational AI which aims at identifying semantics in a conversation with specific importance to situ-

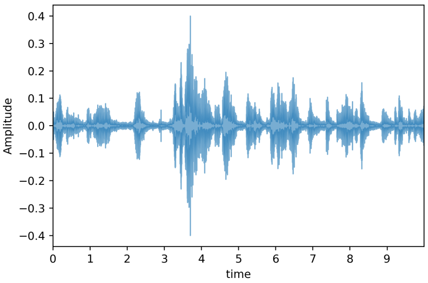
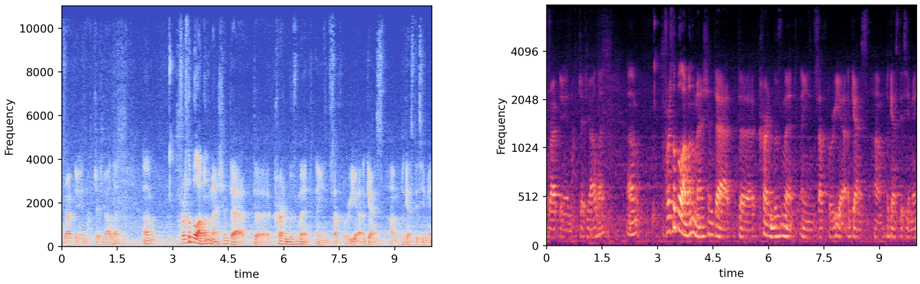


Figure 3.2: Audio Signal

ational context. Acoustic information embedded in time domain, frequency-domain, amplitude domain and spectral energy domains [93] [61] and linguistic information embedded in prosody features [48] [102]. In this thesis, the purpose of speech features is an efficient representation of emotional content in speech *irrespective of the underlying lexical content*. Therefore, we focus on non-lexical features. Literature shows numerous non-lexical audio features used for emotion recognition with different levels of classification based on their domain [83] [1] [99]. Here three classes of speech features are discussed –

- **Prosody features** – Prosody features are such features that can be perceived by humans and describe most of the vocal affective expression. Prosody attributes like duration, intonation and intensity help differentiate between emotional overtones for the same utterances. High arousal emotions such as anger, happiness or surprise are reflected by increased energy of speech signal while disgust and sadness result with decreased energy. Prosody features include the ensemble of features such as - pitch-related features; formants features; energy-related features; timing features; articulation features etc. Some of the important prosody features are the fundamental frequency of the signal, pitch, duration, formant locations and their bandwidths and their derivatives.
- **Spectral features** – Spectral features represent attributes defined in the frequency domain. Spectral features are correlated to the shape of the vocal tract and the rate of change of articular movements. Emotions are described through the embedded spectral energy of the signal. There is a difference in energy distribution across frequency when a person is happy – high energy at higher frequencies; and when an utterance is sad – low energy at the same higher frequencies. Several spectral features have been identified in the literature. Linear prediction coefficients (LPC), least-squares yule-walker equations are some of the earliest known features. The linear prediction coefficients are simply m-order all-pole models of the vocal tract [99]. Spectral band energies, spectral slope and harmonic differences are some of the other spectral features found to have strong correlations to affective arousal. Spectrograms, spectral centroid are the other common descriptors in this class.

  Cepstral features - An optimized representation is obtained if the speech signal is band-passed to constrain the frequency to the audible range of humans. Thereafter, the bandwidths of the filters are scaled to the nonlinear scale of Mel-frequency. This process results in constraining frequency spectra to the auditory canal and thus captures accurately the human auditory perception. This creates a class of frequency coefficients and spectrograms called the Mel-Frequency Coefficients. Consequently, it is possible to obtain the cepstral coefficients – coefficients from a fictional domain inverse of frequency - quefrency. This gives the popular Linear Predictor Cepstral Coefficients (LPCC) and the Mel-Frequency Cepstral Coefficients. Gammatone Frequency Cepstral Coefficients (GFCC) is obtained by applying the Gammatone filter-bank to the power spectrum. Formants are the frequencies of the acoustic resonance of the vocal tract. They are computed as amplitude peaks in the frequency spectrum of the sound.



(a) Power Spectrogram

(b) Mel-frequency Spectrogram

Figure 3.3: Comparison of Spectrograms

The large number of features available for analysis makes the selection of features very important in order to implement a speech emotion recognition pipeline. Literature also provides a wide range of available feature sets for the task of speech emotion recognition. These include *eGeMAPS* feature set, *ComParE2016* feature set, and the *InterSpeech Challenge* feature sets. In this thesis, we propose to use the *eGeMAPSV02* feature set [33] from the openSMILE toolkit [32]. This feature set is selected owing to its compactness and superiority in describing arousal and valence dimensions of emotion. It has been shown to outperform many of the classical feature sets across emotion recognition corpus'. These will be used in the implementation of the feature processing part of the proposed pipelines.

While all the discussed features find relevance, emphasis is provided on selecting features that have been shown to provide satisfactory classification discrimination. For this purpose, classification architectures are surveyed briefly in the next section to identify suitable feature-architecture combinations for the proposed architecture. In the next section, classification techniques for speech emotion recognition are discussed briefly.

3.1.3. Classification Methods

Classification methods for speech emotion recognition have evolved over the past decade from traditional supervised machine learning approaches such as Support Vector Machines(SVM) and Gaussian Mixture Models (GMMs) to more sophisticated time-series representations using Hidden Markov models (HMMs) [99] [29] [48]. With the advancement in deep learning, architectures using Long-short Term Memory (LSTMs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) are becoming extremely popular [94] [63].

Each of the described methods has its pros and cons. Majority of the architectures utilize either prosody or spectral features or a combination of the two for classification. But, their ability to handle many modalities and capture important emotion dynamics make them useful. Using CNNs, audio-visual modalities are combined to obtain useful information for emotion recognition. With RNNs and HMMs, time-varying nature and fleeting emotion can be captured. GMMs provide a mechanism to estimate the probabilistic distribution of an utterance across numerous emotions in a discrete space. Since the application in this thesis focuses on non-lexical speech features, classification techniques that focus on textual interpretations of speech are not discussed here.

The household environment introduces noise from surroundings in IVAs. Noise robustness is thus essential in emotion recognition methods implemented for IVAs. Methods with noisy speech have been extensively studied. The literature presents numerous methods to increase the robustness of the existing set of methods. In [42] authors describe the use of histogram equalization on spectrograms to normalize segments with noise. Similarly, spectral subtraction and filtering have been proposed to remove noise from spectrograms [20]. Several techniques propose augmentation of MFCC features through nonnegative matrix factorization [67], adaptive noise cancellation using channel equalization [95]. Alternatively, several methods operate on raw audio segments such as compressed sensing-based estimation[100], to pre-processing with voice activity detectors [68].

The methods and architectures discussed to provide a very brief and inexhaustive overview of the speech emotion recognition landscape. Within this paradigm, the focus is to obtain features that can adapt to few-shot learning architectures (presented in section 3.3). Before proceeding to few-shot learning architectures, we briefly discuss physiological emotion recognition systems in the next section.

3.2. Emotion Recognition with Physiological Signals

3.2.1. Emotional Expressions in Physiology

Identifying the neurophysiological origin of emotions is a field of active research in psychology and neuroscience. Physiological responses are one of the three indicators of emotional state together with evaluative reports (verbal confirmations and questionnaires) and overt actions (facial expressions, vocal utterances, and bodily gestures) [51]. Everyday human interaction is ingrained with visible (such as heavy breathing in a state of fear or anxiety) or invisible (such as low skin conductance in a state of sadness, relief) physiological overtones. Numerous articles in the literature refer emotions to be autonomic changes under stimuli [53] [49].

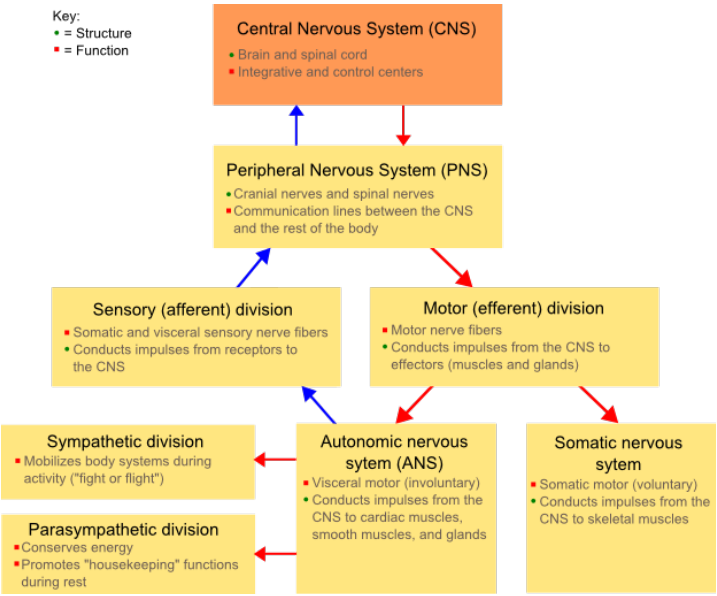

Key:
● = Structure
● = Function
Central Nervous System (CNS)
● Brain and spinal cord
● Integrative and control centers
Peripheral Nervous System (PNS)
● Cranial nerves and spinal nerves
● Communication lines between the CNS and the rest of the body
Sensory (afferent) division
● Somatic and visceral sensory nerve fibers
● Conducts impulses from receptors to the CNS
Motor (efferent) division
● Motor nerve fibers
● Conducts impulses from the CNS to effectors (muscles and glands)
Sympathetic division
● Mobilizes body systems during activity ("fight or flight")
Parasympathetic division
● Conserves energy
● Promotes "housekeeping" functions during rest
Autonomic nervous sytem (ANS)
● Visceral motor (involuntary)
● Conducts impulses from the CNS to cardiac muscles, smooth muscles, and glands
Somatic nervous sytem
● Somatic motor (voluntary)
● Conducts impulses from the CNS to skeletal muscles

Figure 3.4: Taxonomy of Nervous System [55]

Literature from neuroscience points to the Peripheral Nervous System (PNS) as the seat of voluntary and involuntary control of the human body in conjunction with sensed information from a stimulus [16]. This is besides the cognitive responses of the Central Nervous System (CNS) which is responsible for originating an integrated response. Two components of the PNS, branch out – the involuntary functions performed by the Autonomic Nervous System (ANS) and the voluntary functions performed by the Somatic Nervous System (SNS). The Autonomic nervous system (ANS) controls physiological signals. Correspondingly, the respiratory system, cardiovascular system, electrodermal systems and facial motor nucleus are the regions of activation associated with emotional stimuli [44]. Since these are unconscious responses, it makes them devoid of faking, that is there is no difference between the emotions exhibited and experienced through these physiological signals. Further, the exhibited emotion in speech might be deliberately subdued by the user while the emotion from physiological reaction remains uninhibited.

The bulk of the emotion recognition research using physiological signals is found within laboratory settings [13] [28]. Many of the existing studies focus on data collected in isolated environments with clinical instruments. Such measurement methods constrain the collection of physiological signals to a laboratory. While a laboratory setting permits collection of a large volume of high-quality physiological data, it may not exactly allow the replication of natural behaviour and physiology as present in the real-world setting. As established in our problem, the goal of the proposed methods is to identify emotions in a real-world setting. Therefore, it is apt to consider wearables to capture signals unobtrusively and without affecting the user's natural emotions. Therefore, we focus on physiological signals captured by wearables.

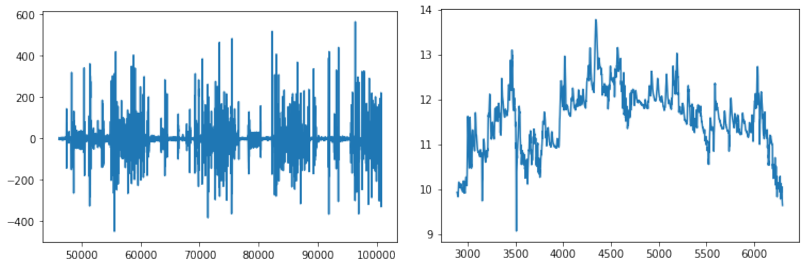
Further, as described in Section 1.2, the physiological signals relevant for this thesis are blood volume pulse and electrodermal activity signals. Commercial devices such as the Empatica E4 can capture both these signals [31]. Apple's and Samsung's smartwatches are a few of the mainstream devices that have started including both these sensors for their health and fitness tracking applications [30] [4]. In the next section, physiological features from the selected two signals are discussed briefly.

3.2.2. Physiological Features

Physiological signals are time-series signals. Unlike audio signals, they have the property of being semi-periodic and partly stochastic. Thus, a small set of features can efficiently describe the characteristics of a time series. Physiological signals are sensitive to acquisition procedures. Wearable-based sensing, adds another challenge to this problem. These signals can be easily corrupted by thermal noise, measurement errors, electromagnetic interference, and motion artifacts. Thus, special attention needs to be paid to the acquisition and preprocessing of the signals to avoid adverse effects. This motivates pre-processing steps. It involves smoothening the signals through low-pass filters, time synchronization and artifact removal. For modern smartwatches and wearables, noise and artifact removal algorithms are usually embedded within the device. Pre-processing and feature extraction steps are physiological signal specific.

PPG-based Features

Photoplethysmography (PPG) signal is measured using a skin-contact based photosensor. An LED illuminates the skin and the photo-diode measures the backscattered light from that patch of skin. Indirectly, this measures the pulsatile component of the arterial blood flow. Since the measurement is contact-based, there are possibilities of motion artifacts and drifts due to too poor contact. This is removed using a high-pass filter or a Butterworth filter. Alternatively, adaptive filters can correct the signal in real-time continuous acquisition scenarios. Typically,



(a) Typical PPG Signal

(b) Typical EDA signal

Figure 3.5: Typical Signals of PPG and EDA

commercial PPG sensors have sampling rates below 100 Hz. PPG-based features can be classified based on the domain of the feature attributes as shown in table 3.1.

Table 3.1: Overview of PPG based features

Feature type	**Example**
Temporal features	statistical mean, median, heart rate (HR), heart rate variability (HRV), statistical features of HRV, number and percentage of RR intervals differing more than 20 ms (NN20, pNN20), Maxima, minima and centroids of frames zero-crossing rate of HR.
Spectral features	Frequency bands of HRV (Ultra Low Frequency, Very Low frequency, Low Frequency, High Frequency), normalized LF/HF ratio, spectral centroid, spectral spread, spectral skewness, spectral kurtosis, spectral slope and spectral variation
Non-linear features	1st and 2nd order Standard Deviations from Poincare plot, sample entropy, approximate entropy, recurrence rate, determinism

EDA-based Features

The electrodermal activity signal is measured using a pair of electrodes to measure the skin resistance between the two bypassing constant current or constant voltage across the skin. This measurement is a function of skin conductance and the number of active sweat glands in the region of measurement. Common EDA sensors are placed at wrists or torso. The EDA signal is a low-frequency signal, therefore some of the noise artifacts are removed by passing it through a low-pass filter followed by smoothening. The EDA signal is made of two components – the skin conductance level (SCL) referring to the baseline variations in skin conductivity. This component is individual to the person and depends on the person's autonomic response to environmental factors. Another is the skin conductance response (SCR) referring to the short term peaks. These are usually the responses of the sympathetic nervous system to emotionally arousing events. Like the PPG/BVP signal, the EDA signal has several kinds of features [90] [2] [27]. Table 3.2 summarizes the popular EDA based features from literature.

Deep Learning Features

An important factor in the derivation of features from physiological signals is the sampling rate of the acquired signal. Low sampling rates may result in insufficient statistics and would rather be uninformative in the emotion recognition task. Higher sampling rates lead to better frequency domain and non-linear features. In this case, deep learning may help in learning features directly from the processed signal.

Table 3.2: Overview of EDA based features

Feature type	Example
Temporal features	number of SCR events, the sum of SCR startle magnitudes and response duration, rise and recovery times.
Spectral features	spectral power values in the low-frequency bands (0-2.4Hz)
Statistical features	amplitude rise, mean, standard deviation, kurtosis, maxima and minima of the signal frame

Deep Learning-based features are derived directly from raw PPG/BVP and EDA signals quantized by the length of the frame of annotation label. Several architectures such as Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), Long-short Term Memory units (LSTMs) and 1-Dimensional Convolutions can be used to generate embedded features [90]. These are of great interest owing to their ability to capture and preserve temporal stationarity while learning discriminative characteristics from the signal itself. In addition to that, these can be readily adapted to many common deep learning classification schemes, thus simplifying the implementation of classifiers. Such embeddings map raw signals to hypothetical vector feature space. These vectors can readily act as classifying features.

3.2.3. Classification Methods

Emotion Recognition and classification algorithms using physiological signals are extensively found in the literature. Traditionally popular classifiers include Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), k-Nearest Neighbour (kNN), random forests (RF), Support Vector Machines (SVMs) amongst others [28] [86] [90]. These classifiers require specific pre-processing and feature extraction procedures. Owing to a large number of features available from various physiological signals, the selection is performed before classification. Usually, they are accompanied by hand-tailored features constructed with feature optimization and reduction techniques like Sequential Forward Selection(SFS), Independent Component Analysis(ICA) and Principal Component Analysis (PCA). Several of these methods may be applied together to obtain a reduced set of features that have higher discrimination power than the original set of features.

Deep learning algorithms have become powerful classification algorithms with neural networks. Popular methods include Multilayer Perceptrons (MLPs), Long-short Term Memory (LSTMs), Convolutional Neural Networks (CNNs) amongst many [89]. Many studies use a combination of these techniques to create neural network architectures that can perform both feature engineering and classification in an end-to-end manner [27]. Some of these have also made the need for feature engineering obsolete and redundant. These methods can learn correlations between multiple channels and modalities to construct discriminative abstractions of features. Such methods as described here have been shown to perform with higher accuracies than traditional machine learning algorithms.

This and the previous section has analyzed the state-of-the-art features and methods for emotion recognition using speech and physiological signals. These provide much of the motivations for the architectures proposed in this thesis later. These methods, however, are implicitly dependent on large amounts of data for classification and do not fulfil the constraints of the research questions. In lieu of this, we look at few-shot learning inspired methods. The next section provides a brief literature review of these methods.

3.3. Emotion Recognition using Few Shot Learning

As described in the previous chapter in section 2.3, few-shot learning is a class of methods that use very few samples of data for a learning task. Similar to the trends in classic emotion recognition, the focus has been on visual medium – facial expressions, images and videos. Methods involving audio and physiological signals have not been examined in depth to the tune of classical emotion recognition literature.

Facial emotion recognition with FSL models is widely studied. In Ciubotaru et al.[21], the authors present one of the earliest approaches of facial emotion recognition with low shot learning models. The authors use four distance metric learning approaches to perform few-shot learning using several different embeddings architectures. This motivates one of the key aspects of our method – optimized embeddings. Multidomain facial expressions pooling is proposed in D. Kollias et al.. [46], with a novel CNN-based architecture FaceBehaviour-Net. A few models proposed in literature also explore still images by exploiting information of occlusion, pose and illumination [52] [50] [104]. The architectures used are Model Agnostic Meta-Learning (MAML), Generative Adversarial Networks (GANs), respectively.

Table 3.3: State-of-the-art Methods for Few Shot Speech Emotion Recognition

Reference	Architecture	Dataset	Performance (Accuracy)
P. Arora et al..[6]	Siamese Networks	IEMOCAP	63.84%
Boigne et al..[12]	Pre-trained embeddings with RNN	IEMOCAP	64.3 %
A. Naman et al..[64]	MAML	EmoFilm	69.71%
Feng et al.. [34]	Siamese Networks	eNTERFACE, CREMA-D, RAVDESS, IEMOCAP	-

In contrast, few shot speech emotion recognition first appears in literature in P. Arora et al.[6], where the authors propose the use of Siamese networks for the preservation of speaker identity in speech emotion classification. Here, a series of modifications to the CNN embedding functions are made to remove identifiable information using several transformations such as perturbation and dimensionality reduction. In Boigne et al.[12], the authors use standard pre-trained models useful for automatic speech recognition as their basis for identifying emotions from audio-visual data. Here, transfer learning is used to adapt speech recognition problems to emotion recognition. Model Agnostic Meta-learning is proposed to perform speech emotion recognition in A. Naman et al.[64]. Feng et al.[34], propose adaptive pair selection for Siamese Network based speech emotion recognition. Here, the embeddings are generated using time-frequency log spectrograms of speech frames. Table 3.3 summarizes the state-of-the-art performances of Few shot speech emotion recognition methods in the literature.

In the works summarized above, the focus of the literature is on the classification of discrete emotion spaces. Moreover, information from modalities other than speech, video and textual information has very rarely been explored with few-shot learning methods. Our research question tries to fill in this gap. As described in the literature review on emotion recognition in Sections 3.1 and 3.2, there are great benefits to the fusion of information from physiology and speech. Moreover to our knowledge, to date, only one article [6] discusses FSL speech emotion recognition approaches for voice-enabled communication systems. Thus, we aim to investigate multimodal emotion recognition with few-shot learning techniques to obtain evidence of the effect of additional modalities given limited samples of data.

3.4. Conclusion

In this chapter, a brief overview of the emotion recognition literature landscape is provided with a specific focus on the constraints of this thesis. While discussing speech emotion recognition, we identify the important prosody, spectral and cepstral features useful in the discrimination of emotions. The combination of these features effectively captures both time-domain and frequency-domain features while emphasizing the functionalities of the human auditory system. These features are later used in the thesis for the classification task. In addition, deep learning-based classification methods are found to be extremely powerful in this task. We also find that the use of elicited speech is most apt for the application considered in this thesis.

Next, physiological signals are examined for the emotion recognition task. With PPG and EDA signals, we identify deep learning-based feature extractors to be power embedding functions given the low sampling rate of the signals in question. In addition, the constraints of wearable sensors highlight the need for better pre-processing techniques.

Finally, a few shot emotion recognition literature is discussed and presented. While much of the work is unimodal and focuses on visual features (videos and images), we identify motivations for multimodal architectures (speech and physiological). In conclusion, this chapter gives an overall summary of necessary emotion recognition literature for this thesis.

Part III - Datasets and Proposed Methodology

4 Datasets

In this chapter, the pertinent datasets used for the analysis of the research questions of this thesis, are presented. The chapter is divided into three sections. In section 4.1, a brief overview of the available datasets for emotion recognition is summarized. Following, a selection criterion is discussed based on the constraints and research questions presented in Sections 1.2 and 1.3. Subsequently, we present two selected datasets briefly discussing their salient features. The datasets are distinct and present different settings and signal qualities and thus, provide unique setups for analyzing the proposed architecture. In section 4.2, we discuss the K-EmoCon dataset. This is followed by the RECOLA dataset in section 4.3. For each of the datasets, we present the data pre-processing steps, feature extraction and embedding generation together with performance metrics.

4.1. Datasets

A preliminary survey of possible emotion recognition databases is carried out based on numerous surveys available in the literature. Several available datasets are analyzed for suitability. First, we look at available modalities. Table 4.1 shows the emotion recognition datasets classified by available modalities. It may be noted that, for generating this list, existing surveys for speech [1, 94, 83, 63, 29], physiological [86, 13, 28]and multimodal [89, 75, 69, 25] emotion recognition are utilized. This is followed by cross-verification on the availability of the databases online. Some of the existing databases, while found in the literature are now unavailable and hence they are eliminated from this list.

4.1.1. Selected Datasets

Off these datasets, we first shortlist the datasets by our selected modalities - namely audio and physiological signals. These are highlighted as shown in Table 4.2. Further, the available datasets of the selected modality are analyzed and compared for suitability to the research problem. Of these datasets, we select datasets that satisfy the constraints in Section 1.2. This means the emotion elicitation requirements are not induced but rather natural and spontaneous. In addition, the nearest simulation of interaction with IVAs is achieved in dyadic (2-party) conversational datasets. This gives another criterion for selecting the datasets. Finally, the availability of both self-reported and external annotations of emotion is the third criterion for the selection of datasets. This is crucial to determine the gold standard and helps in the comparison of the performance of models with a similar setting.

Table 4.1: Overview of Emotion Recognition Databases found in Literature

Modality	**Datasets**
Audio, Video, Physio	RECOLA, K-EmoCon, MAHNOB-HCI, AMIGOS, VerbIO
Audio, Video, Text	SEND, ICT-MMMO, MOUD, Youtube Database
Audio, Video	CREMA-D, EmoReact, IEMOCAP, SEWA, SEMAINE, eNTERFACE, EMDB, RAVDESS, Belfast Natural Database, Chen-Huang database, SAL Database,
Audio, Text	EmoBank, Facebook posts, EmoLex, Affective Text, ISEAR, Wall Street Journal 100
Video, Physio	ASCERTAIN, DEAP, DECAF
Video	RU-FACS-1, MMI Facial Expression, BU-3DFE, FABO, Cohn-Kanade AU-Coded Facial Expression database
Audio	Banse-Scherer database, Danish Emotional Speech, Berlin Corpus, EU-Emotion Voice Database, Emotional Voice Database, TESS, TURES, ISL Meeting corpus, CSC Corpus, AIBO DATABASE, ESMBS, KISMET, BabyEasr, MPEG-4, SJTU Chinese Database, FERMAUS III, BDFALA, ORESTEIA
Physio	CLAS, DECAF, DREAMER, MPED, GAMEENO, PAFEW
Text	IMDB Reviews, Amazon Database, Blogs database, STS-Test, ANET
Images	JACFEE, Affectuve Image Database, IAPS, GAPED,

Therefore, selected datasets must fulfil the following conditions - involve interaction between individuals, and where the emotions are naturalistic and obtained by that interaction. The result of this selection criteria is the two datasets - K-EmoCon [70] and RECOLA [81] owing to their similarity in experimental set-up and the activities in question, namely two-party conversation.

Table 4.2: Emotion Recognition Databases with Audio, Video and Physiological Modalities found in Literature [1]

Dataset	**Elicitation Method**	**Participants**	**Annotation Approach**
RECOLA[81]	Dyadic Interactions	46	S, E
K-EmoCon [70]	Dyadic Interactions	32	S, P, E
MAHNOB-HCI [92]	Individual video viewing	27	S
AMIGOS [62]	Individual/Group Video viewing	40	S, E
VerbIO [107]	Individual public speaking in VR	55	S

4.1.2. Noise Dataset

Analysis of few-shot emotion recognition tasks under real-world noise is a key research question to be analyzed in this thesis. Therefore, in addition to the two datasets selected for the emotion recognition task, a noise dataset is selected to impute the speech recordings with real-world background noise. From a selection of papers on noisy speech emotion recognition [42][100][95], the DEMAND dataset [97] is selected for imputation of noises. The DEMAND

[1] Explanation : Elicitation Method - Stimuli for emotions ; Annotation - S : Self, P : Partner, E : External annotator

(Diverse Environments Multichannel Acoustic Noise Database) is a comprehensive database with acoustic recordings from numerous settings. The recordings in the DEMAND dataset are 16-channel signals at 48kHz and resampled and processed to provide with 16kHz signals. The dataset consists of 6 categories of noises with 4 categories in indoor environments and 2 categories in outdoor environments. The categories of indoor environments are Domestic, Office, Public and Transportation, while the categories for the outdoor environment are Street and Nature.

The environment settings considered for the IVA in this thesis are indoors, particularly in a household. Therefore, the Domestic and Office categories are selected for noise imputation. Within the Domestic category, 2 specific sub-types of recordings are chosen namely – Kitchen and Living room. In addition, from the office category, 2 other types of sub-types of recordings are chosen namely Office and Hallway. These four categories are deemed to simulate the background noises of a household where IVA interactions occur. Table. 4.3 summarizes the properties of the selected noise signals.

Table 4.3: Noise Characteristics of selected samples from DEMAND Dataset

Noise	Loudness(RMS value)	Maximum dBFS	Highest Amplitude
DLIVING	56	-34.995	583
DKITCHEN	266	-10.916	9324
OOFFICE	257	-9.856	10535
OHALLWAY	113	-22.578	2435

The original noise power of the environment is preserved by avoiding post-processing to gain normalization. Out of the separate 15-channel recordings, noise samples are used only from one of the channels to prevent the effect of variations in noise gain across the different microphone channels, since these are independently recorded. For a fair comparison of the effect of noise on performance, all the noise samples from different categories are selected from the same microphone channel.

4.2. K-EmoCon

The K-EmoCon dataset [70] is a multimodal sensor dataset with continuous emotion recognition specifically collected in naturalistic conversational settings. This dataset consists of 32 subjects in 16 pairs each with 2 subjects in a debate setting. The participants of the dataset are students from the Korea Advanced Institute of Science and Technology. Of the 32 participants, audio-visual recordings of 21 participants are available while no video information is available for the rest 11.

The dataset consists of videos, speech audio, accelerometer and physiological data from the subjects. The average duration of the debates is 10 minutes where each participant can speak for two consecutive minutes. The signals in consideration for the thesis are speech, BVP and EDA which have been captured using LookNTell head-mounted camera, Empatica E4 and Polar H7 Bluetooth Heart sensor respectively. The wearable signals captured in the dataset are summarized in Table. 4.4. The dataset records both valence and arousal based on Russell's Circumplex model on a discrete scale of 1-5. In addition, emotion states describing subjective stress is collected on a discrete scale of 1-4. The annotations available from the dataset are summarized in Table. 4.5. A robust 3-tier annotation is generated by self, partner and 5 external observers. The granularity of the dataset annotation shall help in managing segments of multiple lengths if need be.

Table 4.4: Summary of signal modalities in K-EmoCon

Devices	Collected Data	Sampling Rate	Signal range [min, max]
Empatica E4 Wristband	3-axis acceleration	32Hz	[-2g, 2g]
	BVP (PPG)	64Hz	n/a
	EDA	4Hz	[0.01 μ S, 100 μ S]
	Heart Rate	1Hz	n/a
	IBI	n/a	n/a
	Body Temperature	4Hz	[-40°C, 115°C
Neurosky MindWave Headset	Brainwave	125Hz	n/a
	Attention and Meditation	1Hz	[0,100]
Polar H7 Heart Rate Sensor	HR (ECG)	2Hz	n/a
LookNTell Head-Mounted Camera	Audio and Video	n/a	n/a

The debate settings in this dataset can be thought of as simulated interactions that are likely to occur in a household with an IVA. The debate interaction that occurs in this setting is goal-oriented (to debate the topic) and involves a singular chain of thought. Further, emotions are evoked by a spontaneous display of opinions of the opposite participant. Short utterances from a diarized segment of a speaker, act as speech signals where emotions have been evoked from interactions with people around the user naturally. The context of data collection in K-EmoCon resembles the kind of interactions that would naturally occur between an individual and his/her social group. This forms the basis for the first set of experiments where we consider an IVA to passively listen to audio from a conversation and classify the emotions of a specifically intended speaker who has the wearables. Further, this setting examines the specific case of emotions that occur with pre-conceived beliefs. In this simulated set-up, we examine the social interactions of the user while an IVA listens to the speaker and performs emotion recognition without much intervention of the user.

For experiments, we use the self-annotated labels from the K-EmoCon dataset as it provides a relatively balanced distribution of labels across arousal and valence categories.

Table 4.5: Emotion Annotations in K-EmoCon

Emotion annotation categories	Description	Measurement Scale
Arousal/ Valence	Two affective dimensions from Russel's Circumplex Model of Affect	1: very low; 2: low; 3: neutral; 4: high; 5: very high
Cheerful/ Happy/ Angry/ Nervous/ Sad	Emotion states describing subjective stress state	1: very low; 2: low; 3: high; 4: very high

4.2.1. Data Pre-Processing

While some pre-processing steps are already performed on the dataset, some additional processing needs to be performed on the dataset to make it suitable for the intended experimental setup.

The signals utilized for the experiments are debate audio recordings and physiological signals – PPG and EDA. The dataset consists of 16-paired debates with 32 participants which total 172.92 minutes of dyadic interactions. Physiological signals are captured for a longer duration – slightly more before and after the debate duration. The debates are provided with standard UTC +0-time stamps of their speech start and speech end. The authors provide a check on data availability across various signals captured to mention any errors or mismatches in the collected data. Physiological data from participants 2, 3, 6, 7, 26 is stated to be erroneous or absent and hence these participants are rejected from the study.

Since the debate recordings are paired speeches, there is a need for diarization of the recordings to individual speakers to perform a participant-dependent study of emotion recognition. Thus, separation of audios from the two speakers, as well as respective segments of physiological data, need to be identified. The basis for this is Speech Diarization. The debate audio is diarized using several off-the-shelf open-source diarization toolkits available in Python. To ensure the correct split of audio, diarization is performed using multiple toolkits and their qualitative performance is analyzed. These methods are described in the next sub-section.

Audio Diarization

Following are the speech diarization toolkits analyzed for selecting the diarization method :

1. `Vox-Sort` [103]: `Vox-Sort` is an off the shelf diarization tool that provides quick and easy diarization from the source audio file. It provides satisfactory separation of speakers and generates a time segment-based list showing the times the specific speaker spoke.

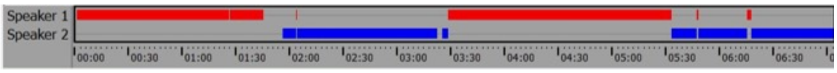


Figure 4.1: Sample Diarization with Vox-Sort

2. `pyannote.audio` [14]: `pyannote.audio` is an open-source speech diarization library in python which uses pre-trained models for performing diarization on a speech sample. It provides great accuracy in determining speaker changes to the accuracy of milliseconds.

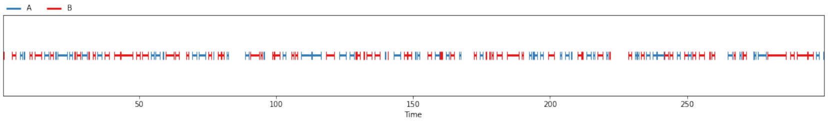
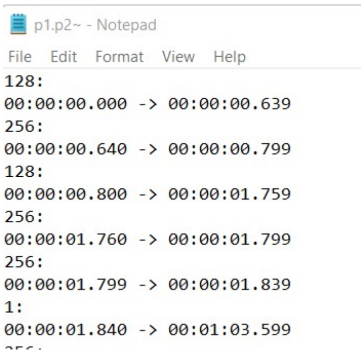


Figure 4.2: Sample Diarization with pyannote.audio

3. `pydiarization` [77]: `pydiarization` is another open-source python library that generates an RTTM file from given audio/video files and diarizes them based on pre-trained models.

To verify the performance of diarization qualitatively following diarizations from the above three methods, manual verification is done with the available videos to check the accuracy of speaker changes and assign the tags generated to participant numbers from the study. We find Vox-Sort a reliable and quick diarization tool that provide a very robust segmentation of speech samples from the debate together with separation of silent segments. Using Vox-Sort, a segmentation file together with speaker code is obtained. A sample of this segmentation is shown in Fig. 4.3. The speaker codes are mentioned as ‘1’ and ‘2’ while silence segments, as well as non-speech sounds, are coded as ’128’ and ‘256’ respectively.



```
128:
00:00:00.000 -> 00:00:00.639
256:
00:00:00.640 -> 00:00:00.799
128:
00:00:00.800 -> 00:00:01.759
256:
00:00:01.760 -> 00:00:01.799
256:
00:00:01.799 -> 00:00:01.839
1:
00:00:01.840 -> 00:01:03.599
```

Figure 4.3: Segmentation Results from Vox-Sort

Separation of specific participant audio and physiological data is performed using these diarization results. Off-the-shelf audio-processing tool `librosa` is used for the selection of audio segments according to the frame times from the diarization result after synchronizing them with the debate time-stamps. Similarly, physiological data and the annotation labels are separated per participant using the diarization results after synchronizing with time-stamps from the modalities with debate time-stamps. This way, out of 16 debate samples, 32 audio, physiological signal and annotation label files are obtained from the corresponding 32 participants. With the 32 participant data available, a cross-verification of audio segments with the debate video recordings is made to verify the correct duration of the speech. Owing to the absence of debate video recordings from participants 11, 12, 17, 18, 27, 28; it is not possible to assign the correct diarization labels to debates involving these participants. These participants are discarded from the experiments. Thus, processed data from 21 participants are available for experiments. The resulting data are summarized in Table. A.1. Post the diarization and time-synchronization of audio signals together with physiological signals, an additional set of audio signals are generated with the imputation of noise, in line with the research questions to test few-shot models in a noisy setting. This is discussed in the next sub-section.

Data Segmentation

To answer the research question on the length of audio segments suitable for the state of the art performance of the proposed model, a suitable segment length must be selected. This is based on the annotation length and the distribution of labels available. Since the unit annotation label is available for a 5-second duration, a sub-split of 5 seconds is possible to lesser duration while preserving the same annotation label. After checking various segment lengths, we select 1s as the duration of the unit sample. This sample length generates enough samples to conduct a few-shot recognition set-up where only a fraction of the original data can be used. Typical fractions of this data are usually 5-10% of the dataset size.

The available processed data from 21 participants, is segmented to 1s. The segmentation involves a split of the audio signal, physiological signals and annotations. The annotation split preserves the annotation from original labels – that is five 1s segments are created by splitting the 5s segment with a specific annotation label. Post segmentation of data, to implement participant-dependent models, a selection of participants is made based on the availability of balanced samples across the two labels for the few-shot models. This is described in the next sub-section.

4.2.2. Selection of Participants

Following the segmentation of data, we perform a participant selection based on the availability of sufficient annotation labels per class. As discussed in 1.4, this thesis aims to perform binary emotion classification on arousal and valence labels. Therefore, the ground-truth self-annotation labels are first converted from 5-level discrete class labels to binary labels with 0 depicting low arousal and valence and 1 depicting high values correspondingly. Here, the neutral label of 3 is removed from the classification as its addition to either of the classes results in a huge imbalance between the two classes. Post the processing of labels across three classes, four participants – 15, 23, 30, 31 – are identified with sufficient samples across the arousal and valence labels to perform experiments on the generated segmentation.

A summary of the resulting selected participant data samples and their class annotations are shown in Table. 4.6.

Table 4.6: Annotation statistics of selected participants for K-EmoCon

Participant	Arousal		Valence	
	0	1	0	1
15	105	111	131	69
23	42	181	109	62
30	83	67	101	67
31	205	110	50	242

4.3. RECOLA

RECOLA [81] stands for the REmote COLlaborative and Affective Interactions. It is a French multi-modal database with 46 speakers and annotations for 5 social behaviours in addition to valence and arousal. The participants are recruited from Freiburg University, Switzerland. The data is collected in a setting where participant pairs are performing collaborative tasks in a virtual environment and communicating with each other. A remote discussion takes place between participants through a task that acts as the source of emotional manipulation. During the tasks, the participants are asked to perform decision making and strategizing on a hypothetical task.

The dataset contains audio-visual information in addition to physiological data collected with spontaneous interactions of subjects with the surroundings. A summary of available signals from the dataset is provided in Table. 4.7. Data records of physiological signals are filtered at 250 Hz. From the collected ECG and EDA signals, HR, HRV and SCL, SCR signals are derived respectively. Signals of interest are audio, ECG and EDA. The audio signal is collected using a unidirectional headset while the ECG and EDA signals are collected using Biopac MP36 unit. The annotation is performed by the participants themselves as well as 6 other independent annotators using standard questionnaires such as SAM and PANAS. Table. 4.8 shows annotation details of RECOLA. The domestic-like setting of the experiments together with an interactive user interface for collaborative tasks is similar to the setting proposed for the thesis work and therefore motivates the selection of this dataset.

Like K-EmoCon, the interactions in RECOLA are spontaneous and dyadic. However, the context of data collection here is that of a joint remote collaborative task. Therefore, the interaction amongst participant pairs in short spontaneous and occurs in bursts. The conversation involves a back and forth of question-answer to arrive at a decision. Emotions are evoked

Table 4.7: Summary of signal modalities in RECOLA

Devices	**Collected Data**	**Sampling Rate**	Resampled Available data
AKG C520L + Audacity	Audio	44.1Hz	22.05kHz
Logitech C270 HD Webcam	Video	n/a	n/a
Biopac MP36 unit	ECG, EDA	1kHz	250 Hz

due to the responses of the opposite participants. This setting is similar to an individual interacting with an IVA with short commands for performing an activity. This forms the second set of experiments for our study where the user actively engages with the IVA for performing an activity. This also provides a test for examining fleeting and short-term emotions arising from interactions as opposed to the stimulation of emotions based on long-occurring beliefs.

In the experiments with RECOLA, the annotation labels from the Gold Standard annotation of external annotators provided with the dataset is used.

Table 4.8: Emotion Annotations in RECOLA

Emotion annotation categories	**Description**	**Measurement Scale**
Arousal/ Valence	Continuous annotations on two affective dimensions from Russel's Circumplex Model of Affect	-1 to +1 with a step-size of 0.01 per 40ms

4.3.1. Data Pre-processing

The RECOLA dataset is constructed such that audio signals from each of the participants are separately available together with their physiological signals and emotion labels. Separate audio channels are provided for each participant during the interaction, thereby eliminating the need for audio diarization. Further, the audio signals are time synchronized with the physiological signals. This reduces the burden of pre-processing and outlier removal. Additionally, the available dataset is without any erroneous values. Therefore it is readily available for use.

Data Segmentation

From the measured data shown in Table. 4.7, the sampling rate of physiological signals is 250 Hz, while the sampling rate of annotations as derived from Table. 4.8 is 25 Hz (with hop-size of 40ms). To obtain a favourable segment size to perform classification, the signals and the annotation labels are segmented at 400 ms. To do this, the gold standard annotation labels provided in the dataset, are first converted to binary labels by a simple mapping function and then averaged over 400 ms. It may be noted that 400ms is chosen since there are no changes in the obtained binary annotation labels which have a change under this segment length. This way the labels are preserved. The original quantization is that of 25 Hz, which is now converted to 2.5 Hz. Thus, one sample consists of 400ms of data. Consequently, Physiological data is also binned into samples of 400ms of data based on the provided timestamps. A cross-verification of annotation labels is performed to ensure the correct assignment of labels. In this way, adequate sample length is obtained without compromising on annotation accuracy.

Since the data from RECOLA is originally meant for a regression task, no further segmentation is made since this would yield complexities in the assignment of annotation labels over different segment lengths.

4.3.2. Selection of Participants

The RECOLA dataset is provided with accurate data across modalities. Here for brevity, out of the 27 participants across the dev, $train$ and $test$ sets, 18 are selected from dev and $train$ sets. Out of the 18 participants, the conversion to binary labels together with segmentation as discussed in the previous section results in a data split between the valence and arousal labels. this is summarized in Table. 4.9. Out of the available dataset, participants $train_3$ and $train_8$ are rejected from the experiments owing to very few samples in one of the labels.

Table 4.9: Annotation statistics of selected participants for RECOLA

Participant	**400ms Segmentation**			
Label Type	Arousal		Valence	
Label	0	1	0	1
dev_1	431	318	136	613
dev_2	285	464	47	702
dev_3	301	448	71	678
dev_4	639	110	495	254
dev_5	487	262	132	617
dev_6	373	376	45	704
dev_7	527	222	349	400
dev_8	219	530	91	658
dev_9	515	234	156	593
$train_1$	524	225	304	445
$train_2$	233	516	168	581
$train_4$	338	411	214	535
$train_5$	542	207	188	561
$train_6$	324	425	56	693
$train_7$	216	533	139	610
$train_9$	202	547	47	706

4.4. Conclusion

This chapter introduced and presented the datasets used to explore the research questions posited in this thesis. We first summarized the large proportion of emotion recognition corpora present in literature across modalities. Thereafter, a subset of the datasets is selected and analyzed for suitability to the tasks of this thesis. K-EmoCon and RECOLA are found to be the most suited datasets which are useful in fulfilling the experimental requirements for answering the research questions. Next, each of the datasets is described in detail. The necessary preprocessing steps for the analysis are presented, together with filtering of participants with erroneous data. This provides the two prepared datasets with two unique settings common to IVAs in households. These datasets are used to train the proposed architecture discussed in the next chapter.

5 Proposed Architecture

We finally discuss the architecture of the proposed model. The motivation for multimodal embeddings stems from the research questions and motivations developed in Section 1.3 and Section 1.2. The architecture must operate in conjunction with an IVA online. The nature of interactions is described in Section 1.1. Given the limited amount and atomicity of interactions with the IVA, the proposed model utilizes few-shot learning techniques.

This chapter is organized as follows. First, in section 5.1 we describe an overview of the proposed architecture of Multimodal Siamese Networks for Emotion Recognition. Then, we discuss the few-shot learning backbone. Then, we present the individual feature embeddings which form the core of the overarching embedding of Siamese Networks. Here, four embeddings from speech, EDA and PPG signals are presented. Then, we discuss the contrastive loss used for the optimization of the model.

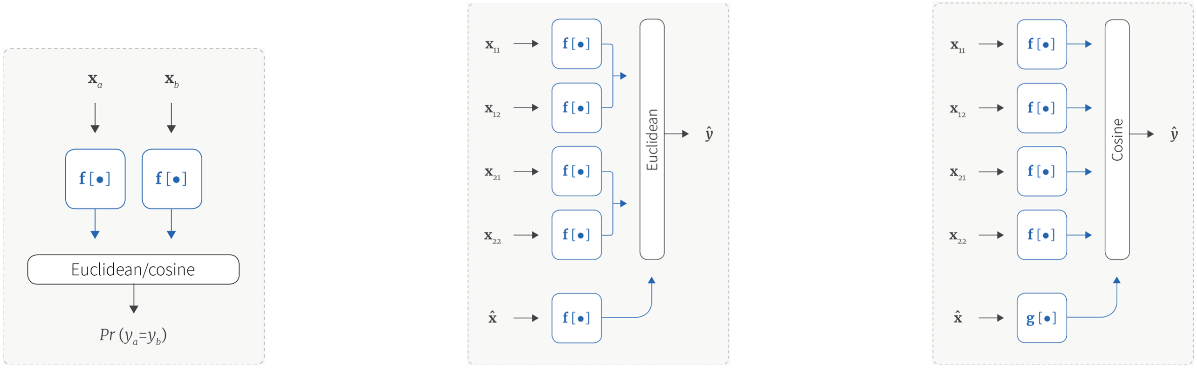
In the next section 5.2, the evaluation metrics used for the analysis of the results are briefly presented together with the motivation of their use specifically for this thesis.

In the final section 6.1, the experimental setup for testing the proposed architecture on the two datasets is presented in detail. This also involves the pre-processing steps for noise imputation, feature generation and dataset splitting. A pipeline setup for the complete process is described concluding with the implementation environment.

5.1. Proposed Architecture

The architecture proposed to solve the research questions posed in section 1.3, is devised considering several problem characteristics and constraints. The original problem of emotion recognition by IVAs using speech and physiological signals can be decomposed by input sources. The audio is input either with IVAs or smartwatches, while the physiological signal is input solely with smartwatches. Therefore, the application in question requires adaptation to speech and physiological signals of any arbitrary sampling rate. The sampling rates of these signals present the first design bottleneck since they are often low and non-uniform across devices. On one hand, the audio recorded by the IVAs is of very high sampling rates of the range of several kiloHertz (for example Amazon Alexa has a sampling rate of 44.1 kHz). While, on the other hand, the physiological signals have quite low sampling rates of under 100 Hz (Samsung Gear smartwatch sampling rates for PPG signal is 100 Hz). This differential sampling rate requires the processing of samples of these two signals independently. Further, feature extraction from both these signals must cater to the common embedding space desirable for emotion classification. In addition to the sampling rates, another important aspect of the input signals is that the audio signals are available in short segments typically interactions with IVAs and smartwatches last for 5-7 seconds or typically 4 commands per instruction as summarized from the literature review presented in section 1.1. This constrains training size and the unit data sample size. Therefore, an architecture must accommodate a sample length appropriate to the sampling rates and duration described so far.

For a training set $\{D_s\}$ for each of the N classes, K samples from each of the N classes are used for training. The embedding model is optimized by maximizing the performance of these classes. The rest of the samples are used to test the model. A good generalization across all classes is achieved this way. Popular metric learning methods are shown in Fig. 5.1. These are Siamese networks, prototypical networks, and matching networks. The metric of each of these methods is the distance between input samples, however, the difference lies in the optimization objective. For Siamese networks, contrastive loss minimization is the objective, while for Prototypical networks, the update and classification lie with the decision of class prototypes generated based on the nearest neighbour mechanism. Matching networks compare similar data samples individually from the support set to perform classification with a sample on the query set. Prior work from the literature review of a few shot emotion recognition algorithms in section 3.3 is the basis of using the Siamese network backbone. This is explained briefly in the next sub-section.



(a) Siamese Network

(b) Prototypical Network

(c) Matching Network

Figure 5.1: Metric Learning Method Architectures (adapted from [111])

Siamese Networks

Siamese Networks [45] is one of the most popular Metric Learning Methods. These operate on the principle of similarity between two samples and predict a probability of whether they belong to the same class or not. A typical architecture is shown in Fig. 5.2. For two samples $x_{s,1}, x_{s,2}$ in a support set $D_{support}$, a mapping is generated using the same embedding function f. The distance metric between the two embeddings is calculated using Euclidean distance given by equation 2.5. Thereafter, this distance is used to optimize the network using a contrastive loss function with the labels y given by –

$$L(y, d_{l_2}) = \frac{1}{2}(1–y) \cdot d_{l_2}^2 + y \cdot \text{max}(0, m - d_{l_2})^2 \tag{5.1}$$

Where m is the loss margin. The Contrastive loss is a distance-based loss that forces similar data points (of the same class) to have a low Euclidean distance and dissimilar data points (from different classes) to have a higher Euclidean distance. Contrastive loss is further discussed in the chapter in section 5.1.4. Siamese networks are the backbone of the proposed multimodal Siamese network architecture. The proposed architecture extends the concept from mono-modal (image) embeddings to multi-modal embeddings. In the next sections, we describe the components step by step.

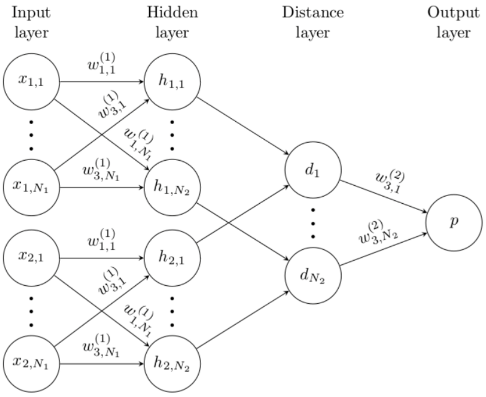


Figure 5.2: Siamese Network Architecture [45]

5.1.1. Multimodal Siamese Network

The proposed Multimodal Siamese Network is an extension of its single modality variant. Specifically, the architecture makes use of the modality-specific embeddings introduced by the authors in [38]. The authors introduce the concept of emotion embeddings. To reinforce the concept of embeddings for few-shot learning, the independent embedding functions are created per modality to create a final joint-multimodal embedding. In this network, 3 modalities contribute to the emotion recognition task. After extracting audio and physiological descriptors from the standardized pre-processing steps, three modality-specific embeddings are used to project unimodal descriptors to a subspace which contributes to the final embedding in question.

As already discussed earlier, Siamese networks are composed of a twin embedding structure with shared weights. The network arms have embedding functions to map the inputs to a common sub-space. This embedding is a high-level feature abstraction of the input. The embedding structure can be chosen to be any arbitrary function that optimizes this loss. This idea is multiplexed across three modalities to create the desired multimodal embeddings in question. This allows signals of any arbitrary sampling rate to be embedded in the desired space as they are segmented to be simultaneously input to the architecture. Further, the independence of embeddings preserves the individual contribution of signal features. The two twin networks have the same parameters and weight updates during training. The sharing of weights between arms ensures that two similar inputs are not mapped to very disparate locations in the sub-space.

Mathematically an embedding function f_m for a modality m maps a sample input from that modality onto a shared coordinate space $\mathbb{R}^E$. If $x_{(\cdot)}$, $e_{(\cdot)}$, and $y_{(\cdot)}$ represent input feature, embedding vector and output prediction of a unimodal input, then three different strategies of feature fusion can be formed. For features from two input sample features x_1 and x_2, for early feature fusion combines features as $[x_1; x_2]$; for decision level fusion a weighted averaging is considered ($ay_1, +by_2$; where a, b are weights over decisions y_1, y_2) and model level fusion suggests a concatenation of embeddings $[e_1; e_2]$. Here we use the model level fusion strategy.

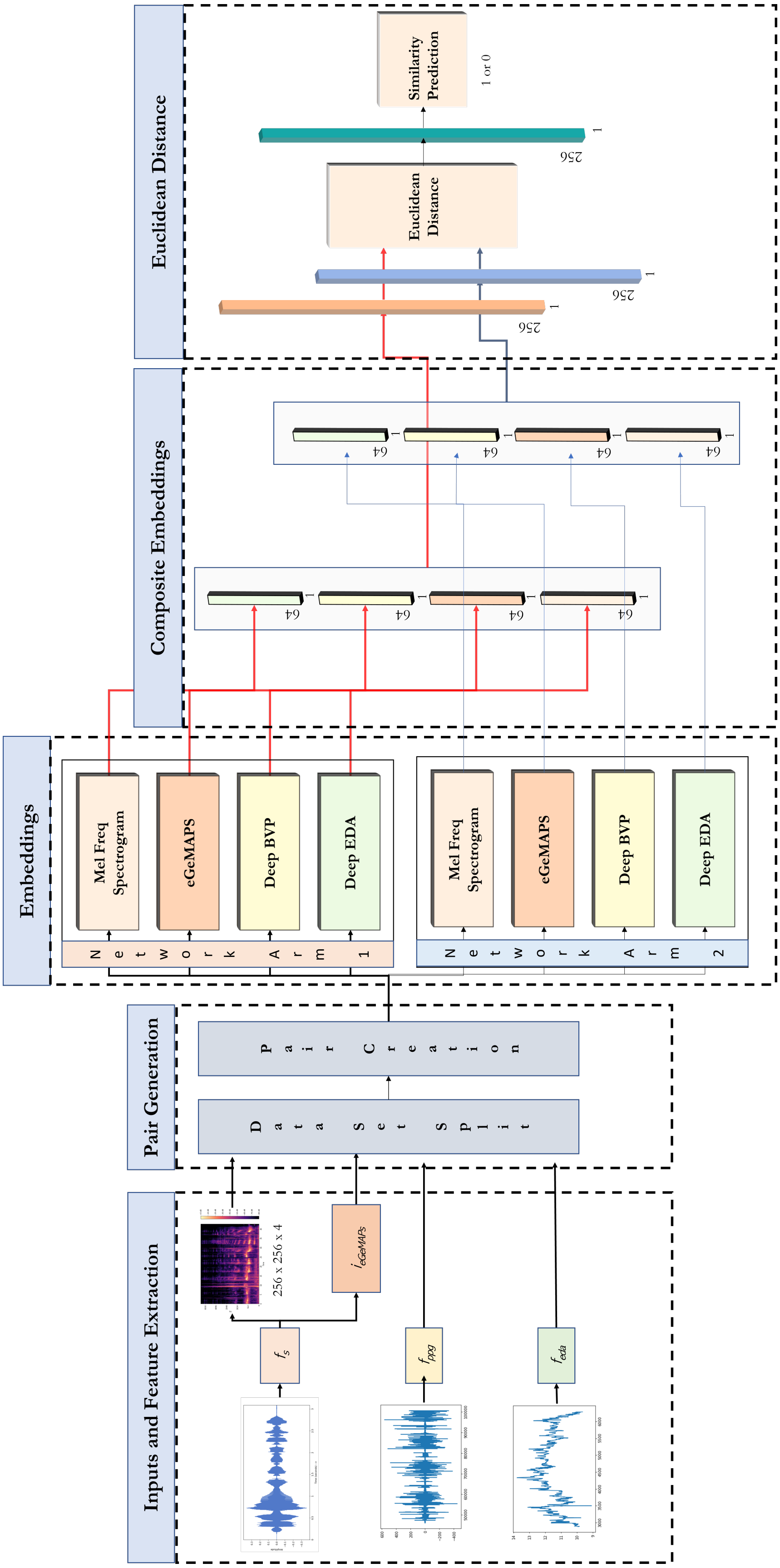


Figure 5.3: Overview of Proposed Architecture including pipeline

As illustrated in Fig. 5.3, the arms comprise four different kinds of embeddings – two embeddings for two different sets of audio features, one embedding for the EDA signal and one embedding for the PPG signal. We use Mel Frequency Spectrograms, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), raw EDA signal and raw PPG signal as inputs to the four feature embeddings. Each of the embeddings can be considered as a black box that maps the input signal from a modality through a specific mapping. All the embeddings are independent of each other. Each of these embeddings generates a feature vector of size 64. In the proposed model, we use model level fusion with the concatenation of embeddings to fuse information of different modalities. Thus, all individual embeddings of size 64 are concatenated to generate a multimodal feature vector of size 256. This concatenated feature vector represents the condensed information from all four modalities for a single sample set.

Following the generation of 2 such embeddings from the two arms of the Siamese Network, the similarity between the two vectors is measured using Euclidean Distance Metric as shown in Eq. 2.5. This distance measure is then converted to a prediction using a fully connected layer with one output and a sigmoid activation. This converts the distance metric to a probability distribution of the input over similarity (1) or dissimilarity(0).

The embedding functions are designed to take signals of any arbitrary length as inputs and map them into feature vectors of size 64. This allows the model to work universally with audio of any length. This model architecture is uniform across experiments with different audio lengths, therefore allowing a comparison of performance based on audio length. In the next sections, detailed architectures of each of the embeddings are discussed. The optimization of the model is performed using the Contrastive Loss which is discussed in the later section.

5.1.2. Audio Embeddings

Audio embeddings for the proposed architecture consist of two types. These two embeddings represent the features discussed in chapter 3 earlier - namely the prosody, spectral and cepstral features. The structure of the two audio embeddings is derived from prior work found in the literature. The following sections discuss the two embeddings in detail.

eGeMAPS embedding

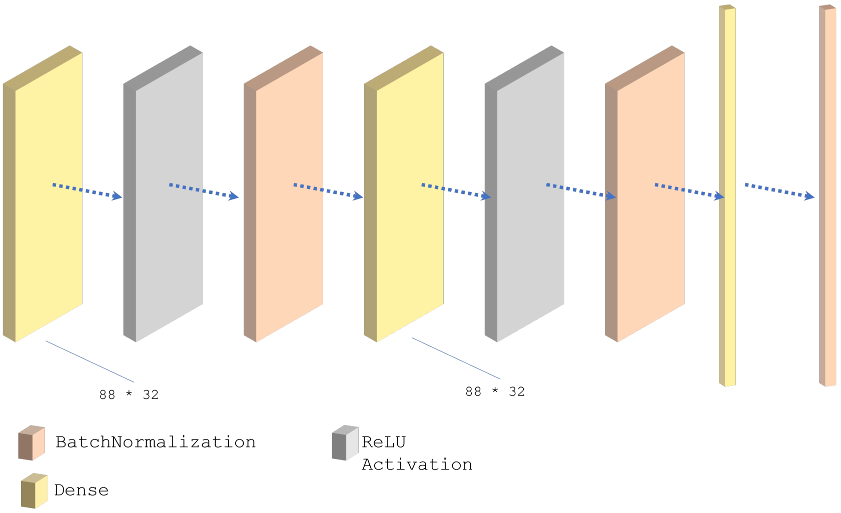


Figure 5.4: Embedding Architecture for eGeMAPS

For the spectro-temporal features, the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS [33]) is extracted using the openSMILE toolkit [32]. This results in a set of 88 acoustic features per sample segment. The features for sample audio, generate a vector of size 88 which acts as input to a fully connected neural network. This embedding consists of two fully connected layers with 32 units each with a ReLU activation function followed by a flattening layer. Another fully connected layer of size 64 maps the flattened input to a vector of size 64. The architecture is shown in Fig. 5.4. This embedding summarizes the prosody, spectral and amplitude features of the audio sample. Therefore, this embedding is complementary to the Mel-frequency spectrogram embeddings presented next. Since these features are speech specific, they make the composite audio embeddings robust against background noise.

Mel-frequency Spectrogram embedding

To capture the cepstral features of the audio sample, Mel-frequency spectrograms are used. As shown in section 3.1.2, Mel-frequency spectrograms are superior to power spectrograms in describing the frequency distribution of audio signals due to the Mel-band scale. Therefore, the spectrograms thus generated are descriptors specific to the human auditory system. It may be noted that this also preserves the audio features against disruption from noise as the noisy features are dispersed across higher frequency bands which are omitted in the Mel-bands. This makes the Mel-frequency spectrograms very powerful features for emotion recognition from audio samples which have background noise.

The embedding function for Mel-frequency spectrograms is a Convolutional Neural Network. Several popular embeddings are found in the literature for speech applications. These include architectures adapted from AlexNet, VGG Net, Inception V3 and ResNet-50 [39] [87]. Each of these architectures has millions of weight parameters and require 10s of GPUs and parameters servers to compute and optimize embeddings. However, we consider embeddings with a much smaller number of parameters and hence the proposed architecture of spectrogram embeddings is a smaller and less dense network. Owing to the limited amount of data, it is also ideal to scale-down models to prevent over-fitting.

Mel spectrograms are used for speech emotion recognition by the authors in [66]. This architecture is used as a base architecture for embedding with several modifications. In the current architecture, three composite convolutional blocks are used followed by fully connected blocks. Each of the convolutional blocks consists of a two-dimensional convolution operation followed by batch normalization, activation and max-pooling layers. The activation function used is ReLU. The first convolutional block contains 96 kernel units and a kernel size of 11 x 11 and a stride of 4 x 4 with a ReLU activation. This layer downsizes the input spectrogram to its feature maps to learn details of the complete spectrogram. To average out the feature generations, a max pool layer of pool size 2x2 and stride 2x2 follow is used in this block. The second convolutional block consists of 256 filters of kernel size 5x5 and stride 1x1 with a ReLU activation. This layer learns the finer details from the previously generated structural feature maps. A max pool layer of size 2x2 and stride 2x2 averages the activation to a lower sized map. Finally, the third convolution layer with 64 kernels of size 3x3 again down-samples the finer feature maps from the previous layer to generate abstract feature maps. Next, the feature maps thus generated are flattened to given a long vector of feature representations. Finally, two fully connected blocks with 4096 and 500 units respectively follow the convolutional blocks. These layers have a ReLU activation. These act as reinforces of the learned features to retain leaned features and condense them to a smaller feature vector. Finally, a fully connected layer of size 64 maps the complete vector to size 64 giving the final embedding. The architecture

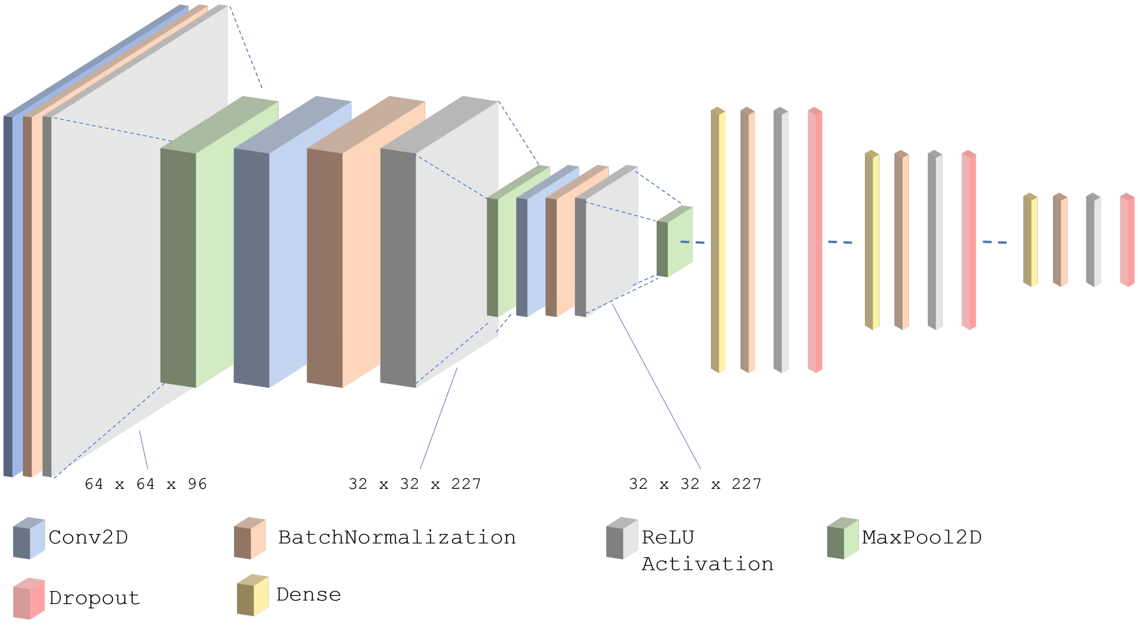


Figure 5.5: Embedding Architecture for Mel Spectrograms [modified from [66]]

discussed here is shown in Fig. 5.5. This simplified architecture generates a good feature representation with far fewer parameter training.

5.1.3. Physiological Embeddings

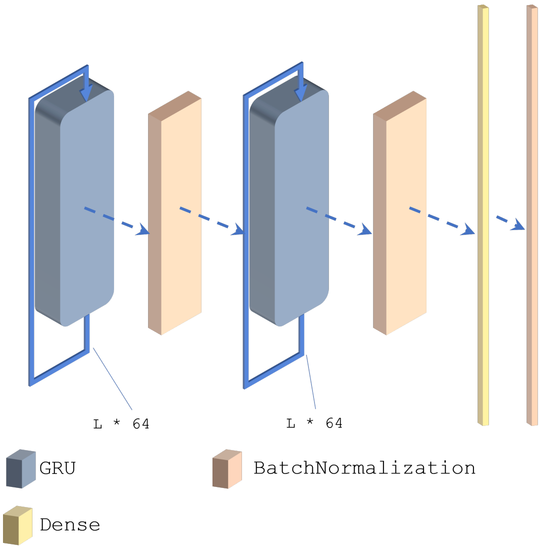


Figure 5.6: Embedding Architecture for EDA and PPG signals

Physiological signals are mapped to embeddings using deep recurrent networks. The deep physiological embedding is designed according to the physiological embeddings proposed by the authors in Han et el. [38]. The authors use GRU networks as the basis of encoding raw segmented physiological signals to embeddings. In the current work, both the embeddings are identical. The embeddings consist of two GRU layers of 64 units each. These layers have a hyperbolic tangent activation function. This activation is used to stabilize the performance of GRUs. Following the GRU encoding, the outputs are flattened to generate a unit dimensional feature followed by a dense layer restricting the output to 64. The embedding directly generated from the physiological signal is beneficial in this case, as the proposed framework is constrained with signals of low sampling frequency. This prevents the use of statistical and frequency-based features for short segments of signals. In Han et al. [38], the authors point to the simplicity of GRUs over LSTMs with fewer parameters as an advantage while generating feature representations. The GRU layers are used to model physiological features from signals that persist over time thus retaining the temporal characteristics of the signal. The embedding structure is illustrated in Fig. 5.6.

5.1.4. Contrastive Loss

Contrastive Loss, introduced in Haskell et al. [37], is a similarity metric that uses the comparison of embeddings instead of actual labels to measure loss. It operates on a pair of data points instead of individual ones. Consider a sample embedding e_a whose label is known. For another embedding e_p, the pair $[e_a, e_p]$ is said to be similar in a metric to be learned if their corresponding class labels are the same. Consequently, for an embedding e_n, the pair $[e_a, e_n]$ is said to be negative if the class labels of the two are different. The objective of the contrastive loss function is to lean representations that yield a small distance d between similar pairs and a large distance between dissimilar pairs. Effectively, this loss function forces the representations to give 0 distance between embedding pairs with the same labels and a distance greater than a margin m for embedding pairs with different labels. The choice of the distance function d can yield different results. Euclidean distance is most commonly found in literature and is used here as the distance metric. For two embeddings e_0 and e_1 with a binary flag label y which is 0 for a negative pair and 1 for a positive pair, this loss is given by –

$$L(e_0, e_1, y) = y\,\|e_0 - e_1\| + (1 - y)max(0, m - \|e_0 - e_1\|) \tag{5.2}$$

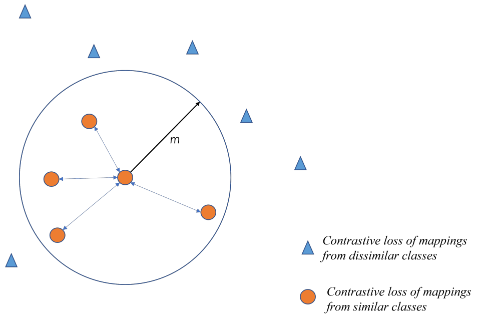

m
Contrastive loss of mappings from dissimilar classes
Contrastive loss of mappings from similar classes

Figure 5.7: Illustration of Contrastive Loss and margin between samples

In the proposed architecture, the embeddings e_0 and e_1 are the composite embeddings generated because of the concatenation of the four embeddings discussed in the previous sections. As shown in Fig. 5.3, the four embeddings of dimensions 64×1 form three modalities are concatenated to generate a composite embedding of dimension 256×1. Euclidean distance is used as the distance metric in the proposed architecture to compute the contrastive loss between the two embeddings. The network outputs the prediction on the comparison of the two embeddings giving a probability similarity flag (denoted by 1) or a dissimilarity. This characteristic similarity is modelled for both arousal and valence identically. The same model can be used for both arousal and valence dimensions.

5.2. Evaluation Metrics

In this section, the evaluation metrics used to analyze the performance of the proposed architecture are introduced. The focus of the discussion is metrics useful for the evaluation of Siamese networks. In the following sections, we list the evaluation metrics used in the experiments in this thesis. These are effective and widely-used criteria for assessing the performance of few-shot learning methods in emotion recognition tasks in literature.

A. Residual Contrastive Loss Residual Contrastive loss is one of the basic measures of performance for Siamese networks. This is because it directly provides a measure of similarity depending on the decided margin set for the training of Siamese networks. Contrastive loss of the query set compared against the support set gives the network's ability to generalize and minimize on new unseen samples. Contrastive loss is given by the equation 5.2.

B. Weighted Binary Accuracy As described in Chapter 4, the annotation labels of the datasets are converted to binary labels to simplify the classification task to a binary one. The small and imbalanced datasets require a weighted metric for the comparison of classification performance. The datasets have a large disparity in distribution for both arousal and valence dimensions along both high and low categories is non-uniform. Therefore we use weighted binary accuracy as one of the metrics of measuring model performance. In the current work, both support set and query set are evaluated for weighted binary accuracy.

C. Weighted Precision, Recall and F1 Score Some of the prior works ([64] [12] [6]) have utilized the binary-weighted Precision, Recall and F1 score metric with their methods. In the current work, these are used to provide a comparison of the performance of the model against various noise compositions and support set sizes.

5.3. Conclusion

In this chapter, we presented the proposed architecture to answer the research questions of this thesis. The Multimodal Siamese Network is presented with details of the embeddings from various modalities and features. The four different embeddings which result in equidimensional feature vectors are explained in detail. Finally, the concept of Contrastive loss is discussed which is used as the loss metric for optimization of the Siamese Network.

Next, the evaluation metrics which will be used to assess the performance of the model are discussed. These include contrastive loss, weighted binary accuracy, AUC score and weighted F1 score. In the next chapter, the experiments involving the two datasets are described in detail together with the results.

Part IV - Experiments & Results

6 Experiments and Results

This chapter presents the implementation details of experiments associated with the two selected datasets on the proposed architecture and discusses the results obtained. Firstly the experimental set-up of the complete pipeline is presented in section 6.1. This section details the various steps associated with noise imputation in sub-section 6.1.1, followed by feature extraction in sub-section 6.1.2. Next, the data set split process is explained in section 6.1.3. Finally, the model implementation details are shown in 6.1.4.

In the subsequent section, we present the complete experiments and results with the K-EmoCon dataset in Section 6.2. In section 6.2.1 the mean baseline results for this dataset are discussed. The results of this dataset with imputed noise is presented in section 6.2.2. Finally, the results of individual participants are presented in section 6.2.3. Next, in Section 6.3, the complete experimental setup and analysis of RECOLA are presented. The organization of sections follows from the results of K-EmoCon. Firstly, in section 6.3.1 the mean baseline results for this dataset are discussed. The results of this dataset with imputed noise is presented in section 6.3.2. Finally, the results of individual participants are presented in section 6.3.3. For consistency in the discussion, all the sub-section discussions are organized by the specific performance metric in question.

Finally, in section 6.4, we compare the proposed model with the state-of-the-art models found in the literature. This concludes the chapter on results.

6.1. Experimental Setup

The main research question posited in Section 1.3 revolves around the proposed Multimodal Siamese Networks. The architecture suited for the problem was discussed in the previous chapter. In this chapter, the related pertinent sub-questions are answered using the selected datasets. This forms the ground hypothesis for the experimental setup. Here, several parameters need to be tested namely – the number of data samples to achieve state-of-the-art performance, the effect of segment length and the impact of real-world noise. These conditions are tested across two settings of the IVA with the two selected datasets.

Following the segmentation of datasets, noise imputation is carried out on the audio signals to simulate the required noise settings of a household. Then, the feature extraction process describes the procedure to create necessary inputs for the embeddings. Then, actual experimental tasks are treated per dataset separately. This pipeline is shown in Fig. 6.1.This setup is described in the following sections.

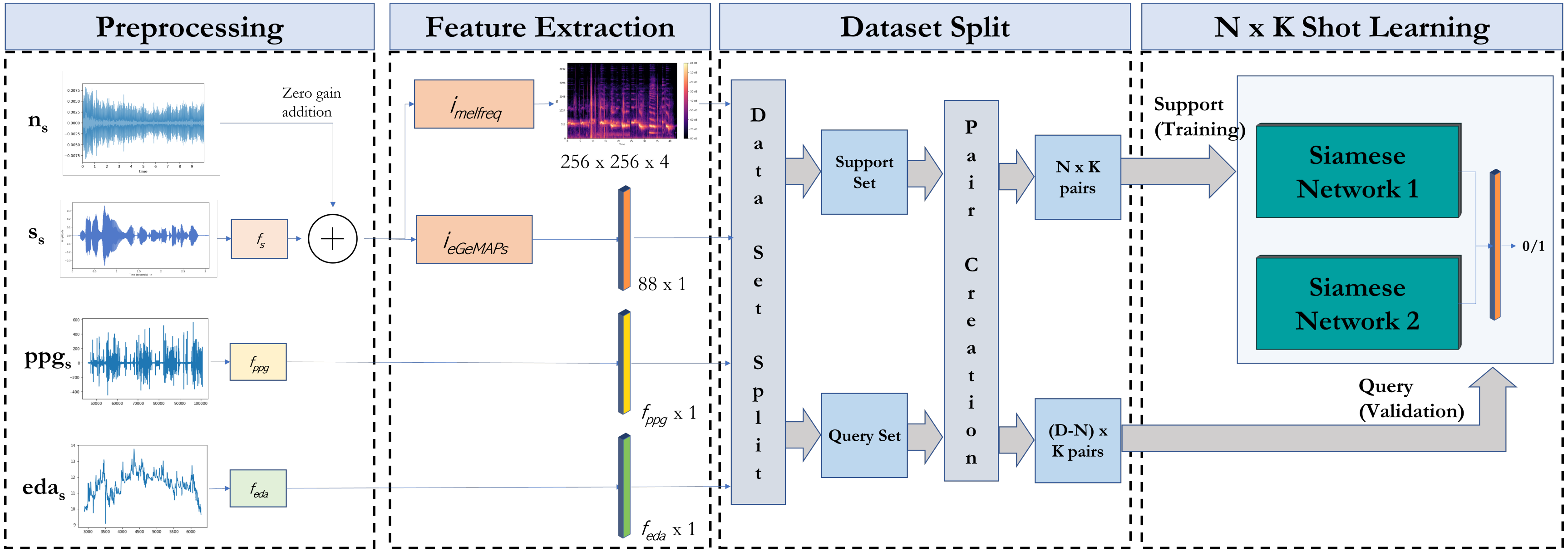


Figure 6.1: Experimental Setup

6.1.1. Noise Imputation

To analyze the performance of few-shot emotion recognition in real-world settings, the signals must be analyzed in conjunction with background noise. To simulate this setting, noise samples from the DEMAND dataset are imputed in audio signals. The length of the audio samples from the DEMAND dataset is multi-channel, however, only a single channel is used to impute the participant audio sample with noise. This setup is based on the conjecture that the noise is a single source and unidirectional, and does not involve any cross-talk amongst multiple noise sources. While this assumption does not capture the actual setting of an IVA, where multiple background noises are present around the device, it simplifies the problem setup. Further, the effect of cross-talk noise on the performance of the proposed model is beyond the scope of the current study. The samples are of the length of 5 minutes. Since most participant audio lengths are around 5 minutes, therefore, the noise sample can be overlayed directly to the audio of the participants. The imputation is performed without applying power gain to the noise signals. This is done using the `pydub` [78] python toolkit. This means the participant audio signal is completely overlaid with the noise signal in an end-to-end manner. Four noise categories are used in this process, namely `DKITCHEN`, `DLIVING`, `OOFFICE` and `OHALLWAY`. This process yields four audio signals per participant for each dataset, in addition to the processed audio signal obtained from the dataset.

6.1.2. Feature Extraction

After the data pre-processing steps described in chapter 4, the corresponding signal features are generated. Feature generation is performed on audio signal alone since the architecture aims to generate GRU embeddings directly from PPG and EDA signals. For this, the physiological signals (PPG and EDA) from the selected participants are padded to the specific segment lengths directly from the raw data to generate the signal inputs for the embedding architectures.

Audio features described in Section 3.1.2 are used in the proposed architecture. The classes presented – prosody, spectral, and cepstral features are generated. Owing to the superiority of mel-frequency spectrograms against power spectrograms to represent human auditory responses, this cepstral spectrogram representation is used as one of the inputs for the embeddings. Mel frequency spectrograms are generated for the sample length used in the respective datasets. Mel-filter banks corresponding to an FFT window length of 2048 samples with a hop-length of 512 samples is generated. The choice of Mel banks is made depending on the segmentation of samples and to maximize the representation of the sample information without leading to empty frames. This is summarized in Table. 6.1.

Table 6.1: Hyper-parameters for Mel Frequency Spectrograms

Dataset	Sample length	Mel-filter banks	Optimal Value
K-EmoCon	1s	$[64, 128, 512, 1024]$	512
RECOLA	0.4s	$[64, 128, 512, 1024]$	128

For K-EmoCon, the number of Mel banks used is 512, while for RECOLA, the number of Mel banks used is 128. The Mel filter banks generate a Mel-frequency spectrogram image. To generate the mel-frequency spectrograms from the audio samples, the off-the-shelf audio-processing library `librosa` [58] is used. The resulting image of the spectrogram is read using `imread` as a 4 channel image to a `numpy` array representation of size $[256 \times 256 \times 4]$. This image is used as feature input for the cepstral embedding of the proposed architecture.

Aside from the cepstral embedding, the temporal and spectral information of the audio signal is encoded through the encore of features discussed in Section 3.1.2. This includes all the frequency-related parameters, energy-related parameters, temporal features, spectral dynamics, and frequency bandwidths. These are complemented with the arithmetic mean, coefficient variations and functional of each of the features. To capture all the mentioned features, the audio feature-set described by the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [33], from the `openSMILE` toolkit [32]. The eGeMAPS parameter set consists of formerly discussed features together with their functional, arithmetic mean and covariance. This results in 88 features per sample. These parameter sets have been evaluated for both the binary valence and arousal labels and have been found to perform substantially well on the existing datasets. These form the spectro-temporal audio feature embedding of the architecture. The complete list of the selected features is provided in the Appendix A.

The above featurization for audio signals is carried out for 5 different settings – one for the participant audio signal as is, and four other signals with four categories of noises mentioned in Section 6.1.1. This process is repeated for both datasets. Thus, for K-EmoCon, this process results in 5 sets of features for analysis with the sample segmentation of 1s. For RECOLA, there are 5 different sets of features with a sample segmentation of 400ms.

6.1.3. Dataset Split

The training and testing setup for both datasets consists of common steps. From the feature generation process discussed in Section 6.1.2, participant dependent features are generated for respective datasets. The four sets of data streams corresponding to the different modalities – PPG, EDA, Mel-Frequency spectrograms and eGeMAPS feature set are constructed. These are inputs for the four different embedding functions described in Chapter 5.

1. **Support and Query Set Creation**: For a given size of K shots for the Support set, K samples of each of the features are selected from each of the label categories $[0, 1]$. This forms the N x K Support Set. The remaining data from the available feature set per participant acts as the Query Set which is unseen in training. The selection of the Support set samples is randomized to avoid bias in the model since the samples are time-sequenced. A good generalization is obtained in this way by random sampling across the whole duration of the audio signal. Following the proposed setup of 2.3.1, for the binary emotion (specifically arousal or valence) classification task T, a feature set $D = \{D_s, D_q\}$, is split into Support set $D_s = \{(x_{support}, y_{support})\}$ of size I and contains $I = N$ x K samples, and a Query set $D_s = \{(x_{query}, y_{query})\}$ of size $D - I$ contains the remaining $D \setminus D_s$ samples. This also ensures, pair creation for the support and query sets is independent.
2. **Creation of Pairs**: As proposed earlier, the Multimodal Siamese Network requires paired data. For this, pairs are created within the support and query sets, with 2 different labels. For samples having the same label i. e., either both 0 or 1, the feature pairs are labelled 1, while for samples with dissimilar labels, the feature sets are labelled 0. These newly assigned labels are not to be confused with the original labels for valence and arousal obtained from the dataset. The new labels of 1 and 0 are defined by convention popular in literature and correspond to the similarity or dissimilarity of pairs, respectively. Multimodal feature pairs are constructed using the index of the generated sample. One input training sample for the Multimodal Siamese network is a paired sample set obtained as a result of the process described above.

6.1.4. Model Implementation

The architecture for the experiments is implemented in Tensorflow v1.0 with the `keras` library. The architectures are trained on a GeForce RTX 2080 Nvidia GPU. For each dataset, training hyper-parameters are identified using a simple test described below. The hyper-parameters include learning rate and batch size. The architectures are trained using the `Adam` optimizer.

Training Hyper-parameters

To find the most optimal learning rate and batch size, a learning rate scheduler is used based on the learning rate range test presented in Smith [91]. Here the learning rate is gradually increased over each batch of training on a logarithmic scale. The implementation is tuned to evaluate contrastive loss values of the query set over the learning rate range to find the most optimal architecture. This process is repeated for several batch sizes resulting in learning rate plots for multiple batch sizes. The learning rate and batch-size corresponding to the minimum query set loss is selected as the optimal set of values. This provides an objective way to reduce the search space and find optimal parameters quickly.

6.2. Experiments with K-EmoCon

For the first task representing the scenario of IVA listening passively to conversations, experiments are performed on K-EmoCon. As described in the previous sections, the pre-processing of the K-EmoCon dataset gives 4 selected participants with 1s sample. The samples are processed with the processes described in the previous sections to obtain the audio features. Thereafter, the segmented dataset is split into support and query sets followed by pair generation within these sets.

Support Set Size

To answer the second research question, the support dataset size must be analyzed. This is determined by the values of K. The rationale for the choice of K is based on the duration of a unit interaction with an IVA/smartwatch obtained from the literature. As found in section 1.1, this is found to be $\mathbf{5}-\mathbf{7s}$ [101] [59]. For our experiments, this is rounded off to $10s$ to cater to the design which is taken as the baseline interaction i. e., *one shot of interaction*. Further, the range of shots is chosen to denote multiples of this interaction in the range $[\mathbf{1},\mathbf{2},\mathbf{3},\mathbf{4},\mathbf{5},\mathbf{6}]$. This duration also helps in designing the sample count of $N \times K$ across the various values of K chosen. Given the data samples shown in Table. 4.6, the value of the support set size i. e., K, with their relation to the interactions and the percentage of participant dataset is shown in Table. 6.2. It is important to note that, for K-EmoCon, the processed dataset samples are of duration $1s$, and the setup described here, is suitable for this sample duration only. This value of K is the basis for the first set of experiments performed on the selected participants. The models are trained separately to predict binary valence and arousal classes.

Table 6.2: Support Set Size for K-EmoCon

Variable	Values					
K	5	10	15	20	25	30
$N \times K$	10	20	30	40	50	60
interactions	1	2	3	4	5	6
% of Data	6.67	13.33	20	26.67	33.33	40

Learning rate Range Test for K-EmoCon

The results of the learning rate range tests for samples from K-EmoCon are summarized in Table. 6.3. for arousal and valence respectively. The optimal batch size for both arousal and valence models is 16.

Table 6.3: Hyper-parameter Tests for K-EmoCon

Model	Hyper-parameter	Range of Values tested	Best value/ Optimal Range
Arousal	learning rate	$[1e^{-6}, 1e^{-2}]$	$2e^{-4}$
	batch size	$[8, 16, 32, 64]$	16
Valence	learning rate	$[1e^{-6}, 1e^{-2}]$	$2e^{-4}$
	batch size	$[8, 16, 32, 64]$	16

- **Arousal Model**: The optimal batch size for arousal models is 16 as shown in Fig. 6.2. The reason for choosing 16 as batch size for the Arousal model is that this batch size minimizes the contrastive loss for not only the query set which is the target of the algorithm but also the support set, which is a visually objective metric for analysis.
- **Valence Model**: Similarly, this applies to Valence as well. The corresponding minima of the trajectory are selected as the working learning rate. This is backed by the fact that the model with 16 as batch size also has the maximum accuracy for both support and query sets. As visible in Fig. 6.3 the graph for batch size 16 gives the minimum contrastive loss amongst all the tried batch sizes.

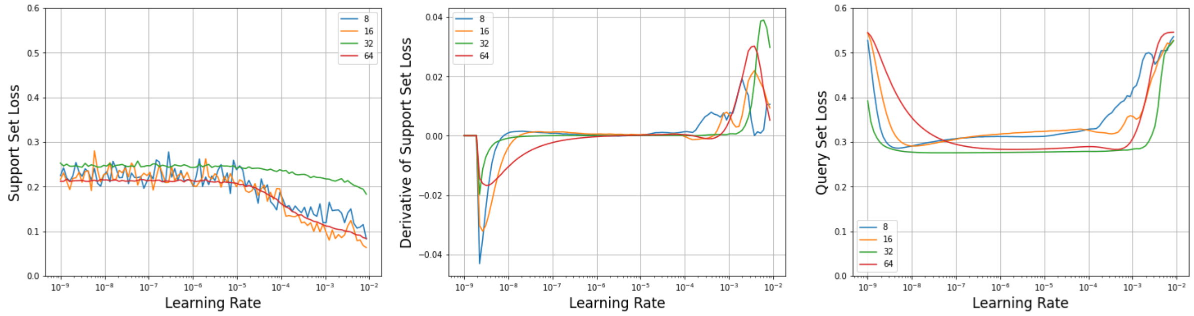


Figure 6.2: K-EmoCon : Learning-rate Range Test for Arousal Model

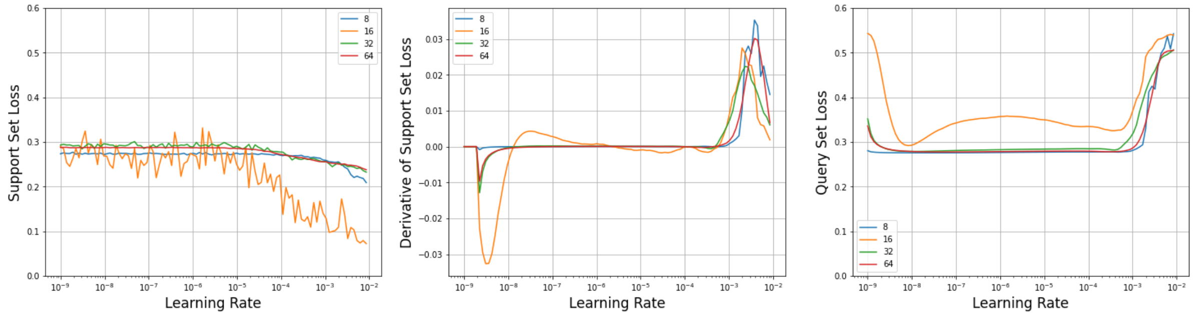


Figure 6.3: K-EmoCon : Learning-rate Range Test for Valence Model

In the following sections, experiments for the four selected participants are discussed.

6.2.1. Results with Baseline Case

In this section, we present the baseline results for the K-EmoCon dataset with our proposed Multimodal Siamese Network. Table. 6.4 shows the results for the arousal and valence models. The values shown here are the mean values obtained across the four participants for each value of K. The table includes the evaluation metrics discussed earlier in section 5.2 - including residual contrastive loss (for support set (L_s) and query set (L_q) *pairs*), binary classification accuracy (for support set (A_s) and query set (A_q) *pairs*). The precision (P), recall (R), f1 score ($F1$) are calculated solely on the query set. Except for the residual contrastive loss, all the other metrics are expressed in percentage. The detailed $Baseline$ results for K-EmoCon are shown in Section B.1.

Table 6.4: K-EmoCon : Performance Metrics for Baseline Models

	Arousal Model							Valence Model						
K	L_s	L_q	A_s	A_q	P	R	$F1$	L_s	L_q	A_s	A_q	P	R	$F1$
5	11.02	9.36	93.75	54.06	55.52	53.42	48.24	6.38	5.47	88.75	51.89	58.65	51.36	42.82
10	7.95	6.72	88.75	56.51	56.66	56.45	55.95	0.11	0.25	98.12	60.75	61.42	60.75	59.61
15	5.40	4.66	94.16	55.02	54.87	53.58	50.89	1.74	1.72	94.16	53.69	56.57	53.69	47.42
20	1.37	1.25	96.56	63.97	62.28	56.97	52.80	0.04	0.27	99.06	60.88	60.41	60.11	59.58
25	0.12	0.30	89.75	61.47	61.85	56.67	53.42	0.10	0.24	91.25	66.91	60.65	60.05	59.70
30	3.55	2.77	89.16	57.25	57.41	57.25	56.71	0.10	0.26	97.91	64.26	62.28	60.89	59.89

This baseline result is used to identify performance trends across the participants for various values of K. This shall help in establishing possible suggestions towards answering the research question for finding the optimum number of samples for state-of-the-art classification. Table. 6.4 highlights the value of K that provides the best average binary classification accuracy on the query set for both arousal and valence models. Since these values are *averaged over the participants*, these provide a general trend of the performance of the model for the dataset. Several observations can be made from these results. To compare the mean performances shown in Table. 6.4 across participants, these results are complemented with line plots showing the aforementioned mean values with the confidence interval across the four participants. These are discussed in the following sections.

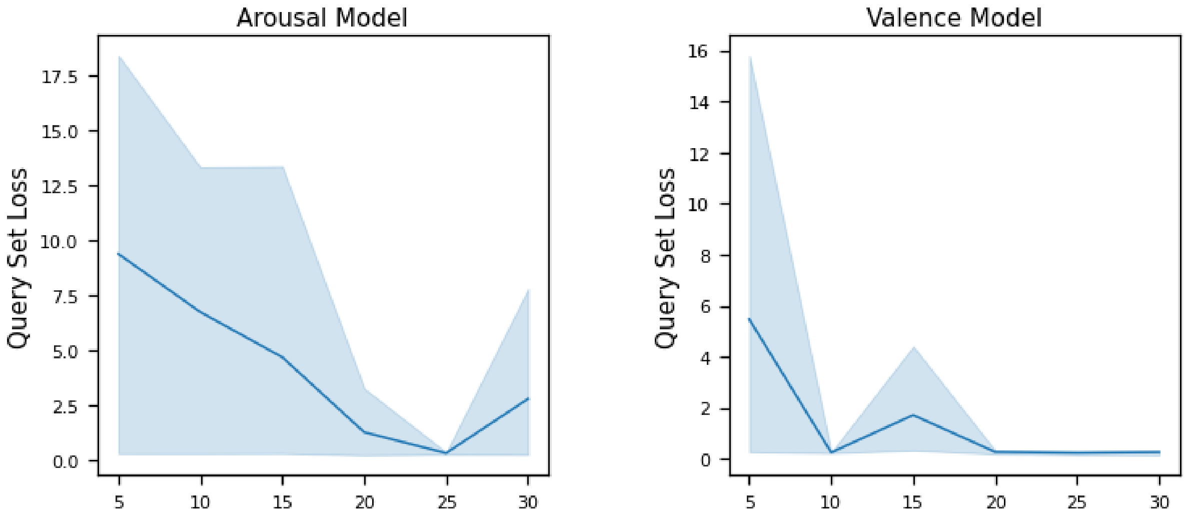


Figure 6.4: K-EmoCon : Distribution of Query set Contrastive Loss (L_q) for Baseline Model across Participants

A. Residual Contrastive Loss

Firstly, we observe the trends in *mean residual contrastive loss* across the value of K for the query set, in Fig. 6.4. The figure shows a line plot showing the spread of contrastive loss values across all the participants for the query set. As expected, an increase in the number of samples K decreases the overall loss as well as the skewness of loss across participants indicating homogeneous performance. An exception here is the arousal model for the values of $K = [30]$ with a larger spread in the loss. This reinforces that the model learns more and more as the support set samples K are increased, for both the emotion dimensions - arousal and valence. This plot also suggests that as K is increased, imbalances within the labels amongst participants have little impact on the model's ability to reduce the residual contrastive loss.

B. Binary Accuracy

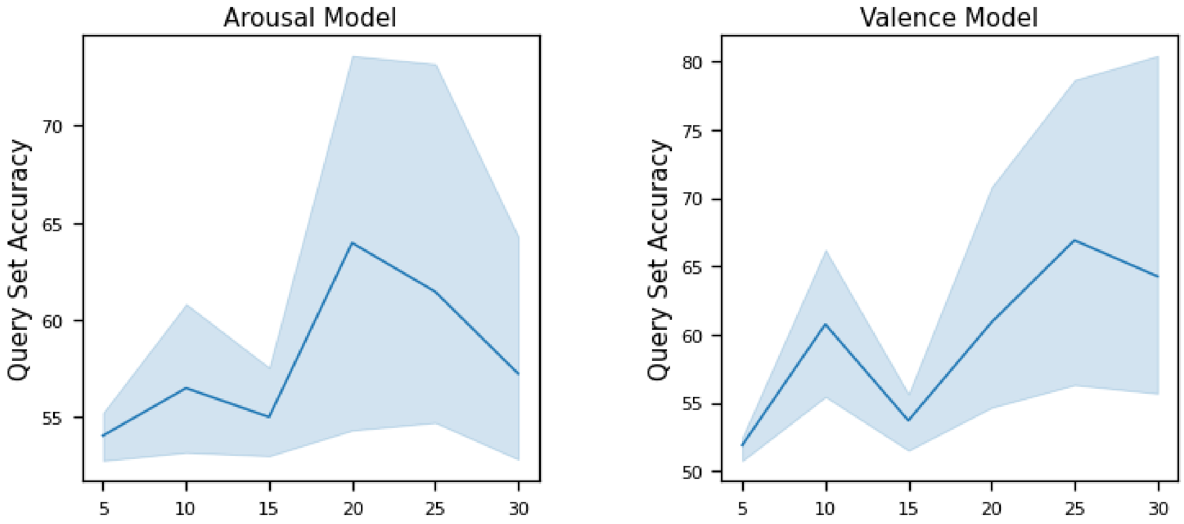


Figure 6.5: K-EmoCon : Distribution of Query set Binary Accuracy (A_q) for Baseline Model across Participants

The *mean binary classification accuracy* of the arousal and valence models for the baseline case, for the query set, are shown in Fig. 6.5. What stands out in this figure is the general pattern of increase and subsequent decrease in binary classification accuracy for both the arousal (with maxima at $K = 20$) and valence models (with maxima at $K = 25$). This drop in accuracy, past maxima may be reasoned by the fact that as the number of input samples increases, after a certain threshold, the model starts over-fitting on the query set. This suggests the inadequacy of the proposed architecture for learning from the increased number of samples. It is also visible, that the spread of accuracy values increases widely for higher K, as against lower values of K. This spread is indicative of the differences in the label distribution between the different participants. It can be concluded from these plots that residual contrastive loss cannot solely determine the model performance objectively, nor does it measure the generalizability of the Siamese networks.

C. Precision, Recall and F1 Score

Finally, the line plot of precision, recall and f1 score for the baseline case are shown in Fig. 6.6. It should be noted that the precision, recall and f1 score values plotted here are the *mean weighted scores* across individual participants, and are thus unbiased of the underlying label distributions of the participants. It is observed that the mean values of precision, recall and

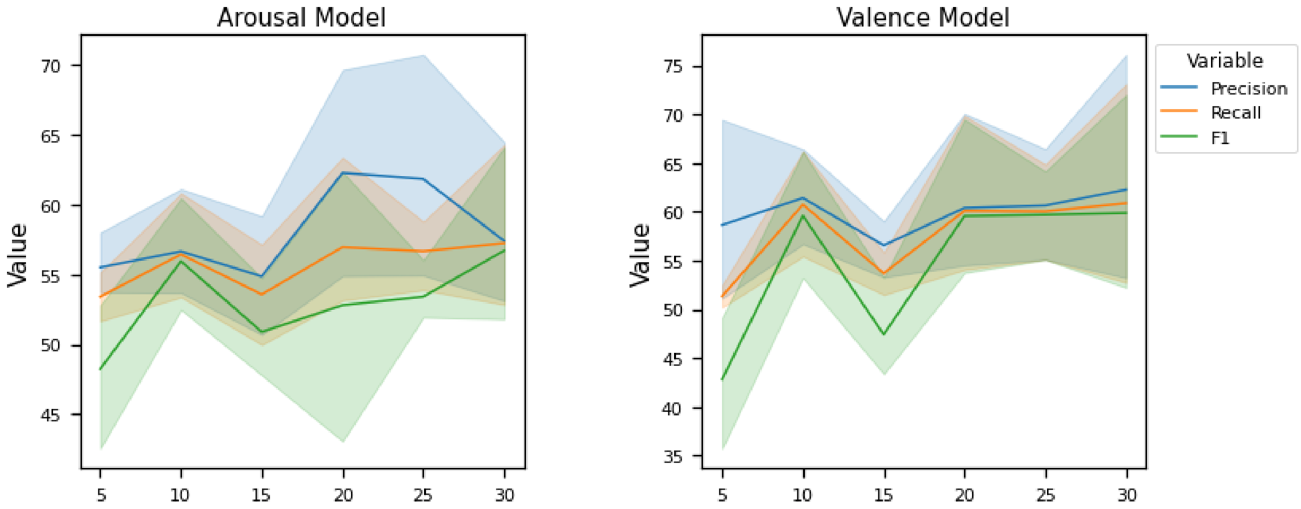


Figure 6.6: K-EmoCon : Distribution of Query set Precision (P), Recall (R) and F1 Score ($F1$) for Baseline Model

f1 score, across the participants, remain largely around 55%. Despite the higher prediction accuracies, the model shows discrepancies in learning across the (generated) pair labels. In other words, the model predicts one of the labels, more correctly than the other. The spread of the values also reiterates this discrepancy. In general, the models have high precision, as compared to recall, which in turn is higher than the f1 score. Here, the F1 scores deviate from lying between the precision and recall values since we use the weighted f1 score values for each of the participant cases to accommodate the label imbalance. For both arousal and valence variables, positive predictions tend to be more accurate, however, the actual relevant prediction is lower. Moreover, the f1 score indicates objectively the poor performance of the models for lower values of K. For instance, $K = [5, 15]$, is a practically insufficient amount of data, for reliable arousal and valence dimensional classification. Both the arousal and valence models barely learn anything and perform similar to or worse than a random classifier with f1 scores less than or around 50%. The highest values of f1 scores are obtained for $K = 30$ for both arousal and valence models (highlighted in red). It can be seen that both $K = [25, 30]$ result in favourable values of precision, recall and f1 score, denoting the generalized learning capability of the model being achieved at around these values of K.

In the next section, we briefly present the performance of the proposed architecture with different types of embedded noises.

6.2.2. Results with Embedded Noise

This section discusses the performances of the models for the prediction of arousal and valence dimensions in presence of different types of noises. The audio samples are corrupted by a noise sample at zero gain indicating that the overall power of the audio signal is maintained. This section, therefore, provides the basis for comparing model performances while examining the effects of noise and support size, simultaneously. This analysis is presented similar to the baseline results, focusing on contrastive loss, binary accuracy and the metric of precision, recall, f1 score individually. Table. 6.5 summarizes the performances of the dataset with the audio segments corrupted with the embedded noises. To concisely discuss and present the results, the mean performances across all participants are described in the table. Mean performances are described for each of the embedded noises separately. Line plots are used to compare and analyze performances across different embedded noises. A detailed look at all the results for individual participants is available in Appendix B in Section B.2.

A comparative discussion of the impact of the different types of impregnated noises on the model performances can now be presented. To cohesively present the results, line plots with confidence bars are used representing the spread of the values (across participants) around the mean values. The plots in the following sections are created using Table. 6.5. Throughout the discussion, we again focus on the mean values of the performance metrics and discuss the skewness of the values across this mean. This spread is the contribution of individual participants.

Table 6.5: K-EmoCon : Performance Metrics for different embedded noises

	Arousal Model							Valence Model						
K	L_s	L_q	A_s	A_q	P	R	$F1$	L_s	L_q	A_s	A_q	P	R	$F1$
DKITCHEN														
5	9.55	8.15	93.75	54.59	56.09	54.59	50.80	8.65	7.34	91.25	53.83	56.24	53.83	49.05
10	5.37	4.56	93.12	56.07	57.01	56.07	53.93	9.68	8.25	96.25	54.76	54.49	53.77	51.42
15	9.70	8.16	92.50	57.58	59.65	57.58	55.50	6.52	5.67	93.75	53.17	55.62	53.17	48.47
20	0.14	0.31	96.25	61.90	62.37	56.47	52.23	0.10	0.32	92.81	55.06	52.79	52.78	52.26
25	2.61	2.08	86.00	59.18	58.72	56.03	53.69	0.12	0.31	94.25	60.07	61.32	58.66	54.87
30	0.08	0.26	99.58	59.16	56.07	55.92	55.61	7.17	5.66	94.79	62.82	63.12	60.57	58.64
DLIVING														
5	0.04	0.38	100.00	51.63	50.34	50.46	47.71	11.25	9.53	95.00	55.40	57.12	55.40	53.07
10	0.13	0.28	90.62	56.00	55.90	54.05	51.88	3.69	3.22	95.62	54.98	54.61	54.31	53.68
15	2.60	2.36	94.17	57.36	57.63	57.23	56.71	5.45	4.73	99.17	52.98	53.09	52.98	52.53
20	16.64	12.00	76.88	57.84	58.51	57.84	57.01	0.05	0.28	97.81	62.43	60.66	60.07	59.26
25	0.10	0.26	97.50	67.25	65.49	59.04	54.25	7.36	5.66	88.25	64.78	67.88	64.02	60.38
30	0.20	0.43	90.83	57.06	58.92	57.06	53.60	7.24	5.68	91.46	61.68	63.41	61.10	58.61
OHALLWAY														
5	13.10	11.08	85.00	54.60	63.42	54.60	44.54	0.19	0.36	95.00	54.36	51.75	48.52	39.37
10	4.62	3.97	86.25	54.63	54.89	54.63	53.98	12.20	10.36	92.50	55.14	58.47	55.14	50.28
15	1.44	1.52	100.00	56.16	58.46	56.16	52.34	2.18	2.04	91.67	58.63	61.58	57.78	52.16
20	0.06	0.30	97.81	60.22	59.44	58.78	57.69	0.20	0.44	95.00	56.11	57.94	53.26	49.21
25	3.57	2.77	91.50	59.92	60.74	56.46	52.74	0.08	0.26	91.75	63.65	60.28	59.60	58.97
30	0.11	0.28	93.96	62.23	65.31	57.28	51.46	0.07	0.30	97.91	56.95	55.55	55.22	54.20
OOFFICE														
5	6.35	5.35	90.00	56.72	56.77	56.20	55.02	3.81	3.39	90.00	53.15	55.41	53.15	47.49
10	6.37	5.42	92.50	56.11	56.38	56.11	55.62	3.13	2.72	86.25	53.84	56.21	53.27	48.77
15	10.03	8.46	90.42	56.21	57.45	56.21	54.64	7.64	6.65	95.41	53.22	55.86	53.22	47.43
20	0.10	0.27	94.69	57.45	56.22	55.81	55.10	0.10	0.33	93.12	55.38	55.34	54.65	53.38
25	0.07	0.28	98.00	61.02	59.55	57.09	53.53	0.08	0.23	91.50	66.52	64.04	63.59	62.78
30	0.09	0.26	96.46	59.73	59.69	59.16	57.58	0.05	0.30	98.75	58.40	57.08	56.97	56.70

A. Residual Contrastive Loss The residual contrastive loss shown in Fig. 6.7 gives an overview of the learning behaviour of the arousal and valence models. The plot consists of loss values for different cases including baseline and the different types of noises, plotted against the values of K. A recurring observation for the residual contrastive loss values for the arousal models is the gradual decrease in the loss values with an increase in K. This decrease is smooth for the $Baseline$ case, while with different noises the trend is disrupted. The case with the noise of type $OHALLWAY$ follows the $Baseline$ closely. An unexpected

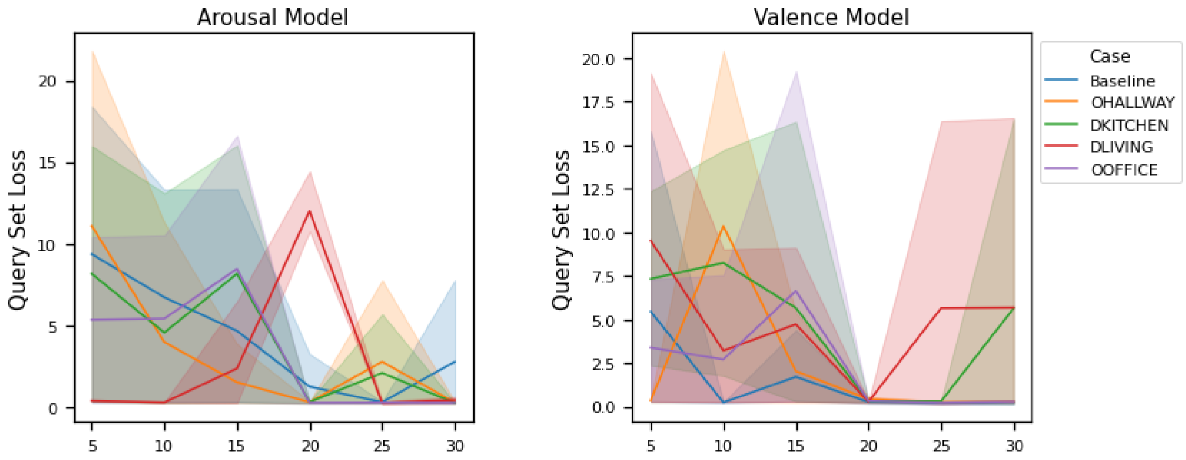


Figure 6.7: K-EmoCon : Summary of Query Set Contrastive Loss (L_q) for different cases

observation is made with the noise of type $DLIVING$, where the residual loss values follow an inverted parabolic trend of increase and subsequent decrease. The minima of the loss values for the arousal model is spread across different values of $K = [20, 25, 30]$.

Now compared to this, the valence models have different behaviour with the residual loss values. Firstly, the $Baseline$ case again follows a decreasing trend in loss with an increase in K. The $Baseline$ case is the one with the least residual loss for the valence models, amongst all the cases. A fluctuating trend of this residual loss with different values of K may indicate the overall difficulty of generalization. As discussed earlier, the spread of the residual loss indicates the effect of differential label distribution across participants. It may also be inferred that the residual loss is not a sufficient metric for the comparison of models for different support set sizes. The results of other performance metrics in conjunction with residual loss can be used to comment on the classification performance of the models.

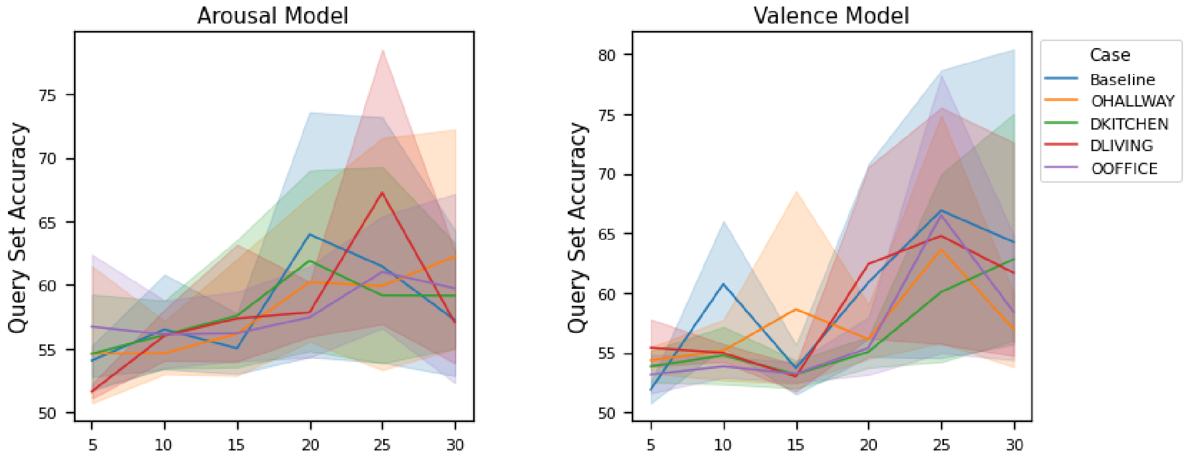


Figure 6.8: K-EmoCon : Summary of Query Set Binary Accuracy (A_q) for different cases

B. Binary Accuracy The line plots of variation of binary accuracy across the value of support set K is shown in Fig. 6.8. It can be observed that the binary accuracy shows an upward trend with increasing K. The mean $Baseline$ accuracy values, however, are not the maximum values when compared with the different cases of noises. The maximum accuracy is achieved with the inclusion of the noise $DLIVING$. Interestingly, the overall spread of accuracy values across participants remains overlapping for different values of K, across all the types of noises. Further, maxima of binary accuracy across participants occur at $K = 25$. This indicates the optimal support set size for the arousal models.

For the valence models, the line plots of binary accuracy show an almost knew-point behaviour for $K = 15$, above which the accuracies increase dramatically. This indicates the minimum support set size for classification. Further, the maxima of accuracies throughout the different cases occur at $K = 25$, echoing the behaviour of arousal models.

For different types of noises, the value of K at which the maximum accuracy occurs is different. This indicates the impact of noise is dependent on the support set composition. If we take a look at the skewness of the mean binary accuracy, the impact of K is readily visible. While lower values of K appear to exhibit similar performances across all the participants, larger values of K differentiate the participants greatly, since the support set composition is skewed. This composition affects the performance of Siamese networks owing to pair-wise comparison in the architecture and the contrastive loss function. The comparison of similarity (or dissimilarity) of pairs, may result in a larger reduction of loss and subsequent higher accuracy for highly skewed label distribution. This is because a skewed label distribution results in highly contrasting pairs - effectively increasing the ability of the network to discriminate the pairs. Another observation, (in conjunction with residual contrastive loss values) is the optimal value of K in presence of noise.

It is found that $K = 25$ achieves in general, the maximum binary accuracy with a reasonable compromise on residual contrastive loss for the K-EmoCon dataset.

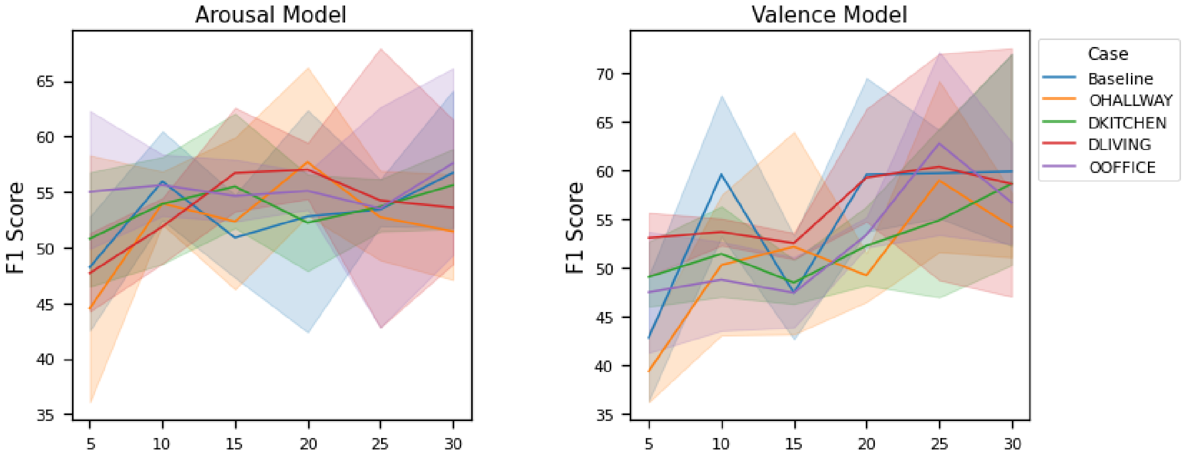


Figure 6.9: K-EmoCon : Summary of Query Set F1 score ($F1$) for different cases

C. F1 Score For brevity of discussion, we focus on the F1 score and not on precision and recall. The f1 score line plots for the arousal and valence are shown in Fig. 6.9. As stated earlier, the mean f1 scores provide an objective outlook at the performance of the proposed model against different settings of K and noises. We first discuss the arousal model results. Immediately visible is the poor scores of the arousal models for all cases for $K = 5$. This strengthens the hypothesis presented earlier, regarding the inadequacy of this value of K to perform any meaningful classification. The performance is less than or close to 50% for all the cases, except for the case of $OOFFICE$. Further, the skewness of the f1 score across the mean is similar for all values of K. The models, throughout perform moderately with average values existing around 50 - 55%, for all the cases, including baseline. Incidentally, the $Baseline$ case scores the least mean f1 score against all the noises for all values of K except 30. This behaviour is rather unexpected and requires further investigation.

For the valence models, the f1 scores provide a generally upward trajectory, as opposed to the behaviour observed with arousal models. Here, the $Baseline$ case also has a maximum f1 score over all the cases with different noises, over most of the values of K. The case with the noise of type $DLIVING$ performs as closely as the $Baseline$ case for higher values of K. This indicates the ability of the model to overcome the degradation introduced by the musical noise with higher support set size. It can also be seen that the valence models possess better F1 scores than their arousal counterparts for the same values of K. This is a departure from the usual trends in emotion recognition models, where the valence models perform poorly as compared to the arousal models, owing to the limited contribution of the acoustic features towards the valence dimension as compared to the arousal dimension.

Summary The examination of the average performance of all the participants of the K-EmoCon dataset gives several important results. From the discussion presented earlier, it can be concluded that the residual contrastive loss is not a clear indicator of performance when comparing models for different support set sizes K. Next, a look at the binary accuracy plots in Fig. 6.8, indicates the optimal support set size for this dataset. The maxima of accuracies occur at $K = 25$, for both the valence and arousal models, thus suggesting a near-optimal value. Further, the subsequent drop in accuracy for $K = 30$ may indicate the overfitting nature of the model against the support set size. This behaviour calls for further inquiry with different architectural hyper-parameters. From the plots of f1 score in Fig. 6.9, it is also visible that the binary accuracy, is not the best indicator of the performance of this dataset, owing to the individual label distributions across participants. Further, it also reinforces the optimum K with an overall maximum at $K = 25$.

An interesting observation throughout the models is the fluctuation in binary accuracy and f1 scores with $K = [5, 10, 15]$. This may be explained by the fact that the batch size of the models is 16. This is because support set size of $K = 10$ results in batches with the disproportionate count of samples - with 20 samples and batch size of 16, the two created batches have 16 samples and 4 samples respectively. This may result in impartial averaging across the two batches resulting in a large increase in accuracy and f1 scores at $K = 10$, which stands out against the other cases.

The large skewness of the performance metrics shown in the figures earlier implies the need for participant-specific analysis. This may give deeper insights into the impacts of label distribution across the various cases. In the next section, we briefly present the participant-specific results of the K-EmoCon dataset across all the cases. The analysis of the results is restricted to the metrics of binary accuracy, precision, recall and f1 score as these describe the individual results concisely.

6.2.3. Results with Individual Participants

We next analyze the results of individual participants from the K-EmoCon dataset against the various values of K and imputed noises. Here, a comparison of the models is made for individual participants, therefore accurately describing the prediction abilities of the model for a single speaker. This section, therefore, highlights the variance in performance while considering the compositions of the query set sizes in conjunction with the test parameters. Table. 6.6 summarizes the *best performing models* of all the participants of the dataset across various cases. The table shows the results of the best performing value of K for all the cases. A detailed look at all the results for individual participants is available in Appendix B.

Table 6.6: K-EmoCon : Performance Metrics for Individual participants [3]

Case	Arousal Model						Valence Model					
	K	A_s	A_q	P	R	$F1$	K	A_s	A_q	P	R	$F1$
	Participant ID - 15											
Baseline	**25**	79.00	57.74	57.99	57.74	57.40	**25**	96.00	70.06	62.11	62.10	62.10
DKITCHEN	**20**	95.00	56.08	57.04	56.08	54.52	**25**	98.00	74.04	69.88	69.87	69.87
DLIVING	**10**	82.50	55.87	56.14	55.87	55.38	**25**	97.00	65.58	66.49	65.58	65.11
OHALLWAY	**20**	93.75	55.52	57.18	55.52	52.80	**25**	92.00	62.10	60.27	58.92	57.52
OOFFICE	**25**	98.00	57.60	57.63	57.60	57.56	**25**	97.00	70.25	66.50	66.46	66.43
	Participant ID - 23											
Baseline	**25**	100.00	78.61	75.01	59.44	51.97	**25**	94.00	59.84	58.69	58.27	57.75
DKITCHEN	**20**	100.00	65.53	65.30	61.32	58.62	**30**	100.00	57.02	57.07	57.02	56.95
DLIVING	**25**	100.00	84.55	75.43	51.69	36.97	**10**	100.00	55.92	56.13	55.92	55.54
OHALLWAY	**30**	100.00	77.84	77.59	59.38	51.35	**30**	100.00	57.50	53.43	53.33	53.01
OOFFICE	**25**	98.00	66.67	68.18	66.67	65.96	**25**	74.00	55.24	56.12	55.24	53.57
	Participant ID - 30											
Baseline	**25**	84.00	56.88	61.71	56.88	51.92	**20**	100.00	57.63	56.13	55.34	53.86
DKITCHEN	**15**	100.00	57.79	65.13	57.79	51.96	**30**	100.00	57.83	52.18	52.17	52.12
DLIVING	**25**	91.00	60.45	60.53	60.45	60.39	**20**	92.50	57.31	59.05	57.31	55.15
OHALLWAY	**20**	100.00	55.60	55.88	55.60	55.08	**25**	100.00	56.91	50.43	50.41	49.66
OOFFICE	**20**	81.25	56.36	57.55	56.36	54.55	**30**	99.17	55.93	56.19	55.93	55.47
	Participant ID - 31											
Baseline	**20**	100.00	70.86	67.81	66.55	65.94	**30**	100.00	88.59	82.41	78.84	78.24
DKITCHEN	**25**	100.00	74.44	70.61	61.85	57.32	**30**	100.00	81.04	81.68	77.71	76.99
DLIVING	**25**	100.00	70.52	70.72	70.52	70.45	**25**	96.00	82.11	80.07	79.07	78.89
OHALLWAY	**20**	100.00	70.76	70.34	70.04	69.92	**25**	93.00	80.89	75.72	74.39	74.05
OOFFICE	**30**	100.00	68.92	66.91	66.60	66.45	**25**	97.00	85.91	78.17	77.98	77.94

For analysis of the performance of individual participants, we omit residual contrastive loss. Within a single participant analysis, the distribution of the query set is uniform and therefore, the residual contrastive loss does not inform any additional information of the model behaviour. Therefore, we focus on binary accuracy and precision, recall and f1 scores for our analysis.

A. Binary Accuracy From Table. 6.6, the binary accuracy of the query set indicates a remarkable result. For the arousal model, we immediately see the better performing participants as opposed to the weaker ones. With 78.61% and 70.86% on the $Baseline$ case, participants 23 and 31 respectively, score dramatically higher accuracies, as compared to participants 15 and 30. This performance is sustained even with the inclusion of different types of noises, as

[3] Abbreviations :- A_s : Support Set Accuracy; A_q : Query Set Accuracy; $P(\%)$: Precision (in %); $R(\%)$: Recall (in %); $F1$: F1 Score (in %).

clearly seen in the results (highlighted in red). Conversely, for the valence model, the best performing participants are 15 and 31 with binary accuracies of 70.06% and 88.59% respectively, on the baseline. Similar to the trend in the arousal model, these participants perform better even in presence of noise.

B. Precision, Recall and F1 Score When we look at the precision, recall and f1 scores amongst the participants, we can see a relatively different picture. While much of the trends are similar to binary accuracy for all the participants, there are some deviations to this. The maximum f1 scores for arousal and valence models are with participants 31. However, amongst other participants, the maxima of f1 scores for arousal and valence models are not consistent with the binary accuracy values. This behaviour highlights the specificity captured by the proposed architecture concerning differences in characteristics of individual participants. This hypothesis, ties into the differences in binary accuracy amongst participants discussed earlier.

What is even intriguing is that the best performances for the two emotion dimensions - arousal and valence may or may not reside with the same participants. This is evident from the discussion above - while participant 31 has the best performances for both arousal and valence dimensions, participants 15 and 23 fares superior only on **one** of the dimensions - valence and arousal, respectively. It is suspected that the individual label distributions of the participants contribute heavily to the accuracy of the arousal and valence models. This observation requires further investigation of labels. This is discussed in the following section.

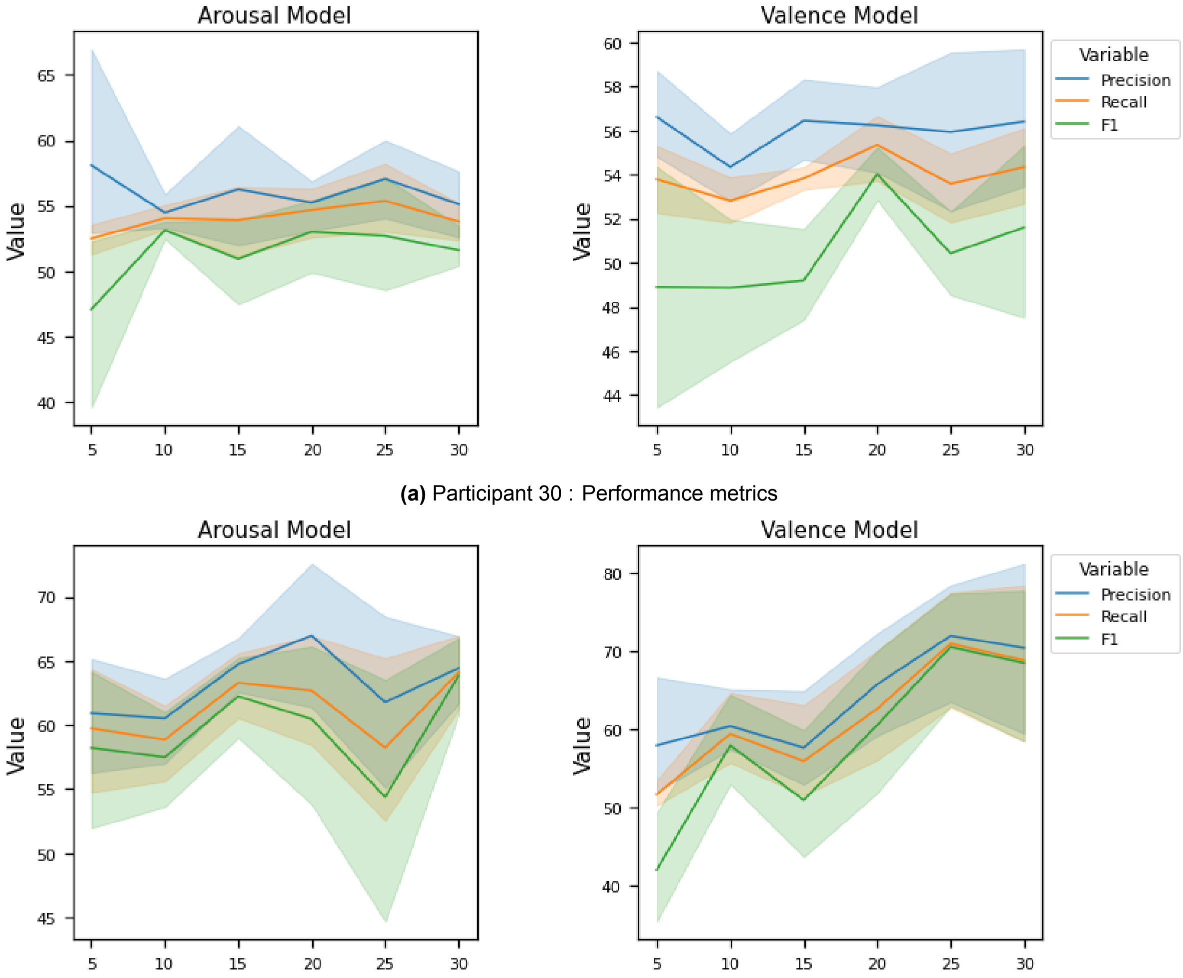
Investigation on Label Distribution

Table 6.7: Distribution of Labels across participants of K-EmoCon

Participant	Arousal		Valence	
	0	1	0	1
15	105	111	131	69
23	42	181	109	62
30	83	67	101	67
31	205	110	50	242

We take a look back at the distribution of labels in the K-EmoCon dataset. Upon analyzing the obtained results with the label distribution shown in Table. 6.7, we can infer several important observations. For instance, the better performing participants for the arousal dimension are participants 23 and 31. Incidentally, the arousal labels for these two participants show the maximum imbalance between the 0 (low) and 1 (high) labels. On the other hand, for the valence models, participants 15 and 31 perform better than the others. Comparing their label distributions, we see a significant imbalance in these participants against the others. To objectively reason the imbalance, we consider an imbalance when a participant has twice the samples (or greater) in one label than the other. Another observation on these lines is the variation of performance with this imbalance. Participant 31 which has the highest skewed label distribution for the valence dimension, has the highest overall performance among the 4 participants.

This highlights the importance of the f1 score in performance analysis since the query set deals with unbalanced label distributions. This imbalance in the ground labels of the samples is sustained in the pair creation process. While, the pairs are generated such that there are an equal number of pairs for each (pair) labels (pair labels of classifier being *similarity* denoted by 1 and *dissimilarity* denoted by 0), the model's ability to predict this is affected by imbalance presented in the query set.



(a) Participant 30 : Performance metrics

(b) Participant 31 : Performance metrics

Figure 6.10: Comparison of Performance metrics of Participant 30 and 31

To attest to this behaviour, we compare the precision, recall and f1 scores of two participants 30 and 31. While participant 31 has a large imbalance in the label distribution, participant 30 has an almost equitable distribution of labels. Fig. 6.10 compares these results for various values of K for the two participants. It is visible that the precision, recall and f1 scores for participant 31 are consistently higher than that of participant 30 for both the dimensions - arousal and valence. Comparing all the performances across different noises, an inverted parabolic behaviour is visible for arousal and valence dimensions. The confidence intervals across the different cases narrow with the increase in K.

6.3. Experiments with RECOLA

For the second task representing the scenario of IVA actively responding to short prompts from the user, experiments are performed on RECOLA. The pre-processing of the RECOLA dataset gives 18 selected participants with $400ms$ sample. In this section, for brevity, results from 4 participants are presented. These are $[dev_1, dev_2, dev_3, dev_4]$. The samples are processed with the processes described in the previous sections to obtain the audio features. Thereafter, the segmented dataset is split into support and query sets followed by pair generation within these sets.

Support Set Size

Similar to the approach taken with K-EmoCon, here we select support set sizes after proportionating the size of the dataset per participant. For the one-shot interaction of **5** $-$ **7s** [101] [59], and sample duration of $0.4s$, the value of K representing one-shot interaction is 10 equivalent of $8s$ (slightly higher than the prescribed duration). The range of shots is chosen to be multiple of this value in the range of $[\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}]$. Given the data samples shown in Table. 4.9, the value of the support set size i. e., K is selected as shown in Table. 6.8. The table also summarizes the support set as a percentage of the participant dataset. This value of K is the basis for the first set of experiments performed on the selected participants. The models are trained separately to predict binary valence and arousal classes.

Table 6.8: Support Set Size for RECOLA

Variable	Values					
K	10	20	30	40	50	60
$N \times K$	20	40	60	80	100	120
interactions	1	2	3	4	5	6
$\%Data$	2.67	5.33	8	10.67	13.33	16

Learning rate Range Test for RECOLA

The results of the learning rate range tests for samples from RECOLA are summarized in Table. 6.9. for arousal and valence respectively. For RECOLA, the large number of samples available in the dataset for each participant allows the choice of large batch sizes as compared to K-EmoCon. Thus here the range of batch sizes tested is large. The optimal batch size for both arousal and valence models is 256.

Table 6.9: Hyper-parameter Tests for RECOLA

Model	**Hyper-parameter**	**Range of Values tested**	**Best value/ Optimal Range**
Arousal	learning rate	$[1e^{-7}, 1e^{-2}]$	$2e^{-4}$
	batch size	$[96, 128, 256]$	256
Valence	learning rate	$[1e^{-7}, 1e^{-2}]$	$2e^{-4}$
	batch size	$[96, 128, 256]$	256

- **Arousal Model**: The optimal batch size for arousal models is 256 as shown in shown in Fig. 6.11. While the steepest descent in query set loss is achieved by other learning rates, it still manages to reduce support set loss to a minimum. Further, it also achieves the maximum accuracy of all the batch sizes.

- **Valence Model**: For the valence model, batch size of 256 achieves a compromise between least support set loss and steepest descent in query set loss. This is visible in Fig. 6.12. Since the learning goal is based on the reduction of contrastive loss, a proper metric shall reduce both support set as well as the query set contrastive loss. In the case of data for valence prediction, this is achieved by the batch size of 256.

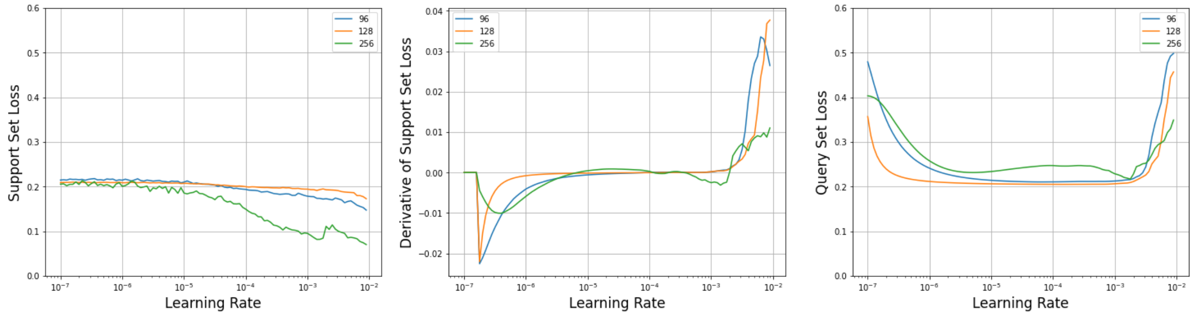


Figure 6.11: RECOLA : Learning rate Range Test Contrastive Loss for Arousal Model

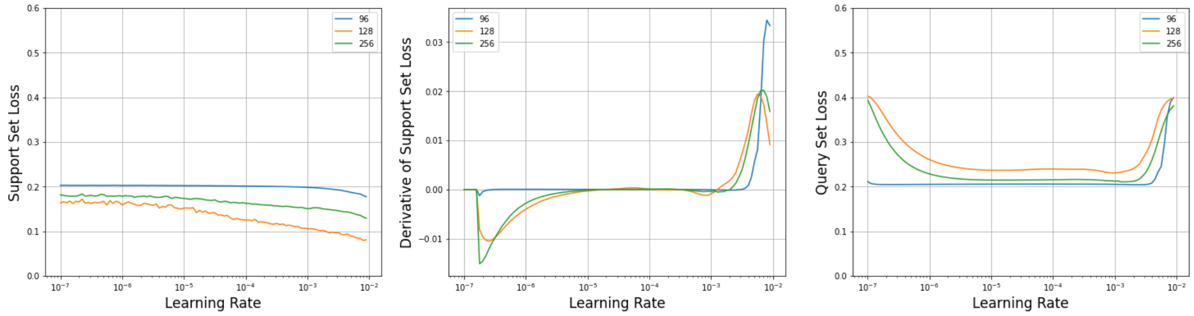


Figure 6.12: RECOLA : Learning rate Range Test Contrastive Loss for Valence Model

In the following sections, experiments for the four selected participants are discussed.

6.3.1. Results with Baseline Case

This section follows thematically from the Baseline results of the K-EmoCon dataset discussed earlier in 6.2.1. In Table. 6.10, the baseline results of our proposed Multimodal Siamese Network for the arousal and valence dimensions are presented. The values in this table represent the mean values obtained from the four selected participants for each value of K. The results are analyzed from the point of view of the following performance metrics - residual contrastive loss (for support set (L_s) and query set (L_q) *pairs*), binary classification accuracy (for support set (A_s) and query set (A_q) *pairs*). The precision (P), recall (R), f1 score $(F1)$ are weighted for the labels. Again, for brevity, only the query set results are presented here. Except for the residual contrastive loss, all the other metrics are expressed in percentage. A detailed look at all the results for individual participants is available in Appendix B in Section B.3.

From Table. 6.10, we see that both the arousal and valence models perform relatively poorly. We discuss the results presented in the table above in conjunction with line plots showing the spread of metrics around the mean. This is explained in the following sections.

Table 6.10: RECOLA : Performance Metrics for Baseline Models [3]

	Arousal Model							Valence Model						
K	L_s	L_q	A_s	A_q	P	R	$F1$	L_s	L_q	A_s	A_q	P	R	$F1$
10	17.56	15.18	67.50	50.79	45.97	50.79	41.48	6.51	5.60	93.12	51.80	46.05	51.30	42.05
20	1.79	1.86	90.62	51.66	47.11	51.64	43.35	7.95	6.68	91.56	52.10	51.12	50.47	39.62
30	1.49	1.61	95.83	53.78	54.52	53.78	48.74	3.74	3.35	89.59	52.45	31.54	50.29	37.78
40	2.42	2.32	85.00	52.09	56.56	52.09	45.35	2.21	2.05	89.84	53.87	48.10	51.33	42.42
50	4.90	4.31	89.12	53.81	47.82	53.81	48.78	15.50	13.19	76.00	52.10	45.31	51.24	40.95
60	0.78	0.99	92.92	51.55	43.98	50.22	37.37	5.56	4.70	79.90	52.58	46.20	51.76	44.98

A. Residual Contrastive Loss

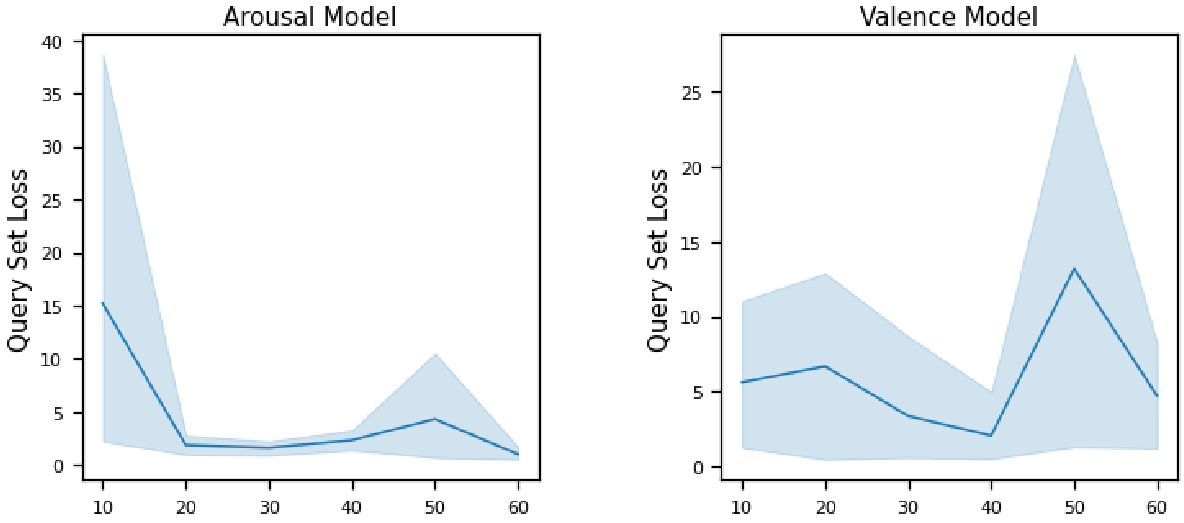


Figure 6.13: RECOLA : Distribution of Query Set Contrastive Loss (L_q) for different cases

The residual contrastive loss values tend to lower with an increase in the number of shots K, however, there are exceptions to this trend with $K = 50$. Several reasons could be attributed to this behaviour. Firstly, it is suspected that early-stopping with some participants triggers the models to stop early in the training phase, thus resulting in a large residual loss. This also indicates that the model fails to learn any substantial discriminating characteristics past some early epochs. Further, we see in Fig. 6.13, that arousal and valence models exhibit different trends in residual losses. There is a clear increase in loss value for larger K indicating difficulties in learning. This higher loss value may also be due to the overfitting behaviour of the model. Incidentally, the valence model has a larger spread of residual loss as compared to arousal models throughout all the values of K. This may be indicative of the relative difficulty in predicting the valence dimension.

B. Binary Accuracy

Fig. 6.14 shows the results of mean performances of the model for arousal and valence dimensions. Here, the variation in K results in an increase and subsequent decrease in binary accuracy values for both dimensions. Interestingly, the arousal model has two peaks at

[3] Abbreviations :- A_s : Support Set Accuracy; A_q : Query Set Accuracy; $P(\%)$: Precision (in %); $R(\%)$: Recall (in %); $F1$: F1 Score (in %).

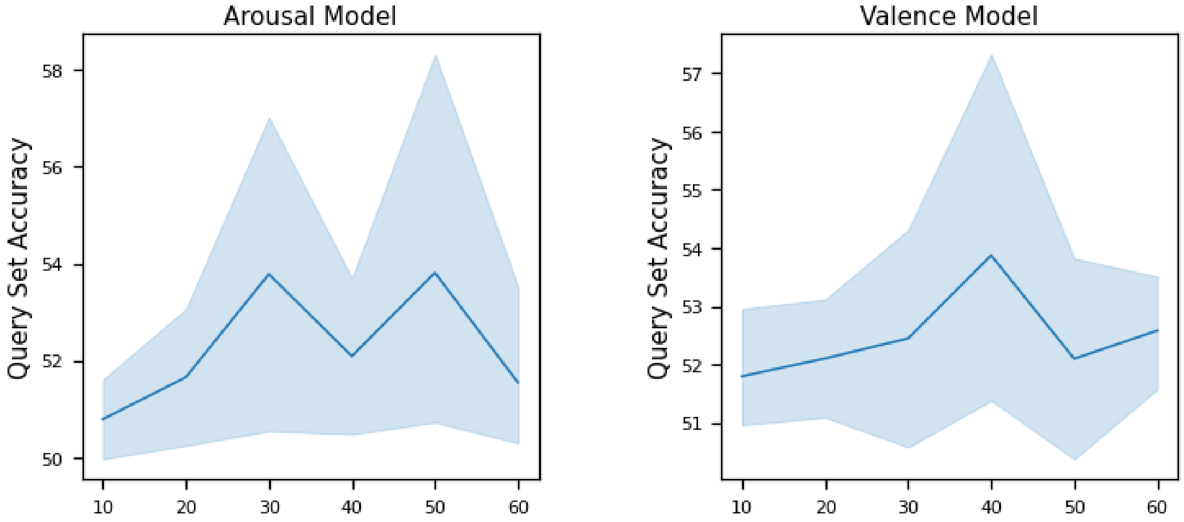


Figure 6.14: RECOLA : Distribution of Query Set Binary Accuracy (A_q) for different cases

$K = [30, 50]$, while the valence model has a single peak at $K = 40$. However, it is also worthwhile to note that the mean values lie significantly in a very narrow range from 50 - 55%, for either of the dimensions. Therefore, while the overall performance may be said to improve, the quantum of improvement is very minimal. For individual participants, the performance of the arousal and valence dimension predictions can be seen to remain within 50 - 58%, as visible from the confidence intervals shown in Fig. 6.14. This indicates that the proposed architecture struggles in learning from the support set, throughout the selected participants. The similarity of this behaviour for both the emotion dimensions, suggests some performance bottleneck associated with the dataset. This behaviour is notably different from the $Baseline$ case for the K-EmoCon dataset. This suggests some bottlenecks associated with the RECOLA dataset.

C. Precision, Recall and F1 Score

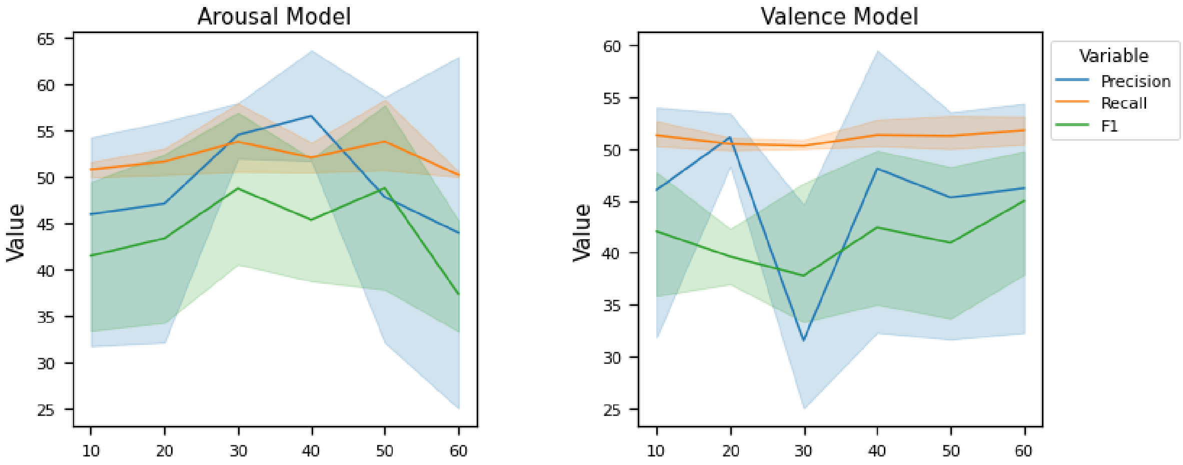


Figure 6.15: RECOLA : Distribution of Query Set Precision (P), Recall (R) and F1 Score ($F1$) for different cases

Next, we take a look at the precision, recall and f1 scores of the arousal and valence predictions for the $Baseline$ case. These results are shown in Fig. 6.15. The line plots denote the mean weighted scores across individual (selected) participants of the RECOLA dataset. The confidence intervals denote the span of these variables for individual participants. Unlike the plots of binary accuracy, Fig. 6.15 indicates metric variables well below 50%. This indicates that the actual predictive power of the algorithm is quite poor. The recall of the predictions for both dimensions is higher in general as compared to the precision. This indicates that the total positive rare (or sensitivity) of the predictions is better, however, over-predicts on the positive class i. e., similarity label (1). In other words, the models are poor at learning similarity of samples, than that of learning dissimilarity. The precision, however, improves briefly for arousal prediction for $K = [30, 40]$, before dropping again. Conversely, the valence dimension lacks any such trend. The f1 scores further indicate the overall performance of the proposed architecture across the two classes is poor, throughout the values of K. The scores are consistently below 50%, indicating a worse performance than a random classifier.

This indicates a case of possible ill-conditioning of the class precision and f1 score. In several instances, one of the classes is never predicted, indicating that the model completely fails at learning anything about that class. This underpins the idea that dissimilarity traits are easily learned by the model as compared to similarity traits, in this specific dataset. This behaviour contradicts heavily that of the K-EmoCon dataset. In the next section, the results of the RECOLA dataset are tested with embedded noise.

6.3.2. Results with Embedded Noise

This section discusses the performances of the models for the prediction of arousal and valence dimensions in presence of different types of noises. The audio samples are corrupted by a noise sample at zero gain indicating that the overall power of the audio signal is maintained. This section, therefore, provides the basis for comparing model performances while examining the effects of noise and support size, simultaneously. This analysis is presented similar to the baseline results, focusing on contrastive loss, binary accuracy and the metric of precision, recall, f1 score individually. Table. 6.11 summarizes the performances of the dataset with the audio segments corrupted with the embedded noises. To concisely discuss and present the results, the mean performances across all participants are described in the table. Mean performances are described for each of the embedded noises separately. while line plots are used to compare and analyze performances across different embedded noises. A detailed look at all the results for individual participants is available in Appendix B in Section B.4.

Line plots with confidence bars are used to represent the spread of the values (across participants) around the mean values. The plots in the following sections are created using Table. 6.11. Throughout the discussion, we again focus on the mean values of the performance metrics and discuss the skewness of the values across this mean.

A. Residual Contrastive Loss From Table. 6.11, the residual contrastive loss values for the different embedded noises show the absence of any specific trends with K. This is quite an important observation, as it indicates the inability of the model to meet the objective of reducing contrastive loss across the board. While a general drop in residual loss may be seen in arousal models, the same cannot be seen with valence models. Fig. 6.16 shows the line plots of the mean values of residual loss together with the spread across participants. This provides a comparison of the $Baseline$ case with that including different embedded noises. The arousal models show an overall trend of drop in residual loss with an increase in K until higher values of K with some exceptions. The minima of loss, for various trails, occurs at $K = [40, 50]$. It is also where the loss becomes homogeneous across participants, indicating similar embedding losses.

Table 6.11: RECOLA : Performance Metrics for different embedded noises [3]

	Arousal Model							Valence Model						
K	L_s	L_q	A_s	A_q	P	R	$F1$	L_s	L_q	A_s	A_q	P	R	$F1$
DKITCHEN														
10	12.24	10.44	82.50	52.55	54.10	52.55	46.14	11.35	9.63	88.12	52.28	55.97	52.28	43.85
20	16.08	14.05	78.44	51.99	47.34	51.99	41.33	7.12	6.01	89.38	51.28	46.31	50.96	40.35
30	15.03	13.20	82.92	50.97	39.02	50.88	39.98	11.16	9.36	76.66	52.84	54.87	52.84	47.96
40	0.28	0.49	99.37	52.82	37.99	49.95	36.74	3.78	3.39	85.78	51.30	51.56	51.30	48.46
50	7.42	6.28	82.00	53.09	52.81	52.99	47.45	3.55	3.18	87.12	53.57	45.62	51.36	43.89
60	4.64	4.09	84.59	52.77	51.03	52.40	46.89	15.77	13.71	82.08	53.38	49.02	51.77	42.69
DLIVING														
10	0.54	0.81	98.75	52.98	47.96	52.17	42.20	12.68	10.86	88.12	52.00	53.14	51.74	45.06
20	1.52	1.64	96.56	54.59	48.67	54.37	46.84	4.87	4.37	88.44	53.49	54.15	53.23	49.88
30	16.97	14.69	78.54	51.05	47.18	51.05	40.74	3.71	3.15	83.12	53.77	40.34	52.16	41.82
40	1.52	1.65	90.31	52.82	57.58	52.82	46.63	1.66	1.67	92.34	51.67	52.68	51.67	45.28
50	1.35	1.48	93.38	55.75	56.14	55.75	54.62	14.19	12.12	78.00	52.12	55.15	52.12	45.59
60	8.09	6.93	70.21	50.59	56.44	50.59	36.64	5.97	5.07	77.92	53.38	47.31	52.17	44.99
OHALLWAY														
10	22.68	19.41	70.00	50.86	46.38	49.90	37.90	4.49	4.02	89.38	54.53	56.08	54.53	50.81
20	0.99	1.15	96.25	56.17	51.58	54.33	46.30	1.65	1.62	93.44	53.79	48.67	53.12	45.11
30	1.21	1.23	95.42	55.26	47.41	52.41	45.20	4.63	4.03	85.21	54.47	56.78	53.89	47.34
40	3.66	3.20	86.88	53.04	46.85	52.74	47.54	18.47	15.88	80.94	52.50	57.63	51.54	40.90
50	0.78	0.96	97.25	54.78	33.29	51.78	38.93	1.51	1.66	91.38	54.12	54.85	53.29	49.98
60	10.73	9.24	84.79	53.72	56.25	52.62	43.95	2.60	2.37	87.19	52.99	45.63	51.77	46.05
OOFFICE														
10	1.34	1.42	95.00	52.80	51.62	51.58	45.69	1.70	1.79	96.88	53.81	54.25	53.76	51.19
20	14.55	12.35	81.25	52.72	55.21	52.72	43.13	10.31	9.02	87.50	53.00	51.91	51.98	43.76
30	1.75	1.78	89.38	52.98	46.87	52.98	47.57	13.18	11.09	82.71	51.95	53.59	51.24	44.35
40	8.77	7.35	77.97	55.77	47.38	53.27	47.85	1.59	1.73	88.91	53.95	48.46	53.87	47.73
50	1.42	1.50	88.62	53.95	47.93	51.64	40.33	4.08	3.67	86.00	55.31	47.71	53.17	46.95
60	11.76	9.96	71.04	53.09	47.67	52.97	40.88	8.26	7.12	84.68	51.65	43.64	49.92	39.49

Contrasting this visible trend, are the results from the valence model, where the loss fluctuates for all the different cases. Interestingly, similar to the observations made in the $Baseline$ case earlier, all the cases of embedded noises indicate a valley-like minimum at $K = 40$ except for the noise of type $DLIVING$, which has rather a maximum. While this is the mean residual loss across participants, the spread of the loss values follows similarly and hence this observation hold in general for any participant in this dataset. The cause of these values comparative of the $Baseline$ is difficult to interpret and require further analysis. It may be suspected that embeddings of this dataset are greatly affected by noise.

B. Binary Accuracy We analyze the mean binary accuracy across the various cases with Fig. 6.17. Firstly, the arousal dimension prediction deviates dramatically with the inclusion of the noise. Surprisingly, some of the cases of imputed noises result in increased binary

[3] Abbreviations :- A_s : Support Set Accuracy; A_q : Query Set Accuracy; $P(\%)$: Precision (in %); $R(\%)$: Recall (in %); $F1$: F1 Score (in %).

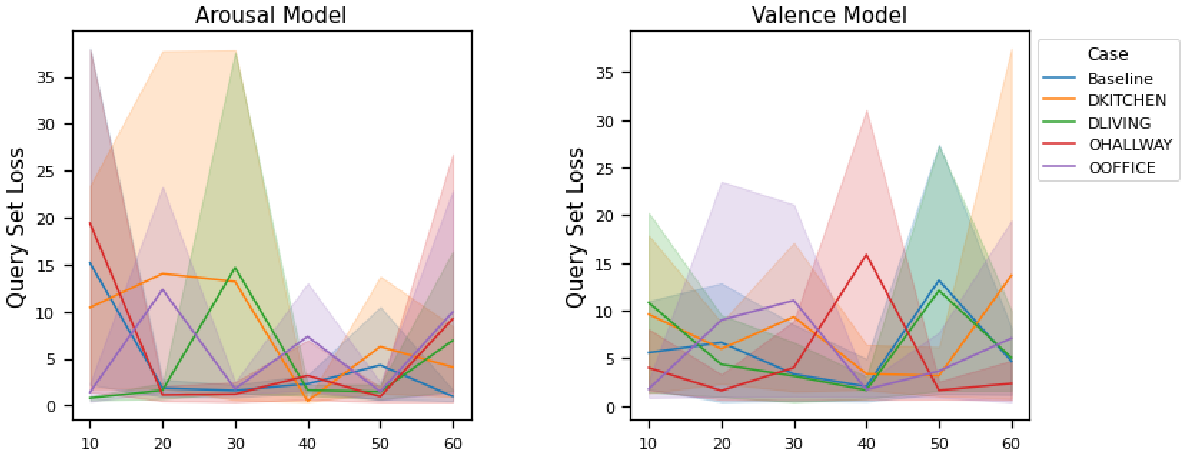


Figure 6.16: RECOLA : Summary of Query Set Contrastive Loss (L_q) for different cases

accuracy for this dimension from the $Baseline$. For instance, from the plots shown, the binary accuracy of the arousal model with the noise of type $OHALLWAY$ is consistently better than the $Baseline$. For higher values of K, the accuracies of other noise types also improve over $Baseline$ (as visible for the cases of $OOFFICE$, $DLIVING$). The accuracies drop at $K = 60$, as seen in the $Baseline$ case. The change of maximum mean accuracy from the $Baseline$ case with 53.81% is different across different noises. For instance, the accuracy is higher for $DLIVING$, $OHALLWAY$ and $OOFFICE$ cases at 55.75%, 56.17% 55.77% respectively.

The predictions on the valence dimension, have similar chaotic behaviour as observed with the arousal dimension. Comparing the cases of different noises with the $Baseline$, there is no particular uniformity. For the case of $OHALLWAY$, the overall mean accuracy drops, while for $DKITCHEN$, it rises with the increase in K. For the remaining cases, $DLIVING$ and $OOFFICE$, the variation of accuracy with K is inconclusive. Further, this is reflected across participants by the visible confidence intervals about the mean. Compared to the $Baseline$ accuracy of 53.87%, the accuracy increased for the case of $OHALLWAY$ and $OOFFICE$ with accuracies 54.53% and 55.31%.

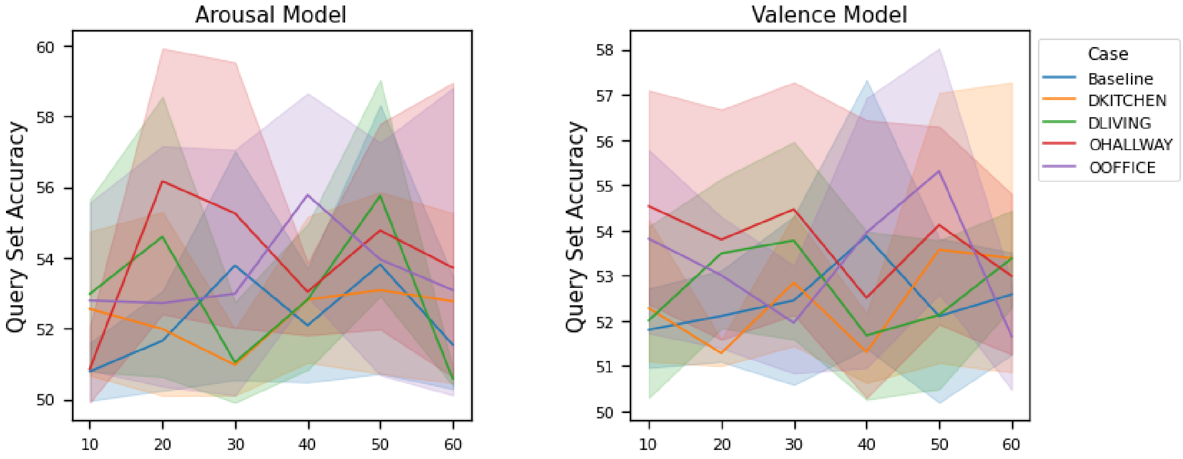


Figure 6.17: RECOLA : Summary of Query Set Binary Accuracy (A_q) for different cases

The results described above, while inconclusive of the trends of accuracy with K agree on the overall low value of accuracy throughout the different cases. For both the emotion dimensions, we see the accuracy span between 50 - 58% at most. This is a moderate performance for the participants. The increased accuracy with imputed noise is rather unusual and suggests that the added noise dramatically affects the audio samples. Alternatively, when talking about embeddings, the artefacts of noise are accentuated with the audio samples rather than being averaged out.

C. F1 Score Fig. 6.18 shows the f1 scores of the different cases compared to the $Baseline$ for the arousal and valence dimensions. The analysis of the arousal dimension predictions shows two sets of behaviours. The $Baseline$ case shows a trend of increase and subsequent decrease in f1 scores with increasing K, which are followed closely by the cases with the noise $OHALLWAY$, $OOFFICE$. This indicates the similarities in prediction patterns for the different cases across the labels. Alternatively, the imputed noises do not skew the embedding space thereby generating predictions consistent with the $Baseline$ case. The other two cases of noise $DKITCHEN$ and $DLIVING$, deviate from these cases. The f1 scores for these cases drop and rise again with increasing K.

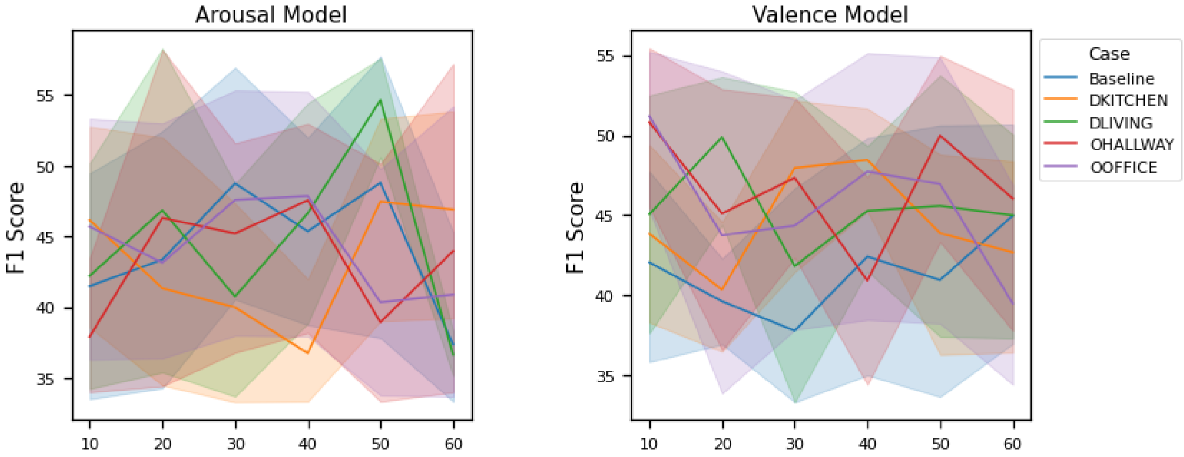


Figure 6.18: RECOLA : Summary of Query Set F1 Score ($F1$) for different cases

The f1 scores for the valence predictions are equally skewed. Compared to the $Baseline$, all the noise cases have higher f1 scores. This indicates that the introduction of noise in audio samples, in this case, wildly skew the resulting embedding space. The maximum mean f1 score for the $Baseline$ case occurs at $K = 60$, which differs from the other cases wildly. For instance, the maximum mean f1 scores with imputed noise of type $OHALLWAY$ and $OOFFICE$ occurs at $K = 5$.

Over and above the general observations discussed above, it is important to note that the overall f1 scores obtained with the RECOLA dataset are extremely poor. Most of the values occur well below 50%. As speculated earlier, these results could be stemming from ill-conditioned precision and f1 scores. A direct reason for this could be that the model simply does not learn to predict one of the labels at all for certain cases. This is specifically true for cases where the precision is also considerably lower than the recall.

Summary The discussion on the performance of the proposed architecture for the RECOLA dataset provides specific insights. Firstly, we again see similar patterns of diminishing residual loss with an increase in support size K for the arousal predictions. However, the same cannot be seen with the valence predictions. The plots of residual contrastive loss on the query set in Fig. 6.16 suggest a lot of chaotic behaviour in the models. The variance of the loss is very high across the participants. Next, when we compare this with the binary accuracy plots in Fig. 6.17, again, we see no specific trend in the distribution of mean accuracy across the various values of K. When we look at the results at a larger scale, the accuracy improves minimally over some values of K and then again drops at very high K. The change in binary accuracy observed for both the emotion dimensions is gradual and minimal. However, we also see the spread of the confidence interval across the 4 participants close to +/- 3-4%. This spread begs the question of analysis of participants for identifying the cause of the poor performance. Finally, with the f1 scores shown in Fig. 6.18, the differential performance observations are greatly enhanced across the noise types. Visually the variation in f1 scores is around +-15-20% for both the arousal and valence predictions. Since the values are derived from weighted scores across labels, it is suspected that the models fail on one of the labels significantly more than the other. Here, the labels in question are that of similarity (1) and dissimilarity (0). This punctuates the mean performances observed so far quite oddly, as it indicates that individual participants have greatly different responses to the proposed architecture.

A clear difference with this dataset from K-EmoCon is the absence of any major trends in both the binary accuracy and f1 scores. Not only is this attributed in the $Baseline$, but is also evident with the cases involving different noises. While some drop in performance is expected with the introduction of noise, the absence of any trends with the increasing value of K across all the cases as comparably visible in Fig. 6.17 and Fig. 6.18 stipulates that the proposed architecture only learn moderately from the dataset and significant distinguishing characteristics of the individual samples are either not captured or incompatible with the architecture. To test this stipulation, it is important to look at individual participants for more insights. IN the next section, we briefly present the results of individual participants.

6.3.3. Results with Individual Participants

To dive deeper into the reason for the poor average performance of the proposed Multimodal Siamese Network with the RECOLA dataset, we look at individual participants for insights. In Table. 6.12, the results of the selected participants for the best performing values of K are shown. To present the results concisely, the results from all the shots are not shown in the table here. Appendix B lists the results of all the individual participants in-depth.

From Table. 6.12, a lot of the observations made so far in the discussion of the mean performance metrics can be reasoned. Since the table presents results only with the best performing values of K, we can compare the participants across this dimension as well. In the next sections, we focus on binary accuracy and precision, recall and f1 scores to identify the causes of the average performances.

A. Binary Accuracy A closer look at the query set binary accuracy values in Table. 6.12 shows that while the average performances of the models are moderate, the individual performances are slightly better. For the prediction of arousal dimension, participant dev_2 performs significantly better than the rest with $Baseline$ accuracy of 60.47%. The addition of noise disrupts this accuracy only mildly, however, it changes the K for which the model achieves accuracy, wildly. For instance, participant dev_2, the case of $DKITCHEN$ has maxima in accuracy at $K = 50$, similar to that of $Baseline$, however for $DLIVING$ and $OHALLWAY$, the value drops to $K = 20$. Similarly, the best prediction on valence dimension is achieved with participant dev_3, with $Baseline$ accuracy of 59.00%. Incidentally, the degradation in accuracy with imputed noise is only minimal. But, like the variation in K observed in arousal models, here too, the value of K fluctuates wildly for all the types of noises.

Table 6.12: RECOLA : Performance Metrics for Individual participants [3]

Case	**Arousal Model**						**Valence Model**					
	K	A_s	A_q	P	R	$F1$	K	A_s	A_q	P	R	$F1$
	Participant ID - dev_1											
Baseline	**60**	100.00	54.38	25.00	50.00	33.33	**50**	99.50	53.42	25.00	50.00	33.33
DKITCHEN	**20**	96.25	56.97	56.97	56.97	56.97	**30**	51.67	52.10	52.11	52.10	52.03
DLIVING	**10**	100.00	57.13	59.25	57.13	54.53	**50**	95.50	54.79	54.80	54.79	54.77
OHALLWAY	**50**	98.00	58.02	25.00	50.00	33.33	**60**	93.33	54.57	54.58	54.57	54.57
OOFFICE	**40**	78.75	57.30	57.32	57.30	57.28	**20**	97.50	54.08	25.00	50.00	33.33
	Participant ID - dev_2											
Baseline	**50**	99.50	60.47	60.77	60.47	60.20	**50**	84.00	54.22	54.71	54.22	53.00
DKITCHEN	**50**	93.00	57.00	59.13	57.00	54.34	**40**	91.87	52.63	52.73	52.63	52.19
DLIVING	**20**	97.50	60.49	61.06	60.49	59.98	**20**	100.00	56.29	55.63	55.24	54.47
OHALLWAY	**20**	100.00	61.57	61.61	61.57	61.53	**10**	95.00	58.21	59.80	58.21	56.43
OOFFICE	**60**	88.75	61.66	62.64	61.66	60.90	**40**	90.62	57.69	58.88	57.69	56.22
	Participant ID - dev_3											
Baseline	**30**	95.83	55.19	55.28	55.19	55.02	**40**	98.75	59.00	62.38	51.11	36.69
DKITCHEN	**60**	59.17	52.82	52.83	52.82	52.77	**60**	99.58	59.13	63.77	52.68	40.76
DLIVING	**40**	86.87	55.27	55.49	55.27	54.84	**30**	67.50	57.26	58.95	57.26	55.13
OHALLWAY	**20**	95.00	54.99	55.06	54.99	54.84	**30**	92.50	58.74	65.35	58.74	53.76
OOFFICE	**40**	78.12	52.53	53.48	52.53	49.02	**50**	100.00	58.55	25.00	50.00	33.33
	Participant ID - dev_4											
Baseline	**20**	81.25	52.96	53.01	52.96	52.74	**60**	86.25	53.76	54.58	53.76	51.59
DKITCHEN	**30**	96.67	52.52	52.63	52.52	51.99	**30**	83.33	53.40	53.73	53.40	52.33
DLIVING	**10**	100.00	53.22	25.00	50.00	33.33	**60**	72.50	53.23	54.66	53.23	49.34
OHALLWAY	**40**	94.38	53.69	53.85	53.69	53.21	**10**	100.00	53.77	55.49	53.77	49.84
OOFFICE	**30**	94.17	53.46	53.90	53.46	52.10	**20**	93.75	54.51	54.58	54.51	54.36

This observation also helps in marking the reason for which the mean performances shown in Fig. 6.17 are lower across all K. This is because the value of K for which a participant model achieves maxima varies with the different imputed noises. This suggests that the impact of noise on a single sample is more pronounced and results in a larger variation in the embedding space compared to a sample from the $Baseline$ case. This results in a change in support set size K for the same or altered achievable accuracy for the given participant.

B. Precision, Recall and F1 Score Similar to the observations made in the case of binary accuracy, Table. 6.12 shows evidence of better precision, recall and f1 scores for some of the participants. The best f1 scores for a participant has varying values of K. For instance, in case

[3] Abbreviations :- A_s : Support Set Accuracy; A_q : Query Set Accuracy; $P(\%)$: Precision (in %); $R(\%)$: Recall (in %); $F1$: F1 Score (in %).

of arousal model predictions, participant dev_2 has the best $Baseline$ precision, recall and f1 score $[60.77\%, 60.47\%, 60.20\%]$ at $K = 50$; and for participant dev_1, the best $Baseline$ precision, recall and f1 score $[25\%, 50\%, 33.33\%]$ occurs at $K = 60$. Interestingly, the best performing f1 score for dev_1 is less than 50%. This resounds the case of ill-conditioned precision and f1 score. Similarly, we can see that for several cases across participants, some of the values of K resulted in ill-conditioned precision and f1 scores.

This discussion shows that while accurate discrimination of the similarity and dissimilarity classes is possible with the proposed architecture, it may be deteriorated by the presence of noise to the point where a class is not learned at all. Alternatively, the ill-conditioning may result from the inability of the proposed model to model the sample to embedding space, correctly.

6.4. Comparison with State of the Art

In the previous sections, we discussed in depth the results obtained with the two datasets representing the two test cases of our research question. Here, we compare these results with the state-of-the-art models. For the K-EmoCon dataset, the comparison is shown in Table. 6.13. It can be seen that on this dataset, the proposed model performs better than the existing works while relying on considerably less amount of support set. This value corresponds to the optimal performance for $K = 20$. Similarly, the comparison for RECOLA. In this case, we see that the performance of the proposed model fares moderately against the state-of-the-art models. However, it can be argued that against the fraction of amount of data used in the support set, the performance compares to achieve better valence prediction than some of the state-of-the-art models.

Table 6.13: Comparison with State-of-the-art Methods

Dataset	Reference	Modalities	Classes	$D_s : D_q$	Arousal	Valence
K-EmoCon	J. Quan et al.[80]	Audio	High, Low	70:30	53.02%	54.80%
	P. Gupta et al.[36]	Physio	5 classes	90:10	56.30%	31.03%
	Proposed Model	EDA, BVP, Audio	High, Low	**25:75**	**63.97**%	**66.91**%
RECOLA	M. Neumann et al.[65]	Audio	High, Low	80:20	60.77%	52.30%
	X. Cai et al.[17]	ECG, EDA, BVP, HR	High, Low	84:16	64.20%	50.00%
	J. Wu et al.[106]	Audio	High, Low	50:50	58.80%	34.80%
	Proposed Model	EDA, BVP, Audio	High, Low	**25:75**	**53.81**%	**53.87**%

Part V - Discussion & Future Work

7

Discussion and Future Work

In this final chapter, we present a discussion of the results of the experiments performed in this thesis. The goal of the proposed model is to answer the research questions presented in Section 1.3. We discuss the results of the experiments keeping these questions in mind in the Discussion section 7.1. Next, we present some of the limitations of the current work and motivate the possibilities of further exploration for production and deployment in Future Work 7.2. Finally, the thesis is concluded with the Conclusion section 7.3.

7.1. Discussion

In this section, we discuss the results obtained from our experiments with the two datasets with the proposed Multimodal Siamese Network. The results are summarized to answer the research questions posed in Section 1.3. The problem setting posed in Section 1.2 has several challenges. While several attempts have been made to solve similar problems, the approach presented in this thesis focuses on tackling problems associated with actual use-case settings and therefore differs largely from many of the previous works explored in this domain. We attempt to analyze the problem in a two-fold setting - with two datasets representing the behaviour of two types of conversation contexts. The K-EmoCon dataset represents the case where an IVA passively listens to the conversations, and the RECOLA dataset represents the case where an individual is in active conversation with the IVA. The characteristic settings in which the data is collected for each of these datasets is similar to the proposed settings. This setup of the problem statement is unique compared to existing models. Further, a very small amount of data is expected to be available for learning. Therefore, the proposed model in the thesis relies on few-shot learning for the basis of classification, with a specific focus on Siamese Networks.

We visit the various research questions with their respective objectives and analyse the results of the observations made across the two datasets in this section. An analysis of the proposed architecture is presented first while discussing the effect of certain choices on performance. Next, we analyse the optimal value of support set size for the two datasets. The analysis of optimality of sample duration is discussed next. This also provides other valuable insights regarding the inherent characteristics of the datasets. Finally, we analyse how the noise plays its effect on the proposed model and reason the robustness of the model against four different noise types.

7.1.1. Choice of Architecture

For analysing the performance of the emotion recognition pipeline on any dataset, it is important to consider the architectural characteristics of the pipeline. One of the major goals of this thesis is the fusion of modalities to augment system performance in realistic settings. We aim to solve the problem of emotion recognition with audio from everyday household conversations of individuals.

The *first research sub-question* deals with effective integration of physiological signals as additional modality. Our proposed model uses Siamese networks with multimodal embeddings for this task. This architecture fundamentally eliminates the need for a large amount of data for training. For the Siamese Networks, this problem becomes that of similarity and dissimilarity of samples. Specifically, the architectures learn to discriminate between classes. Further, multimodality is achieved with the concept of emotion embeddings. We create physiological embeddings parallel to the audio embeddings to perform the task classification. Signals obtained from wearable devices - BVP and EDA - are sources of physiological embeddings, while two separate audio embeddings are created out of a feature set of eGeMAPS and Mel-frequency spectrograms. These design choices are utilised to learn as much contextual information about emotions.

The construction of the four embeddings used to create a branch of Siamese Networks is equally important. To help in distilling discriminating features from each of the signals effectively, the embeddings focus on attributes of these signals. For instance, the temporal features of EDA and BVP signals within a sample are embedded using GRU embeddings. As already discussed, these embeddings are suited for use in multimodal context as these can be used to map the input physiological signal to a common space while being computationally more efficient than LSTM. However, it may be noted that LSTMs are useful in remembering longer signal sequences. This factor is important to note when dealing with signals of different sampling rates. If we analyze the results of the two datasets from this perspective, we may find that contributions of the physiological embeddings may have been hindered by the use of GRUs in the case of the RECOLA dataset. The K-EmoCon dataset has physiological signals of a lower sampling rate - with BVP signal at 64 Hz and EDA signal at 4Hz. These result in shorter sequences of signals per unit duration of length. On the other hand, RECOLA consists of signals at 250 Hz. This means, for a given duration, a signal sample from RECOLA is a considerably larger sequence than that of K-EmoCon. Consequently, an LSTM would appear to be a better discriminator specifically for cases where physiological signals are sampled at a large rate.

Aside from the physiological embeddings, the audio embeddings are the main modality for emotion classification. Here, too, the choice of architectures is dependent on the feature target space. The efficiency of mel-spectrograms in capturing emotional information from human audio is utilized with the help of CNN embeddings. The architecture relies on previous work and mimics the spectrogram embedding architectures similar to VGG-16. It may be noted that the simplified architecture proposed in the current work follows from the constraints established in the research questions - where the goal of the proposed solution is to be deployed on IVAs. This imposes a bottleneck on the complexity of the overall architecture. A desirable model achieves a good classification accuracy with limited computation power and prediction time.

Finally, it is worthwhile to note that this architecture allows the use of signals of any sampling rate. Across the two datasets, the sampling rates of the physiological signals vary widely, which is well accommodated by the architecture. For this purpose, the algorithm uses dynamic allocation of input size for embeddings.

7.1.2. Optimality of Number of Samples

The *second research sub-question* aims at identifying the optimal number of shots K for the sample set for the adequate performance of the proposed model against the state of art. The observations for this research question differ wildly for the two datasets.

For the K-EmoCon dataset, the mean $Baseline$ performances indicate that the proposed model achieves optimal prediction accuracy for $K = 25$ for the arousal and valence dimensions respectively. Comparing this mean $Baseline$ with individual participants, it can be observed that the results are largely consistent with this observation. Three out of four participants attain maxima in accuracies at $K = 25$. While statistically, this is not a significant analysis, we analyze the qualitative homogeneity of the results across individual participants. The agreement of optimality of $K = 25$, suggests the existence of optima across other participants of this dataset.

For the RECOLA dataset, the results of mean $Baseline$ performances are somewhat inconclusive. The maxima for this case is at $K = 50$ & 40 respectively for arousal and valence dimensions respectively. However, this is only slightly better than the performances at other values of K. More importantly, it is difficult to differentiate optimal performances from the mean values. When we take a look at the individual participants, the results are again quite skewed. For two of the participants the maxima for valence prediction performance occurs with $K = 50$, while with the other participants, it is lower with $K = 30$ & 40. The arousal prediction performance is even more confusing, with the maxima spread over multiple values of K. Therefore, optimality is wildly dependent on individual participants.

The observations made above also reveal the subjective differences presented by the type of audio in question. As described in Chapter 4, the K-EmoCon dataset involves continuous speech segments in a debate setting, while the RECOLA dataset contains spontaneous interactions between individuals. The poor performance on RECOLA reveals that the proposed model is unable to capture the emotional information from such spontaneous interactions and is perhaps better suited to decipher continuous speech.

7.1.3. Effect of Segment Length

The *third research sub-question* aims at comparing the effect of sample length on performances of the proposed model. This analysis requires a comparison of the performances of the model across the two datasets. As discussed earlier, the sample duration for the K-EmoCon dataset is 1s and for the RECOLA dataset, it is 0.4s. It is important to note that the experiments described in this thesis are limited and do not provide an objective measure of comparison of sample length. This is because the characteristics of the two datasets vary wildly. The differences in the sampling rate of physiological signals between the two datasets themselves render this comparison futile. However, we compare the results qualitatively and comment only on the probable impact of the contributions of audio embeddings towards the performance. To objectively compare the results, experiments need to be performed within a single dataset.

We compare the two datasets by matching the support set size of the support set D_s. In other words, the amount of training data is the same for a given instance of comparison despite having different sample duration. This is represented by the following values of support set size (per class) $K_{KEmoCon} = 20$ and $K_{RECOLA} = 50$. Both of these cases result in the total support set duration of 40s. The mean performances of the respective datasets on their selected participants are shown in Table. 7.1.

From Table. 7.1, a notable difference in the mean performances of the two datasets is

Table 7.1: Qualitative comparison of mean $Baseline$ performance across datasets

Dataset	**Duration**		Arousal Model					Valence Model				
		K	A_s	A_q	P	R	$F1$	A_s	A_q	P	R	$F1$
K-EmoCon	50s	20	96.56	63.97	62.28	56.97	52.80	99.06	60.88	60.41	60.11	59.58
RECOLA	50s	50	89.12	53.81	47.82	53.81	48.78	76.00	52.10	45.31	51.24	40.95

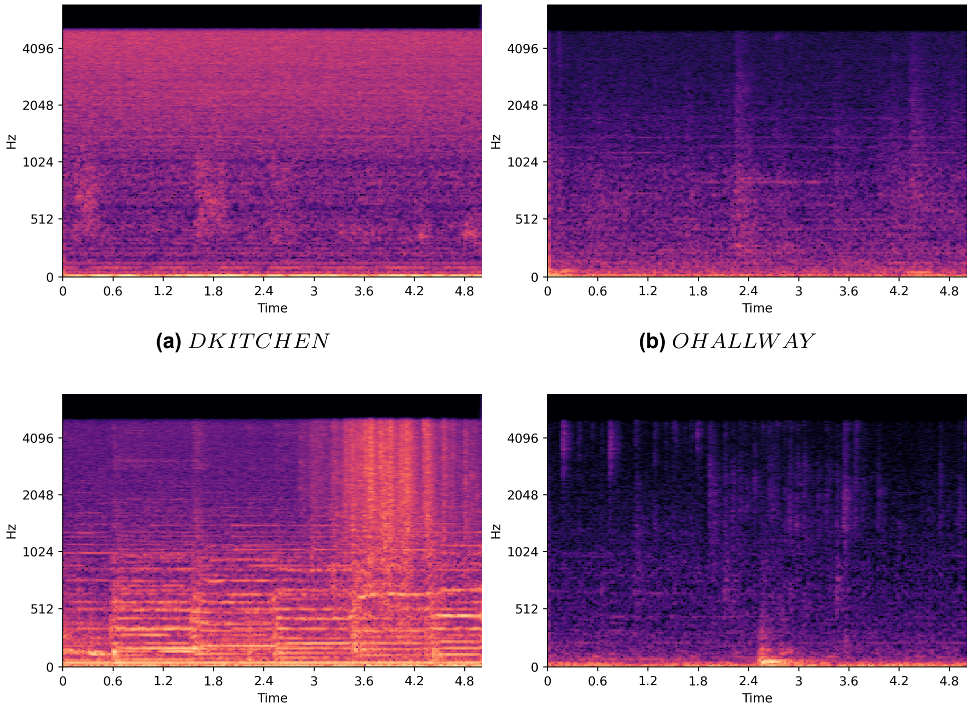
visible. Incidentally, even the support set accuracies achieved by the two datasets show that K-EmoCon performs better than RECOLA. Qualitatively, higher performances are obtained with K-EmoCon as compared to RECOLA for the same amount of support set size of the 50s. Remarkably, K-EmoCon performs better than RECOLA, even though the physiological signals of RECOLA have a higher sampling rate, implying larger data points to learn from. This comparison also throws light on the comparative contributions of the different embeddings on the resultant composite embedding of the Multimodal Siamese network. It may suggest that audio embeddings are relatively dominant contributors towards learning discriminative features than physiological embeddings. As previously stated, this could also be stemming from the architectural shortcomings of the proposed multimodal network. It is also important to note that for experiments with K-EmoCon, self-annotated labels are used as opposed to aggregate external annotations which are used for the experiments with RECOLA. Therefore, despite the observable differences in support set accuracy, it is not completely suggestive that unit sample duration affects the performance of the emotion recognition algorithm. This hypothesis requires substantial experimentation with several control conditions.

7.1.4. Robustness against Noise

Finally, we examine the *final research question* which aims at the analysis of the model robustness against noise. Several architectural, as well as feature choices, may have contributed to model robustness. When examining the effect of noise on the embeddings, the sources of disruption of the $Baseline$ embedding space are mel-frequency spectrogram embeddings as well as the eGeMAPS embeddings.

The eGeMAPS embeddings are selectively affected by the impact of most of the noise artefacts. The extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS), which contribute to the eGeMAPS embedding, are designed to capture dynamic characteristics of audio samples and selectively distil parameters for voiced and unvoiced regions. Some of the temporal features such as mean length and standard deviation of voiced and unvoiced regions, together with some of the cepstral features such as *Mel-Frequency Cepstral Coefficients (MFCC) 1-4*, *Spectral flux* and *Formant bandwidths*, may have contributed to heightened differentiation of speech in presence of noise. This is because most of the noise types are non-verbal in nature. Therefore, the contributing effects of these noises are selective on the parameters of eGeMAPS.

The mel-frequency spectrogram embeddings typically contribute to the chaotic behaviour of the proposed model in presence of noise. This is because of the mel-scale which mirror human perception of audio. Therefore, the effects of noise are distinctly unique on mel-frequency spectrograms. Fig. 7.1 shows the Mel-Frequency spectrograms of the four types of noise used in the experiments for a duration of 5s. It is visible that each noise has specific artefacts which contribute directly to the spectrogram embeddings. However, when we analyse each of these spectrograms carefully, we can see some temporal as well as spectral statistics of the noise. Typically, the $DKITCHEN$ type of noise appears to be spread evenly across higher frequen-



(a) $DKITCHEN$ **(b)** $OHALLWAY$

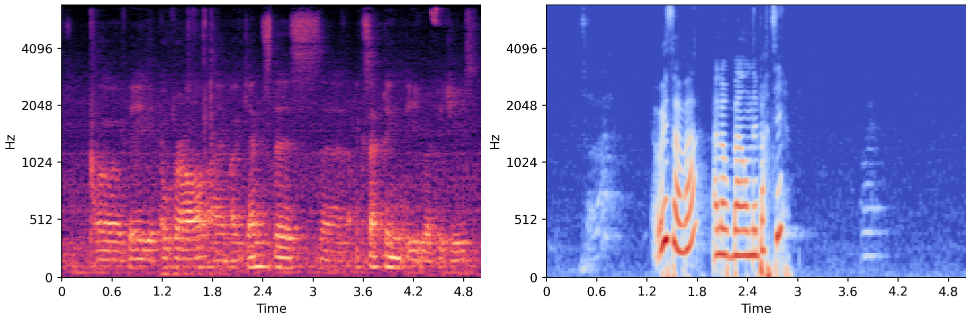
(c) $DLIVING$ **(d)** $OOFFICE$

Figure 7.1: Characteristic Mel-Frequency Spectrograms of the noises

cies as compared to lower frequencies. In contrast, the contributions of $OOFFICE$ is mainly in the lower frequencies, although the artefacts are not temporally constant. The $OHALLWAY$ noise has a sparse distribution of noise across the lower frequencies with some unique non-temporal instances. Finally, the most coloured of all noises is $DLIVING$. This distribution of noise is observably stationary as can be seen from the repeated patterns - owing to the presence of musical tones.

The analysis of these spectrograms immediately provides some insight into the behaviour of the spectrogram embedding in presence of noise. In addition to the robustness of the selected audio features of the embeddings, it may also be pointed out that the architecture of Siamese networks presents an inherent quality of robustness. Particularly, the principle of pair comparison (for similarity and dissimilarity), partly contributes to the elimination of effects of noise, since the objective of the twin networks is neither learning specific artefacts in spectrograms, nor patterns in the eGeMAPS parameter set. It is rather the contrast in the pairs, which contributes to the learning of the model. For a noise, such as $DKITCHEN$ and $OHALLWAY$ which exhibits properties of statistically stationarity, the contributions for the label pairs of dissimilarity are effectively cancelled across the two arms of the Siamese Network.

Further, if we compare the effects of noise on the two datasets, it is readily visible, no why the two datasets behave differently with imputed noise. Fig. 7.2 shows a sample of mel-frequency spectrogram from the two datasets for a duration of 5s. These spectrograms show dramatically different characteristics. Participant 15 from the K-EmoCon dataset exhibits a continuous audible speech signature while for participant dev_1, the speech is short and spontaneous. The verbal speech presence is present throughout the 5 seconds of duration for the former, while it is only partly present for the latter. It indicates that the per-frame contribution



(a) Participant 15 (with mel-banks = 512) **(b)** Participant dev_1(with mel-banks = 128)

Figure 7.2: Comparison of Mel-Frequency Spectrograms of datasets

of the former is more than the latter. This behaviour may also reflect on why the average performance of the K-EmoCon dataset is higher than that of RECOLA, in addition to the reasons suggested earlier. This is a fundamental deduction on the performance of the proposed model. It implies that the nature of interaction in a given audio signal may greatly affect the performance of the model.

7.2. Future Work

This thesis explored a novel methodology of Multimodal Siamese Networks for emotion recognition. While a lot of prior research work has focused on improving emotion recognition models with numerous datasets, this work comprises one of the first comprehensive experiments examining the few-shot emotion recognition problem while analysing the optimality of support set size as well as the effects of noise. This work attempts to investigate few-shot learning principles within realistic settings. The research contributions of this work reflect the ability of the proposed architecture to perform as good as the state-of-the-art architectures while only using a fraction of data for training. Further, to the best of our knowledge, this also marks the first study to compare the performance of a model against two different settings of data - one with continuous speech and the other with spontaneous speech. While the results obtained with the proposed model compared to be better than state-of-the-art for at least one of the datasets, it may still improve on the performance for the other dataset. In this section, we reflect on the gaps in the current work and present possibilities of future research work.

Analysis on Extended Dataset The current work presents the results of experiments conducted on four participants from each dataset. The K-EmoCon dataset presents a bottleneck for testing all the participants as we do not have the necessary number of samples for each of the classes. Alternatively, we restrict the number of participants for RECOLA to remain consistent with the participant count of the K-EmoCon dataset. This makes the results of the current work statistically skewed for the selected participants. The mean performances presented here, therefore do not reflect the behaviour of the complete dataset. The performance of the model can only be considered consistent on individual participants. Therefore, additional tests on the performance of the existing model with other datasets which contain sufficient samples of data may help in comparing the proposed architecture concretely against the state-of-the-art. Several available datasets have already been suggested in Chapter 4.

Ablation of the Physiological Embeddings It has been postulated earlier that the relative contributions of the audio embeddings are more than that of the physiological embeddings. This hypothesis could benefit from an ablation study of the physiological embeddings. In the current work, the concatenation of the embeddings from the different signals and features produces the final contrastive embeddings from the two arms of the network. Since the embeddings are independent of each other, an ablation of physiological embeddings could point out the sources of the contribution of these embeddings as well as the degree to which they affect the performance of the architecture. This study also provides useful insights into the production viability of the integration of physiological signals from wearable devices with IVAs.

Analysis of Effect of sample duration While the analysis of the effect of sample duration on the performance of the arousal and valence prediction is presented qualitatively here, it requires in-depth experimentation to concretely comment on the effects of sample duration. This work requires several control variables to be maintained while experimenting on a single dataset. The chosen dataset requires annotations at different scales, or perhaps a scheme to summarize annotations for a given sample duration.

Novel Architectures The proposed model relies on contrastive loss from pairs of labelled data to predict the class similarity or dissimilarity. This approach of pair comparison, do not account for the degree of exhibited emotion in the sample. When dealing with more than two classes of emotions on a dimensional emotion model, the degree of emotion exhibited may be used to differentiate similar samples more readily than others. This can be performed using other types of metric learning methods. For instance, a variant of the Siamese network - the Triplet network, uses the same architecture, with three embeddings generated from triplets of samples - one representing an anchor, another representing a positive sample and the third representing a negative sample. This architecture may be used to augment the capabilities of the current architecture. Additionally, the proposed architecture may be optimized by the ranking of pairs using a statistical score of the samples. Samples exhibiting stronger similarity may be ranked to obtain better pairs for training. Similarly, prototypical networks may be used to aggregate embeddings with similar cosine distances, thereby performing a clustering scheme for classification.

Privacy While affective computing in the wild opens doors for numerous applications in monitoring emotions and well being individual privacy is an important concern. In the current study, the data from the users are openly available for scientific research. However, for algorithms and systems in production, the continuous capture of audio and physiological data by IVAs and smartwatches poses a threat of misuse and malpractice. Therefore, techniques of privacy preservation should be accounted for in the commercialization of such algorithms. To provide sufficient data for model training while keeping the user identity anonymous, several techniques are possible. Firstly, audio anonymization using identity-removing transformations such as dimensionality reduction, the multiplicative perturbation may be applied intermediately to the inputs. Alternatively, several metric learning methods may also apply homomorphic encryption to the datasets.

7.3. Conclusion

The goal of this thesis has been the exploration of few-shot emotion recognition algorithms for IVAs in a domestic setting. While several of the works in the literature have attempted to solve this problem, many of these rely on extensive amounts of training data with high sampling rates. Further, many approaches consider laboratory settings of data collection which are obtrusive and may affect the emotions exhibited by the individual. Our approach departs from these methods and can operate with audio samples which reflect actual interactions with IVAs.

The proposed model achieves the research objectives laid down in this thesis, with specific architectural design and training strategies with the help of an adaptive Siamese Network architecture with multimodal embeddings. The architecture readily accepts raw physiological signals to transform them into physiological feature embeddings. Further, it uses spectrograms and eGeMAPS feature sets to capture audio statistics for the respective embeddings. These multimodal embeddings are concatenated to generate a composite embedding used for the comparison of sample pairs. Finally, the contrastive loss is used as the optimization objective for the algorithm. This architecture answers the first research question on the integration of physiological networks in the few-shot domain. We test the proposed model on two datasets representing two specific cases of IVA interactions. The K-EmoCon dataset represents a case of an IVA listening passively to a user, while the RECOLA dataset represents a case where the user interacts actively with IVAs. With each dataset, we test the 6 different support set-sizes. The support set represents the training set available for the model. The second research question can be answered with these tests, with $K = 25$ for K-EmoCon. Although the value of K for RECOLA is inconclusive, moderately high performance is obtained for $K = 40$ across participants. For each of the datasets, observations show moderate to high-performance scores. The performances on the K-EmoCon dataset are 63.97% and 66.91% for arousal and valence dimensions. For RECOLA, the performances are 53.81% and 52.87% for the arousal and valence dimensions. This performance is comparable to the state of the art models. The third research question is answered quantitatively by comparing the performances of the two datasets, represented by two different sample duration. While this discussion is not concretely built on quantitative analysis, an overall comparison across participants, definitely suggests the superiority of higher sample duration on prediction performance for the two emotion dimensions. The fourth research question is examined with the experiments of the datasets with the imputation of various types of noise. It is observed that the models sustainingly perform better for the K-EmoCon dataset, however, there is some degradation in accuracy for RECOLA. This is analysed further to find the source of the performance bottleneck.

The proposed work provides insights into the effects of sample duration, noise as well training set size on the proposed few-shot emotion recognition algorithm. While it lacks evaluation across all the participants of the dataset, it provides a potent direction to study these parameters in conjunction with the emotion recognition problem. These parameters are fundamentally important to determine ways to optimize the existing emotion recognition pipelines. Further, comparisons from the two datasets also provide insights on the kind of interaction on which better emotion recognition is possible. We believe this work contributes towards the design of emotionally intelligent IVAs.

References

[1] Mehmet Berkehan Akçay and Kaya Oğuz. “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers”. In: *Speech Commun.* 116.December 2019 (2020), pp. 56–76. ISSN: 01676393. DOI: 10.1016/j.specom.2019.12.001.

[2] Mouhannad Ali et al. “Emotion recognition involving physiological and speech signals: A comprehensive review”. In: *Stud. Syst. Decis. Control*. Vol. 109. 2018, pp. 287–302. ISBN: 9783319589961. DOI: 10.1007/978-3-319-58996-1_13.

[3] William Apple, Lynn A. Streeter, and Robert M. Krauss. “Effects of pitch and speech rate on personal attributions”. In: *J. Pers. Soc. Psychol.* 37.5 (1979), pp. 715–727. ISSN: 00223514. DOI: 10.1037/0022-3514.37.5.715.

[4] *Apple Watch Series 5: Sports & Fitness In-Depth Review | DC Rainmaker*. URL: https://www.dcrainmaker.com/2019/11/apple-watch-series-5-sports-fitness-in-depth-review.html (visited on 11/25/2021).

[5] Yasmine Arafa and Abe Mamdani. “Virtual Personal Service Assistants: Towards real-time characters with artificial hearts”. In: *Int. Conf. Intell. User Interfaces, Proc. IUI* (2000), pp. 9–12.

[6] Priya Arora and Theodora Chaspari. “Exploring Siamese neural network architectures for preserving speaker identity in speech emotion classification”. In: *Proc. 4th Work. Multimodal Anal. Enabling Artif. Agents Human-Machine Interact. MA3HMI 2018 - conjunction with ICMI 2018* (2018), pp. 15–18. DOI: 10.1145/3279972.3279980.

[7] Dare A. Baldwin and Louis J. Moses. “The Ontogeny of Social Information Gathering”. In: *Child Dev.* 67.5 (1996), pp. 1915–1939. ISSN: 00093920. DOI: 10.1111/j.1467-8624.1996.tb01835.x.

[8] Tanja Bänziger and Klaus R. Scherer. “The role of intonation in emotional expressions”. In: *Speech Commun.* 46.3-4 (2005), pp. 252–267. ISSN: 01676393. DOI: 10.1016/j.specom.2005.02.016.

[9] Anton Batliner et al. “Desperately Seeking Emotions or: Actors, Wizards, and Human Beings”. In: *ISCA Tutor.* September (2000), pp. 195–200. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.7104.

[10] Anton Batliner et al. “From emotion to interaction: Lessons from real human-machine-dialogues”. In: *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.* 3068 (2004), pp. 1–12. ISSN: 03029743. DOI: 10.1007/978-3-540-24842-2_1.

[11] Frank Bentley et al. “Understanding the Long-Term Use of Smart Speaker Assistants”. In: *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.* 2.3 (2018), pp. 1–24. ISSN: 2474-9567. DOI: 10.1145/3264901.

[12] Jonathan Boigne, Biman Liyanage, and Ted Östrem. “Recognizing More Emotions with Less Data Using Self-supervised Transfer Learning”. In: (2020). arXiv: 2011.05585. URL: http://arxiv.org/abs/2011.05585.

[13] Patricia J. Bota et al. “A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals”. In: *IEEE Access* 7 (2019), pp. 140990–141020. ISSN: 21693536. DOI: 10.1109/ACCESS.2019.2944001.

[14] Herve Bredin et al. “Pyannote.Audio: Neural Building Blocks for Speaker Diarization”. In: *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* 2020-May (2020), pp. 7124–7128. ISSN: 15206149. DOI: 10.1109/ICASSP40776.2020.9052974. arXiv: 1911.01255.

[15] *Buy Apple Watch Series 7 - Apple*. URL: https://www.apple.com/shop/buy-watch/apple-watch (visited on 11/25/2021).

[16] John T. Cacioppo, Louis G. Tassinary, and Gary G. Berntson. *Handbook of psychophysiology, fourth edition*. 4th ed. Cambridge: Cambridge University Press, 2016, pp. 1–716. ISBN: 9781107415782. DOI: 10.1017/9781107415782. URL: https://www.cambridge.org/core/books/handbook-of-psychophysiology/EACAC4007D68C77D20B912D18C78A370.

[17] Xiong Cai et al. “Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition”. In: *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* 2021-June (2021), pp. 5734–5738. ISSN: 15206149. DOI: 10.1109/ICASSP39728.2021.9413907. arXiv: 2010.13350.

[18] Marc Cavazza and C Emilio Vargas. “How Was Your Day ? A Companion ECA”. In: *Proc. Multiagent 9th Int. Syst. Conf. Auton. Agents (AAMAS 2010)* Aamas (2010), pp. 1629–1630.

[19] Wei Yu Chen et al. “A closer look at few-shot classification”. In: *7th Int. Conf. Learn. Represent. ICLR 2019* 2018 (2019), pp. 1–17. arXiv: 1904.04232.

[20] Farah Chenchah and Zied Lachiri. “Speech emotion recognition in noisy environment”. In: *2nd Int. Conf. Adv. Technol. Signal Image Process. ATSIP 2016* (2016), pp. 788–792. DOI: 10.1109/ATSIP.2016.7523189.

[21] Anca-Nicoleta Ciubotaru et al. “Revisiting Few-Shot Learning for Facial Expression Recognition”. In: (2019). arXiv: 1912.02751. URL: http://arxiv.org/abs/1912.02751.

[22] Phil Cohen et al. “On the future of personal assistants”. In: *Conf. Hum. Factors Comput. Syst. - Proc.* 07-12-May- (May 2016), pp. 1032–1037. DOI: 10.1145/2851581.2886425. URL: https://dl.acm.org/doi/10.1145/2851581.2886425.

[23] Giovanna Colombetti. “From affect programs to dynamical discrete emotions”. In: *Philos. Psychol.* 22.4 (2009), pp. 407–425. ISSN: 09515089. DOI: 10.1080/09515080903153600.

[24] *Compare the Privacy Practices of the Most Popular Smart Speakers with Virtual Assistants | Common Sense Education*. URL: https://www.commonsense.org/education/articles/compare-the-privacy-practices-of-the-most-popular-smart-speakers-with-virtual-assistants (visited on 11/25/2021).

[25] Sidney K. D’Mello and Jacqueline Kory. “A review and meta-analysis of multimodal affect detection systems”. In: *ACM Comput. Surv.* 47.3 (2015). ISSN: 15577341. DOI: 10.1145/2682899.

[26] Sidney DMello and Art Graesser. “AutoTutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back”. In: *ACM Trans. Interact. Intell. Syst.* 2.4 (2012), pp. 1–39. ISSN: 21606463. DOI: 10.1145/2395123.2395128.

[27] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. “Human emotion recognition: Review of sensors and methods”. In: *Sensors (Switzerland)* 20.3 (2020). ISSN: 14248220. DOI: 10.3390/s20030592.

[28] Maria Egger, Matthias Ley, and Sten Hanke. “Emotion Recognition from Physiological Signal Analysis: A Review”. In: *Electron. Notes Theor. Comput. Sci.* 343 (2019), pp. 35–55. ISSN: 15710661. DOI: 10.1016/j.entcs.2019.04.009. URL: https://doi.org/10.1016/j.entcs.2019.04.009.

[29] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. “Survey on speech emotion recognition: Features, classification schemes, and databases”. In: *Pattern Recognit.* 44.3 (2011), pp. 572–587. ISSN: 00313203. DOI: 10.1016/j.patcog.2010.09.020. URL: http://dx.doi.org/10.1016/j.patcog.2010.09.020.

[30] *Electrocardiogram Monitoring Cleared for Galaxy Watch Active2 by South Korea’s Ministry of Food and Drug Safety – Samsung Global Newsroom*. URL: https://news.samsung.com/global/electrocardiogram-monitoring-cleared-for-galaxy-watch-active2-by-south-koreas-ministry-of-food-and-drug-safety (visited on 11/25/2021).

[31] Empatica. *E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors*. 2020. URL: https://www.empatica.com/research/e4/ (visited on 11/27/2021).

[32] Florian Eyben and Björn Schuller. “openSMILE:)” in: *ACM SIGMultimedia Rec.* 6.4 (2015), pp. 4–13. ISSN: 1947-4598. DOI: 10.1145/2729095.2729097.

[33] Florian Eyben et al. “The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing”. In: *IEEE Trans. Affect. Comput.* 7.2 (2016), pp. 190–202. ISSN: 19493045. DOI: 10.1109/TAFFC.2015.2457417.

[34] Kexin Feng and Theodora Chaspari. “Few-shot Learning in Emotion Recognition of Spontaneous Speech Using a Siamese Neural Network with Adaptive Sample Pair Formation”. In: *IEEE Trans. Affect. Comput.* (2021), pp. 1–8. ISSN: 19493045. DOI: 10.1109/TAFFC.2021.3109485. arXiv: 2109.02915.

[35] Nico H. Frijda and Batja Mesquita. “The Analysis of Emotions”. In: *What Dev. Emot. Dev.* Springer, Boston, MA, 1998, pp. 273–295. DOI: 10.1007/978-1-4899-1939-7_11. URL: https://link-springer-com.tudelft.idm.oclc.org/chapter/10.1007/978-1-4899-1939-7%7B%5C_%7D11.

[36] Priyansh Gupta et al. “Emotion Recognition During Social Interactions Using Peripheral Physiological Signals”. In: *Lect. Notes Data Eng. Commun. Technol.* Vol. 75. Springer Singapore, 2022, pp. 99–112. ISBN: 9789811637285. DOI: 10.1007/978-981-16-3728-5_8.

[37] Raia Hadsell, Sumit Chopra, and Yann LeCun. “Dimensionality reduction by learning an invariant mapping”. In: *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 2 (2006), pp. 1735–1742. ISSN: 10636919. DOI: 10.1109/CVPR.2006.100.

[38] Jing Han et al. “EmoBed: Strengthening Monomodal Emotion Recognition via Training with Crossmodal Emotion Embeddings”. In: *IEEE Trans. Affect. Comput.* 12.3 (2021), pp. 553–564. ISSN: 19493045. DOI: 10.1109/TAFFC.2019.2928297. arXiv: 1907.10428.

[39] Shawn Hershey et al. “CNN architectures for large-scale audio classification”. In: *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* (2017), pp. 131–135. ISSN: 15206149. DOI: 10.1109/ICASSP.2017.7952132. arXiv: 1609.09430.

[40] Matthew B. Hoy. “Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants”. In: *Med. Ref. Serv. Q.* 37.1 (2018), pp. 81–88. ISSN: 15409597. DOI: 10.1080/02763869.2018.1404391.

[41] Yang Hu et al. “What Can Knowledge Bring to Machine Learning? – A Survey of Low-shot Learning for Structured Data”. In: 1.1 (2021). arXiv: 2106.06410. URL: http://arxiv.org/abs/2106.06410.

[42] Łukasz Juszkiewicz. “Improving noise robustness of speech emotion recognition system”. In: *Stud. Comput. Intell.* 511 (2014), pp. 223–232. ISSN: 1860949X. DOI: 10.1007/978-3-319-01571-2_27.

[43] M L Knapp and J A Hall. *Nonverbal communication in human interaction*. 1972, p. 512. ISBN: 0495568694. URL: http://books.google.com/books?id=gAmpPwAACAAJ%7B%5C&%7Ddq=isbn:0534625630.

[44] R. Benjamin Knapp, Jonghwa Kim, and Elisabeth André. “Physiological Signals and Their Use in Augmenting Emotion Recognition for Human–Machine Interaction”. In: *Cogn. Technol.* 9783642151835. 2011, pp. 133–159. ISBN: 9783642151842. DOI: 10.1007/978-3-642-15184-2_9. URL: http://link.springer.com/10.1007/978-3-642-15184-2%7B%5C_%7D9.

[45] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. “Siamese Neural Networks for One-Shot Image Recognition”. In: (2015). URL: http://www.cs.utoronto.ca/%7B~%7Dgkoch/files/msc-thesis.pdf.

[46] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. “Face Behavior a la carte: Expressions, Affect and Action Units in a Single Network”. In: (2019). arXiv: 1910.11111. URL: http://arxiv.org/abs/1910.11111.

[47] Andreas Komninos and Sofia Stamou. “HealthPal : An Intelligent Personal Medical Assistant for Supporting the Self-Monitoring of Healthcare in the Ageing Society”. In: *Proc. 4th Int. Work. Ubiquitous Comput. Pervasive Healthc. Appl.* (2006).

[48] Shashidhar G. Koolagudi and K. Sreenivasa Rao. “Emotion recognition from speech: A review”. In: *Int. J. Speech Technol.* 15.2 (2012), pp. 99–117. ISSN: 13812416. DOI: 10.1007/s10772-011-9125-1.

[49] Sylvia D. Kreibig. “Autonomic nervous system activity in emotion: A review”. In: *Biol. Psychol.* 84.3 (2010), pp. 394–421. ISSN: 03010511. DOI: 10.1016/j.biopsycho.2010.03.010.

[50] Soumya Kuruvayil and Suja Palaniswamy. “Emotion recognition from facial images with simultaneous occlusion, pose and illumination variations using meta-learning”. In: *J. King Saud Univ. - Comput. Inf. Sci.* (June 2021). ISSN: 22131248. DOI: 10.1016/j.jksuci.2021.06.012.

[51] Peter J. Lang, Margaret M. Bradley, and Bruce N. Cuthbert. “Emotion and Motivation: Measuring Affective Perception”. In: *J. Clin. Neurophysiol.* 15.5 (Sept. 1998), pp. 397–408. ISSN: 0736-0258. DOI: 10.1097/00004691-199809000-00004. URL: http://journals.lww.com/00004691-199809000-00004.

[52] Mihee Lee et al. “Fast and Effective Adaptation of Facial Action Unit Detection Deep Model”. In: (2019). arXiv: 1909.12158. URL: http://arxiv.org/abs/1909.12158.

[53] Robert W. Levenson. “Blood, Sweat, and Fears: The Autonomic Architecture of Emotion”. In: *Ann. N. Y. Acad. Sci.* 1000 (2003), pp. 348–366. ISSN: 00778923. DOI: 10.1196/annals.1280.016.

[54] Rosemarijn Looije, Mark A. Neerincx, and Fokie Cnossen. “Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors”. In: *Int. J. Hum. Comput. Stud.* 68.6 (2010), pp. 386–397. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2009.08.007. URL: http://dx.doi.org/10.1016/j.ijhcs.2009.08.007.

[55] P. Low. *Overview of the Autonomic Nervous System - Neurologic Disorders - MSD Manual Professional Edition*. 2020. URL: https://www.merckmanuals.com/professional/neurologic-disorders/autonomic-nervous-system/overview-of-the-autonomic-nervous-system%20https://www.msdmanuals.com/professional/neurologic-disorders/autonomic-nervous-system/overview-of-the-autonomic-nervous-system (visited on 11/25/2021).

[56] Ewa Luger and Abigail Sellen. “”Like having a really bad pa”: The gulf between user expectation and experience of conversational agents”. In: *Conf. Hum. Factors Comput. Syst. - Proc.* (2016), pp. 5286–5297. DOI: 10.1145/2858036.2858288.

[57] Alison Marczewski, Adriano Veloso, and Nivio Ziviani. “Learning transferable features for speech emotion recognition”. In: *Themat. Work. 2017 - Proc. Themat. Work. ACM Multimed. 2017, co-located with MM 2017* (2017), pp. 529–536. DOI: 10.1145/3126686.3126735.

[58] Brian McFee et al. “librosa: Audio and Music Signal Analysis in Python”. In: *Proc. 14th Python Sci. Conf.* Scipy (2015), pp. 18–24. DOI: 10.25080/majora-7b98e3ed-003.

[59] Donald McMillan et al. “Situating wearables: Smartwatch use in context”. In: *Conf. Hum. Factors Comput. Syst. - Proc.* 2017-May (2017), pp. 3582–3594. DOI: 10.1145/3025453.3025993.

[60] Albert Mehrabian. “Pleasure-Arousal-Dominance: A general framework for describing and measuring individual differences in temperament”. In: *Curr. Psychol.* 14.4 (1996), pp. 261–292. ISSN: 10461310. DOI: 10.1007/bf02686918. URL: https://link-springer-com.tudelft.idm.oclc.org/article/10.1007/BF02686918.

[61] Christine Meyer, Florian Eyben, and Thomas Zielke. “DEEP NEURAL NETWORKS FOR ACOUSTIC EMOTION RECOGNITION : RAISING THE BENCHMARKS Andr ´ Dept . of Mechanical and Process Engineering , D ¨ usseldorf University of Applied Sciences , Germany Dept . of Electrical Engineering , D ¨ usseldorf University of Appl”. In: *English* (2011), pp. 5688–5691. ISSN: 15206149.

[62] Juan Abdon Miranda-Correa et al. “AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups”. In: *IEEE Trans. Affect. Comput.* 12.2 (2021), pp. 479–493. ISSN: 19493045. DOI: 10.1109/TAFFC.2018.2884461. arXiv: 1702.02510.

[63] Mumtaz Begum Mustafa et al. “Speech emotion recognition research: an analysis of research focus”. In: *Int. J. Speech Technol.* 21.1 (2018), pp. 137–156. ISSN: 15728110. DOI: 10.1007/s10772-018-9493-x. URL: http://dx.doi.org/10.1007/s10772-018-9493-x.

[64] Anugunj Naman and Liliana Mancini. “Fixed-MAML for Few Shot Classification in Multilingual Speech Emotion Recognition”. In: (2021). arXiv: 2101.01356. URL: http://arxiv.org/abs/2101.01356.

[65] Michael Neumann and N. Goc Thang Vu. “CRoss-lingual and Multilingual Speech Emotion Recognition on English and French”. In: *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* 2018-April (2018), pp. 5769–5773. ISSN: 15206149. DOI: 10.1109/ICASSP.2018.8462162. arXiv: 1803.00357.

[66] Stavros Ntalampiras. “Speech emotion recognition via learning analogies”. In: *Pattern Recognit. Lett.* 144 (2021), pp. 21–26. ISSN: 01678655. DOI: 10.1016/j.patrec.2021.01.018.

[67] Alexey Ozerov and Cédric Fevotte. “Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation”. In: *IEEE Trans. Audio, Speech Lang. Process.* 18.3 (2010), pp. 550–563. ISSN: 15587916. DOI: 10.1109/TASL.2009.2031510.

[68] Meghna Pandharipande et al. “Robust front-end processing for emotion recognition in noisy speech”. In: *2018 11th Int. Symp. Chinese Spok. Lang. Process. ISCSLP 2018 - Proc.* (2018), pp. 324–328. DOI: 10.1109/ISCSLP.2018.8706699.

[69] Maja Pantic et al. “Multimodal emotion recognition from low-level cues”. In: *Cogn. Technol.* 9783642151835 (2011), pp. 115–132. ISSN: 16112482. DOI: 10.1007/978-3-642-15184-2_8.

[70] Cheul Young Park et al. “K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations”. In: *Sci. Data* 7.1 (2020). ISSN: 20524463. DOI: 10.1038/s41597-020-00630-y. arXiv: 2005.04120.

[71] Rosalind Picard. “Affective computing”. In: *Pattern Anal. Appl.* 1.1 (1998), pp. 71–73. ISSN: 1433-7541. DOI: 10.1007/bf01238028.

[72] Robert Plutchik. “The nature of emotions”. In: *Am. Sci.* 89.4 (Nov. 2001), pp. 344–350. ISSN: 00030996. URL: http://www.jstor.org/stable/27857503.

[73] Martin Porcheron et al. “Voice interfaces in everyday life”. In: *Conf. Hum. Factors Comput. Syst. - Proc.* 2018-April (2018), pp. 1–12. DOI: 10.1145/3173574.3174214.

[74] Soujanya Poria et al. “A review of affective computing: From unimodal analysis to multimodal fusion”. In: *Inf. Fusion* 37 (2017), pp. 98–125. ISSN: 15662535. DOI: 10.1016/j.inffus.2017.02.003. URL: https://doi.org/10.1016/j.inffus.2017.02.003.

[75] Soujanya Poria et al. “Merging SenticNet and WordNet-Affect emotion lists for sentiment analysis”. In: *Int. Conf. Signal Process. Proceedings, ICSP* 2 (2012), pp. 1251–1255. DOI: 10.1109/ICoSP.2012.6491803.

[76] Alisha Pradhan, Amanda Lazar, and Leah Findlater. “Use of intelligent voice assistants by older adults with low technology use”. In: *ACM Trans. Comput. Interact.* 27.4 (2020). ISSN: 15577325. DOI: 10.1145/3373759.

[77] *pydiarization · PyPI*. URL: https://pypi.org/project/pydiarization/ (visited on 11/30/2021).

[78] *pydub/API.markdown at master · jiaaro/pydub · GitHub*. URL: https://github.com/jiaaro/pydub/blob/master/API.markdown (visited on 12/13/2021).

[79] Jie Lin Qiu, Wei Liu, and Bao Liang Lu. “Multi-view emotion recognition using deep canonical correlation analysis”. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* 11305 LNCS (2018), pp. 221–231. ISSN: 16113349. DOI: 10.1007/978-3-030-04221-9_20. arXiv: 1908.05349.

[80] Jingyu Quan, Yoshihiro Miyake, and Takayuki Nozawa. “Incorporating interpersonal synchronization features for automatic emotion recognition from visual and audio data during communication”. In: *Sensors* 21.16 (2021). ISSN: 14248220. DOI: 10.3390/s21165317.

[81] Fabien Ringeval et al. “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions”. In: *2013 10th IEEE Int. Conf. Work. Autom. Face Gesture Recognition, FG 2013* i (2013), pp. 1–8. DOI: 10.1109/FG.2013.6553805.

[82] James A. Russell. “A circumplex model of affect”. In: *J. Pers. Soc. Psychol.* 39.6 (1980), pp. 1161–1178. ISSN: 00223514. DOI: 10.1037/h0077714.

[83] Kashfia Sailunaz et al. “Emotion detection from text and speech: a survey”. In: *Soc. Netw. Anal. Min.* 8.1 (2018), pp. 1–26. ISSN: 18695469. DOI: 10.1007/s13278-018-0505-2. URL: https://doi.org/10.1007/s13278-018-0505-2.

[84] *Samsung Galaxy Watch4 44mm Black | Samsung Netherlands*. URL: https://www.samsung.com/nl/watches/galaxy-watch/galaxy-watch4-black-bluetooth-sm-r870nzkaeub/ (visited on 11/25/2021).

[85] Klaus R. Scherer. “A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology”. In: *6th Int. Conf. Spok. Lang. Process. ICSLP 2000* (2000).

[86] Philip Schmidt et al. “Wearable-based affect recognition—a review”. In: *Sensors (Switzerland)* 19.19 (2019), pp. 1–42. ISSN: 14248220. DOI: 10.3390/s19194079.

[87] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. In: *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 07-12-June (2015), pp. 815–823. ISSN: 10636919. DOI: 10.1109/CVPR.2015.7298682. arXiv: 1503.03832.

[88] Nicu Sebe and Ira Cohen. “Multimodal approaches for emotion recognition: a survey”. In: *Electron. ...* 5670 (2005), pp. 56–67. URL: http://proceedings.spiedigitallibrary.org/proceeding.aspx?articleid=857894.

[89] Nusrat J. Shoumy et al. “Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals”. In: *J. Netw. Comput. Appl.* 149.February 2019 (2020), p. 102447. ISSN: 10958592. DOI: 10.1016/j.jnca.2019.102447. URL: https://doi.org/10.1016/j.jnca.2019.102447.

[90] Lin Shu et al. “A review of emotion recognition using physiological signals”. In: *Sensors (Switzerland)* 18.7 (2018). ISSN: 14248220. DOI: 10.3390/s18072074.

[91] Leslie N. Smith. “Cyclical learning rates for training neural networks”. In: *Proc. - 2017 IEEE Winter Conf. Appl. Comput. Vision, WACV 2017* April (2017), pp. 464–472. DOI: 10.1109/WACV.2017.58. arXiv: 1506.01186.

[92] Mohammad Soleymani et al. “A multimodal database for affect recognition and implicit tagging”. In: *IEEE Trans. Affect. Comput.* 3.1 (2012), pp. 42–55. ISSN: 19493045. DOI: 10.1109/T-AFFC.2011.25.

[93] Yaxin Sun, Guihua Wen, and Jiabing Wang. “Weighted spectral features based on local Hu moments for speech emotion recognition”. In: *Biomed. Signal Process. Control* 18 (2015), pp. 80–90. ISSN: 17468108. DOI: 10.1016/j.bspc.2014.10.008. URL: http://dx.doi.org/10.1016/j.bspc.2014.10.008.

[94] Monorama Swain, Aurobinda Routray, and P. Kabisatpathy. “Databases, features and classifiers for speech emotion recognition: a review”. In: *Int. J. Speech Technol.* 21.1 (2018), pp. 93–120. ISSN: 15728110. DOI: 10.1007/s10772-018-9491-z. URL: http://dx.doi.org/10.1007/s10772-018-9491-z.

[95] Ashish Tawari and Mohan Trivedi. “Speech emotion analysis in noisy real-world environment”. In: *Proc. - Int. Conf. Pattern Recognit.* 2010, pp. 4605–4608. ISBN: 9780769541099. DOI: 10.1109/ICPR.2010.1132.

[96] George Terzopoulos and Maya Satratzemi. “Voice assistants and smart speakers in everyday life and in education”. In: *Informatics Educ.* 19.3 (2020), pp. 473–490. ISSN: 16485831. DOI: 10.15388/infedu.2020.21.

[97] Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. “DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments”. In: *Meet. Acoust.* (2013), pp. 1–6. URL: https://doi.org/10.5281/zenodo.1227121%7B%5C#%7D.XJ3KmVbUOjd.mendeley%7B%5C%%7D0Ahttp://parole.loria.fr/DEMAND/.

[98] Trager, George L. “Paralanguage: A first approximation.” In: *Stud. Linguist.* 13 (1958), pp. 1–12.

[99] Dimitrios Ververidis and Constantine Kotropoulos. “Emotional speech recognition: Resources, features, and methods”. In: *Speech Commun.* 48.9 (2006), pp. 1162–1181. ISSN: 01676393. DOI: 10.1016/j.specom.2006.04.003.

[100] Emmanuel Vincent et al. “From Blind to Guided Audio Source Separation”. In: *IEEE Signal Process. ...* April (2013), pp. 1–14. URL: http://hal.inria.fr/docs/00/92/23/78/PDF/vincent%7B%5C_%7DSPM14.pdf.

[101] Aku Visuri et al. “Quantifying sources and types of smartwatch usage sessions”. In: *Conf. Hum. Factors Comput. Syst. - Proc.* 2017-May (2017), pp. 3569–3581. DOI: 10.1145/3025453.3025817.

[102] Bogdan Vlasenko et al. “Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications”. In: *Comput. Speech Lang.* 28.2 (2014), pp. 483–500. ISSN: 08852308. DOI: 10.1016/j.csl.2012.11.003. URL: http://dx.doi.org/10.1016/j.csl.2012.11.003.

[103] *VoxSort Diarization: who spoke when?* URL: https://www.voice-sort.com/%7B%5C#%7Doverview (visited on 11/30/2021).

[104] Wenxuan Wang et al. “Learning to Augment Expressions for Few-shot Fine-grained Facial Expression Recognition”. In: 14.8 (2020), pp. 1–17. arXiv: 2001.06144. URL: http://arxiv.org/abs/2001.06144.

[105] Yaqing Wang et al. “Generalizing from a Few Examples: A Survey on Few-shot Learning”. In: *ACM Comput. Surv.* 53.3 (2020). ISSN: 15577341. DOI: 10.1145/3386252. arXiv: 1904.05046.

[106] Jingyao Wu et al. “Multimodal Affect Models: An Investigation of Relative Salience of Audio and Visual Cues for Emotion Prediction”. In: *Front. Comput. Sci.* 3.December (2021), pp. 1–12. ISSN: 26249898. DOI: 10.3389/fcomp.2021.767767.

[107] Megha Yadav et al. “Exploring individual differences of public speaking anxiety in real-life and virtual presentations”. In: *IEEE Trans. Affect. Comput.* X.X (2020), pp. 1–15. ISSN: 19493045. DOI: 10.1109/TAFFC.2020.3048299.

[108] Xi Yang, Marco Aurisicchio, and Weston Baxter. “Understanding affective experiences with conversational agents”. In: *Conf. Hum. Factors Comput. Syst. - Proc.* (2019), pp. 1–12. DOI: 10.1145/3290605.3300772.

[109] Zhihong Zeng et al. “A survey of affect recognition methods: Audio, visual, and spontaneous expressions”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 31.1 (2009), pp. 39–58. ISSN: 01628828. DOI: 10.1109/TPAMI.2008.52.

[110] Biqiao Zhang. “Improving the generalizability of emotion recognition systems: Towards emotion recognition in the wild”. In: *ICMI 2016 - Proc. 18th ACM Int. Conf. Multimodal Interact.* (2016), pp. 582–586. DOI: 10.1145/2993148.2997625.

[111] W. Zi, L. S. Ghoraie, and S. Prince. *Tutorial #2: few-shot learning and meta-learning I*. 2019. URL: https://www.borealisai.com/en/blog/tutorial-2-few-shot-learning-and-meta-learning-i/ (visited on 11/29/2021).

A Supplementary Material

A.1. List of Features in the eGeMAPSv02

1. `F0semitoneFrom27.5Hz_sma3nz_amean'`
2. `F0semitoneFrom27.5Hz_sma3nz_stddevNorm`
3. `F0semitoneFrom27.5Hz_sma3nz_percentile20.0`
4. `F0semitoneFrom27.5Hz_sma3nz_percentile50.0`
5. `F0semitoneFrom27.5Hz_sma3nz_percentile80.0`
6. `F0semitoneFrom27.5Hz_sma3nz_pctlrange0-2`
7. `F0semitoneFrom27.5Hz_sma3nz_meanRisingSlope`
8. `F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope`
9. `F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope`
10. `F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope`
11. `loudness_sma3_amean`
12. `loudness_sma3_stddevNorm`
13. `loudness_sma3_percentile20.0`
14. `loudness_sma3_percentile50.0`
15. `loudness_sma3_percentile80.0`
16. `loudness_sma3_pctlrange0-2`
17. `loudness_sma3_meanRisingSlope`
18. `loudness_sma3_stddevRisingSlope`
19. `loudness_sma3_meanFallingSlope`
20. `loudness_sma3_stddevFallingSlope`
21. `spectralFlux_sma3_amean`
22. `spectralFlux_sma3_stddevNorm`
23. `mfcc1_sma3_amean`
24. `mfcc1_sma3_stddevNorm`
25. `mfcc2_sma3_amean`

26. `mfcc2_sma3_stddevNorm`
27. `mfcc3_sma3_amean`
28. `mfcc3_sma3_stddevNorm`
29. `mfcc4_sma3_amean`
30. `mfcc4_sma3_stddevNorm`
31. `jitterLocal_sma3nz_amean`
32. `jitterLocal_sma3nz_stddevNorm`
33. `shimmerLocaldB_sma3nz_amean`
34. `shimmerLocaldB_sma3nz_stddevNorm`
35. `HNRdBACF_sma3nz_amean`
36. `HNRdBACF_sma3nz_stddevNorm`
37. `logRelF0-H1-H2_sma3nz_amean`
38. `logRelF0-H1-H2_sma3nz_stddevNorm`
39. `logRelF0-H1-A3_sma3nz_amean`
40. `logRelF0-H1-A3_sma3nz_stddevNorm`
41. `F1frequency_sma3nz_amean`
42. `F1frequency_sma3nz_stddevNorm`
43. `F1bandwidth_sma3nz_amean`
44. `F1bandwidth_sma3nz_stddevNorm`
45. `F1amplitudeLogRelF0_sma3nz_amean`
46. `F1amplitudeLogRelF0_sma3nz_stddevNorm`
47. `F2frequency_sma3nz_amean`
48. `F2frequency_sma3nz_stddevNorm`
49. `F2bandwidth_sma3nz_amean`
50. `F2bandwidth_sma3nz_stddevNorm`
51. `F2amplitudeLogRelF0_sma3nz_amean`
52. `F2amplitudeLogRelF0_sma3nz_stddevNorm`
53. `F3frequency_sma3nz_amean`
54. `F3frequency_sma3nz_stddevNorm`
55. `F3bandwidth_sma3nz_amean`
56. `F3bandwidth_sma3nz_stddevNorm`
57. `F3amplitudeLogRelF0_sma3nz_amean`
58. `F3amplitudeLogRelF0_sma3nz_stddevNorm`
59. `alphaRatioV_sma3nz_amean`
60. `alphaRatioV_sma3nz_stddevNorm`
61. `hammarbergIndexV_sma3nz_amean`
62. `hammarbergIndexV_sma3nz_stddevNorm`
63. `slopeV0-500_sma3nz_amean`

64. `slopeV0-500_sma3nz_stddevNorm`
65. `slopeV500-1500_sma3nz_amean`
66. `slopeV500-1500_sma3nz_stddevNorm`
67. `spectralFluxV_sma3nz_amean`
68. `spectralFluxV_sma3nz_stddevNorm`
69. `mfcc1V_sma3nz_amean`
70. `mfcc1V_sma3nz_stddevNorm`
71. `mfcc2V_sma3nz_amean`
72. `mfcc2V_sma3nz_stddevNorm`
73. `mfcc3V_sma3nz_amean`
74. `mfcc3V_sma3nz_stddevNorm`
75. `mfcc4V_sma3nz_amean`
76. `mfcc4V_sma3nz_stddevNorm`
77. `alphaRatioUV_sma3nz_amean`
78. `hammarbergIndexUV_sma3nz_amean`
79. `slopeUV0-500_sma3nz_amean`
80. `slopeUV500-1500_sma3nz_amean`
81. `spectralFluxUV_sma3nz_amean`
82. `loudnessPeaksPerSec`
83. `VoicedSegmentsPerSec`
84. `MeanVoicedSegmentLengthSec`
85. `StddevVoicedSegmentLengthSec`
86. `MeanUnvoicedSegmentLength`
87. `StddevUnvoicedSegmentLength`
88. `equivalentSoundLevel_dBp`

A.2. Participant Data

Table A.1: Results from Vox-Sort Diarization of K-EmoCon

Audio File	Speaker	Diarized Audio (s)
P1.P2	Participant 1	5:56
	Participant 2	7:22
P3.P4	Participant 3	6:02
	Participant 4	4:04
P5.P6	Participant 5	5:50
	Participant 6	4:06
P7.P8	Participant 7	4:25
	Participant 8	5:38
P9.P10	Participant 9	6:49
	Participant 10	5:10
P11.P12	Participant 11	5:17
	Participant 12	4:50
P13.P14	Participant 13	5:07
	Participant 14	4:42
P15.P16	Participant 15	5:09
	Participant 16	4:25
P17.P18	Participant 17	6:05
	Participant 18	4:03
P19.P20	Participant 19	4:24
	Participant 20	5:20
P21.P22	Participant 21	4:16
	Participant 22	5:21
P23.P24	Participant 23	5:29
	Participant 24	5:32
P25.P26	Participant 25	4:46
	Participant 26	5:39
P27.P28	Participant 27	4:39
	Participant 28	5:29
P29.P30	Participant 29	4:38
	Participant 30	5:09
P31.P32	Participant 31	4:38
	Participant 32	4:40

B

Detailed Results

B.1. K-EmoCon : Detailed Baseline Results

Table B.1: Performance Metrics for K-EmoCon - Baseline

Id	K	Arousal Model					Valence Model				
		A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)	A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)
15	5	85.00	52.18	53.34	52.18	47.67	90.00	52.37	50.63	50.26	41.79
	10	95.00	53.02	53.06	53.02	52.83	97.50	62.57	62.74	62.57	62.44
	15	90.00	52.39	54.19	52.39	46.69	93.33	53.53	58.50	53.53	45.57
	20	97.50	54.78	54.97	54.78	54.33	96.25	53.66	53.66	53.66	53.65
	25	79.00	57.74	57.99	57.74	57.40	96.00	70.06	62.11	62.10	62.10
	30	72.50	52.76	52.84	52.76	52.41	98.33	55.44	51.77	51.70	51.19
23	5	100.00	53.76	54.94	51.17	39.67	95.00	52.48	52.68	52.48	51.62
	10	72.50	56.07	55.83	55.83	55.83	95.00	56.21	56.30	56.21	56.06
	15	88.33	54.02	55.27	54.02	51.12	100.00	56.34	56.63	56.34	55.85
	20	100.00	76.32	71.51	52.63	39.32	100.00	56.72	57.12	56.72	56.09
	25	100.00	78.61	75.01	59.44	51.97	94.00	59.84	58.69	58.27	57.75
	30	90.83	52.94	53.43	52.94	51.21	95.83	55.93	57.66	55.93	53.30
30	5	90.00	54.58	54.64	54.58	54.41	90.00	52.52	56.26	52.52	44.16
	10	87.50	53.76	54.35	53.76	52.14	100.00	54.67	57.12	54.67	50.40
	15	98.33	54.92	49.18	49.18	49.13	95.00	54.32	59.57	54.32	47.05
	20	88.75	53.91	54.83	53.91	51.63	100.00	57.63	56.13	55.34	53.86
	25	84.00	56.88	61.71	56.88	51.92	77.00	52.78	52.78	52.78	52.77
	30	93.33	55.21	55.21	55.21	55.20	97.50	57.08	57.27	57.08	56.81
31	5	100.00	55.74	59.14	55.74	51.20	80.00	50.18	75.04	50.18	33.73
	10	100.00	63.18	63.40	63.18	63.02	100.00	69.53	69.54	69.53	69.52
	15	100.00	58.74	60.85	58.74	56.64	88.33	50.57	51.57	50.57	41.20
	20	100.00	70.86	67.81	66.55	65.94	100.00	75.49	74.72	74.71	74.70
	25	96.00	52.63	52.69	52.63	52.39	98.00	84.94	69.03	67.07	66.20
	30	100.00	68.08	68.18	68.08	68.03	100.00	88.59	82.41	78.84	78.24

B.2. K-EmoCon : Detailed Results with different types of noise

Table B.2: Performance Metrics for K-EmoCon with noise - DKITCHEN

Id	K	Arousal Model					Valence Model				
		A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)	A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)
15	5	100.00	51.45	52.24	51.45	46.75	80.00	52.89	56.68	52.89	45.12
	10	95.00	52.55	54.55	52.55	46.69	85.00	55.28	55.54	55.28	54.75
	15	91.67	53.74	54.56	53.74	51.58	93.33	54.09	58.79	54.09	47.02
	20	95.00	56.08	57.04	56.08	54.52	96.25	52.76	50.36	50.31	48.44
	25	76.00	54.55	54.68	54.55	54.21	98.00	74.04	69.88	69.87	69.87
	30	98.33	54.91	55.49	54.91	53.68	79.17	55.37	61.55	55.37	48.48
23	5	75.00	52.58	53.24	52.58	50.05	100.00	54.66	58.74	54.66	48.67
	10	77.50	55.85	56.10	55.85	55.41	100.00	54.28	50.35	50.33	49.66
	15	78.33	53.33	53.43	53.33	53.02	93.33	52.11	52.12	52.11	52.09
	20	100.00	65.53	65.30	61.32	58.62	98.75	54.81	48.12	48.15	47.97
	25	88.00	53.85	55.25	53.85	50.55	94.00	54.00	59.47	54.00	46.24
	30	100.00	60.63	57.20	57.18	57.16	100.00	57.02	57.07	57.02	56.95
30	5	100.00	52.86	55.60	52.86	46.28	100.00	55.66	56.50	55.66	54.19
	10	100.00	56.11	57.41	56.11	54.08	100.00	51.69	53.54	51.69	44.43
	15	100.00	57.79	65.13	57.79	51.96	88.33	54.61	58.03	54.61	49.20
	20	92.50	53.54	51.22	50.88	47.29	95.00	56.02	56.02	56.02	56.01
	25	80.00	53.88	54.33	53.88	52.67	91.00	54.44	59.15	54.44	47.70
	30	100.00	55.15	51.03	51.03	51.02	100.00	57.83	52.18	52.17	52.12
31	5	100.00	61.48	63.30	61.48	60.11	85.00	52.12	53.03	52.12	48.24
	10	100.00	59.76	59.99	59.76	59.53	100.00	57.79	58.52	57.79	56.86
	15	100.00	65.45	65.49	65.45	65.43	100.00	51.89	53.54	51.89	45.57
	20	97.50	72.46	75.93	57.61	48.51	81.25	56.64	56.66	56.64	56.62
	25	100.00	74.44	70.61	61.85	57.32	94.00	57.79	56.78	56.35	55.66
	30	100.00	65.96	60.58	60.58	60.58	100.00	81.04	81.68	77.71	76.99

Table B.3: Performance Metrics for K-EmoCon with noise - DLIVING

Id	K	Arousal Model					Valence Model				
		$A_s(\%)$	$A_p(\%)$	$P(\%)$	$R(\%)$	$F1(\%)$	$A_s(\%)$	$A_p(\%)$	$P(\%)$	$R(\%)$	$F1(\%)$
15	5	100.00	51.93	51.93	51.93	51.92	100.00	58.68	61.13	58.68	56.29
	10	82.50	55.87	56.14	55.87	55.38	85.00	54.14	54.62	54.14	52.94
	15	100.00	54.52	54.15	53.99	53.54	100.00	52.91	52.97	52.91	52.64
	20	83.75	55.31	56.47	55.31	53.20	100.00	62.28	58.03	57.78	57.46
	25	99.00	53.49	55.27	53.49	49.20	97.00	65.58	66.49	65.58	65.11
	30	88.33	53.73	56.67	53.73	47.98	100.00	58.94	56.67	56.62	56.54
23	5	100.00	52.34	46.36	47.66	42.52	80.00	54.04	58.40	54.04	47.18
	10	90.00	55.88	49.11	49.26	46.96	100.00	55.92	56.13	55.92	55.54
	15	90.00	53.57	53.76	53.57	52.98	96.67	51.04	51.11	51.04	50.31
	20	67.50	58.02	58.05	58.02	57.99	98.75	55.11	55.47	55.11	54.36
	25	100.00	84.55	75.43	51.69	36.97	70.00	55.81	63.03	55.81	48.71
	30	81.67	54.24	56.34	54.24	50.09	94.17	55.51	55.58	55.51	55.37
30	5	100.00	51.43	52.17	51.43	46.87	100.00	53.77	53.77	53.77	53.77
	10	90.00	53.46	53.46	53.46	53.46	97.50	54.33	51.67	51.67	51.66
	15	86.67	55.28	55.57	55.28	54.70	100.00	53.82	53.82	53.82	53.81
	20	75.00	56.58	56.67	56.58	56.44	92.50	57.31	59.05	57.31	55.15
	25	91.00	60.45	60.53	60.45	60.39	90.00	55.60	61.91	55.60	48.82
	30	93.33	54.81	56.98	54.81	51.00	71.67	53.95	62.92	53.95	44.27
31	5	100.00	50.82	50.91	50.82	49.52	100.00	55.12	55.17	55.12	55.03
	10	100.00	58.81	64.89	57.63	51.74	100.00	55.51	56.01	55.51	54.58
	15	100.00	66.08	67.05	66.08	65.60	100.00	54.15	54.45	54.15	53.36
	20	81.25	61.47	62.83	61.47	60.42	100.00	75.00	70.08	70.08	70.08
	25	100.00	70.52	70.72	70.52	70.45	96.00	82.11	80.07	79.07	78.89
	30	100.00	65.46	65.68	65.46	65.34	100.00	78.31	78.45	78.31	78.28

Table B.4: Performance Metrics for K-EmoCon with noise - OHALLWAY

Id	K	Arousal Model					Valence Model				
		A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)	A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)
15	5	90.00	50.73	53.45	50.73	38.62	80.00	53.42	39.53	39.74	39.43
	10	92.50	54.04	54.93	54.04	51.87	82.50	53.04	62.77	53.04	41.99
	15	100.00	53.72	56.76	53.72	47.87	90.00	55.56	58.27	55.56	51.58
	20	93.75	55.52	57.18	55.52	52.80	91.25	54.94	58.18	54.94	49.98
	25	82.00	54.91	55.75	54.91	53.21	92.00	62.10	60.27	58.92	57.52
	30	80.83	54.60	63.37	54.60	45.70	100.00	61.11	58.44	58.33	58.20
23	5	90.00	51.88	59.59	51.88	39.77	100.00	53.09	57.13	53.09	45.34
	10	65.00	58.45	58.53	58.45	58.36	95.00	55.84	55.85	55.84	55.84
	15	100.00	53.08	53.11	53.08	52.94	86.67	52.11	59.93	52.11	40.38
	20	97.50	58.99	54.35	53.97	52.94	97.50	54.44	50.41	50.37	49.26
	25	100.00	77.12	76.07	63.28	58.14	82.00	54.69	54.70	54.69	54.66
	30	100.00	77.84	77.59	59.38	51.35	100.00	57.50	53.43	53.33	53.01
30	5	60.00	50.71	75.18	50.71	34.89	100.00	55.06	60.33	51.27	37.57
	10	87.50	52.65	52.73	52.65	52.29	97.50	52.35	55.67	52.35	44.18
	15	100.00	52.80	56.83	52.80	44.64	90.00	53.19	55.95	53.19	47.06
	20	100.00	55.60	55.88	55.60	55.08	98.75	54.55	52.27	52.27	52.27
	25	86.00	52.80	55.82	52.80	45.78	100.00	56.91	50.43	50.41	49.66
	30	95.00	55.39	59.48	55.39	50.00	93.33	52.56	53.44	52.56	49.34
31	5	100.00	65.08	65.45	65.08	64.87	100.00	55.85	50.00	50.00	35.13
	10	100.00	53.39	53.39	53.39	53.38	95.00	59.34	59.57	59.34	59.10
	15	100.00	65.03	67.13	65.03	63.93	100.00	73.67	72.15	70.27	69.62
	20	100.00	70.76	70.34	70.04	69.92	92.50	60.51	70.91	55.45	45.34
	25	98.00	54.85	55.33	54.85	53.82	93.00	80.89	75.72	74.39	74.05
	30	100.00	61.11	60.78	59.77	58.80	98.33	56.64	56.88	56.64	56.26

Table B.5: Performance Metrics for K-EmoCon with noise - OOFFICE

Id	K	**Arousal Model**					**Valence Model**				
		A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)	A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)
15	5	80.00	50.49	50.55	50.49	48.93	100.00	53.68	54.06	53.68	52.58
	10	82.50	52.81	52.98	52.81	52.12	97.50	52.46	60.44	52.46	41.23
	15	71.67	55.65	58.29	55.65	51.81	90.00	52.29	56.88	52.29	42.72
	20	100.00	56.78	55.66	55.65	55.62	96.25	54.29	54.89	54.29	52.85
	25	98.00	57.60	57.63	57.60	57.56	97.00	70.25	66.50	66.46	66.43
	30	95.00	52.13	53.15	52.13	47.95	100.00	67.81	66.47	66.44	66.42
23	5	100.00	57.98	57.90	55.87	52.84	100.00	52.17	54.89	52.17	44.46
	10	90.00	57.11	57.18	57.11	56.99	85.00	54.25	52.23	51.96	50.48
	15	96.67	53.59	53.81	53.59	52.91	98.33	54.55	55.18	54.55	53.11
	20	97.50	53.49	53.76	53.49	52.67	92.50	52.26	52.34	52.26	51.83
	25	98.00	66.67	68.18	66.67	65.96	74.00	55.24	56.12	55.24	53.57
	30	99.17	65.41	65.80	65.41	65.19	96.67	52.82	52.96	52.82	52.26
30	5	80.00	52.86	52.86	52.86	52.86	80.00	55.70	56.18	55.70	54.82
	10	97.50	54.17	54.38	54.17	53.61	97.50	53.69	53.70	53.69	53.66
	15	93.33	54.44	54.50	54.44	54.27	98.33	53.19	54.82	53.19	48.87
	20	81.25	56.36	57.55	56.36	54.55	85.00	55.77	57.69	55.77	52.82
	25	96.00	55.61	52.83	52.80	52.68	98.00	54.69	55.38	54.69	53.18
	30	91.67	52.48	52.88	52.48	50.73	99.17	55.93	56.19	55.93	55.47
31	5	100.00	65.57	65.79	65.57	65.45	80.00	51.06	56.53	51.06	38.10
	10	100.00	60.34	60.96	60.34	59.77	65.00	54.95	58.46	54.95	49.72
	15	100.00	61.15	63.20	61.15	59.58	95.00	52.84	56.58	52.84	45.03
	20	100.00	63.18	57.91	57.76	57.56	98.75	59.22	56.42	56.27	56.02
	25	100.00	64.18	59.55	51.31	37.91	97.00	85.91	78.17	77.98	77.94
	30	100.00	68.92	66.91	66.60	66.45	99.17	57.02	52.69	52.69	52.66

B.3. RECOLA : Detailed Baseline Results

Table B.6: Performance Metrics for RECOLA - Baseline

Id	**K**	**Arousal Model**					**Valence Model**				
		A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)	A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)
dev_1	10	57.50	50.07	55.02	50.07	33.73	95.00	51.03	52.24	51.03	43.38
	20	92.50	50.07	25.00	50.00	33.33	90.00	51.05	54.31	51.05	39.66
	30	97.50	50.51	51.09	50.51	42.83	64.17	51.16	51.16	51.16	51.15
	40	88.13	53.25	54.18	53.25	50.50	96.88	52.30	25.00	50.00	33.33
	50	96.50	50.00	25.00	50.00	33.33	99.50	53.42	25.00	50.00	33.33
	60	100.00	54.38	25.00	50.00	33.33	57.92	50.85	51.10	50.85	47.94
dev_2	10	40.00	49.86	24.97	49.86	33.27	90.00	50.89	52.37	50.89	41.79
	20	88.75	50.42	56.93	50.42	35.21	100.00	53.28	50.71	50.21	39.56
	30	100.00	58.85	58.86	58.85	58.85	99.17	53.73	25.00	50.00	33.33
	40	67.50	50.37	50.81	50.37	42.59	93.75	53.52	54.20	53.52	51.56
	50	99.50	60.47	60.77	60.47	60.20	84.00	54.22	54.71	54.22	53.00
	60	93.33	50.08	75.02	50.08	33.51	97.92	53.26	25.00	50.00	33.33
dev_3	10	82.50	51.44	52.03	51.44	47.59	87.50	53.28	54.58	53.28	49.71
	20	100.00	53.17	53.48	53.17	52.12	81.25	51.13	52.54	51.13	43.21
	30	95.83	55.19	55.28	55.19	55.02	98.33	54.90	25.00	50.00	33.33
	40	94.38	54.15	54.42	54.15	53.44	98.75	59.00	62.38	51.11	36.69
	50	69.50	52.90	53.34	52.90	51.31	70.50	50.75	51.55	50.75	43.49
	60	99.17	50.94	25.00	50.00	33.33	77.50	52.45	54.14	52.45	47.04
dev_4	10	90.00	51.78	51.85	51.78	51.33	100.00	51.99	25.00	50.00	33.33
	20	81.25	52.96	53.01	52.96	52.74	95.00	52.95	46.91	49.51	36.07
	30	90.00	50.58	52.85	50.58	38.26	96.67	50.00	25.00	50.00	33.33
	40	90.00	50.59	66.82	50.59	34.88	70.00	50.67	50.83	50.67	48.10
	50	91.00	51.88	52.16	51.88	50.27	50.00	50.00	50.00	50.00	33.99
	60	79.17	50.78	50.88	50.78	49.30	86.25	53.76	54.58	53.76	51.59

B.4. RECOLA : Detailed Results with different types of noise

Table B.7: Performance Metrics for RECOLA with noise - DKITCHEN

Id	K	Arousal Model					Valence Model				
		$A_s(\%)$	$A_p(\%)$	$P(\%)$	$R(\%)$	$F1(\%)$	$A_s(\%)$	$A_p(\%)$	$P(\%)$	$R(\%)$	$F1(\%)$
dev_1	10	62.50	50.07	52.80	50.07	33.96	80.00	51.30	58.13	51.30	38.37
	20	96.25	56.97	56.97	56.97	56.97	95.00	51.26	25.00	50.00	33.33
	30	99.17	50.36	25.00	50.00	33.33	51.67	52.10	52.11	52.10	52.03
	40	99.37	56.54	25.00	50.00	33.33	76.88	50.51	50.72	50.51	46.67
	50	94.00	50.00	45.27	49.62	34.58	80.00	51.60	53.01	51.60	45.17
	60	96.67	50.00	40.30	48.51	34.69	99.58	51.71	51.83	51.71	50.91
dev_2	10	95.00	56.31	58.65	56.31	53.14	92.50	55.00	58.22	55.00	50.11
	20	80.00	50.77	56.05	50.77	37.05	88.75	51.26	54.93	51.26	40.11
	30	37.50	49.86	24.96	49.86	33.27	80.83	50.79	54.57	50.79	37.95
	40	98.12	51.04	52.64	50.22	35.45	91.87	52.63	52.73	52.63	52.19
	50	93.00	57.00	59.13	57.00	54.34	85.00	50.53	50.92	50.53	44.62
	60	97.50	56.41	57.39	56.41	54.90	79.58	52.67	55.46	52.67	45.75
dev_3	10	90.00	52.54	52.57	52.54	52.38	90.00	50.89	54.94	50.89	38.24
	20	41.25	49.93	24.98	49.93	33.30	87.50	50.91	52.41	50.91	41.89
	30	98.33	51.15	53.49	51.15	41.32	90.83	55.08	59.07	55.08	49.53
	40	100.00	52.65	49.33	49.56	44.85	88.13	51.33	51.35	51.33	51.13
	50	67.00	52.92	54.39	52.92	48.61	96.00	58.86	25.00	50.00	33.33
	60	59.17	52.82	52.83	52.82	52.77	99.58	59.13	63.77	52.68	40.76
dev_4	10	82.50	51.30	52.38	51.30	45.10	90.00	51.92	52.57	51.92	48.67
	20	96.25	50.28	51.36	50.28	37.99	86.25	51.69	52.89	51.69	46.07
	30	96.67	52.52	52.63	52.52	51.99	83.33	53.40	53.73	53.40	52.33
	40	100.00	51.04	25.00	50.00	33.33	86.25	50.74	51.46	50.74	43.85
	50	74.00	52.43	52.46	52.43	52.29	87.50	53.29	53.55	53.29	52.43
	60	85.00	51.86	53.61	51.86	45.21	49.58	50.00	25.00	50.00	33.33

Table B.8: Performance Metrics for RECOLA with noise - DLIVING

Id	**K**	**Arousal Model**					**Valence Model**				
		A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)	A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)
dev_1	10	100.00	57.13	59.25	57.13	54.53	100.00	51.85	51.85	51.85	51.82
	20	97.50	56.62	56.62	56.62	56.61	96.25	52.53	52.62	52.53	52.11
	30	45.00	49.64	24.91	49.64	33.17	70.00	51.37	52.42	51.37	45.49
	40	90.62	54.00	54.00	54.00	53.99	86.87	50.44	50.71	50.44	45.27
	50	97.00	55.80	56.43	55.80	54.69	95.50	54.79	54.80	54.79	54.77
	60	45.00	50.39	52.55	50.39	37.09	69.58	51.87	52.05	51.87	50.76
dev_2	10	100.00	50.34	55.52	50.34	35.13	90.00	50.96	55.90	50.96	37.97
	20	97.50	60.49	61.06	60.49	59.98	100.00	56.29	55.63	55.24	54.47
	30	95.00	50.22	58.39	50.22	34.19	100.00	54.66	25.00	50.00	33.33
	40	90.62	50.67	67.48	50.67	35.05	93.75	51.03	51.88	51.03	44.74
	50	99.50	60.44	60.53	60.44	60.36	83.00	50.75	52.29	50.75	40.76
	60	83.33	50.39	61.46	50.39	34.61	72.92	53.57	57.53	53.57	46.55
dev_3	10	95.00	51.23	52.06	51.23	45.82	100.00	55.13	54.46	54.10	53.17
	20	92.50	50.35	52.02	50.35	37.45	96.25	54.00	55.23	54.00	51.12
	30	78.33	53.60	53.62	53.60	53.56	67.50	57.26	58.95	57.26	55.13
	40	86.87	55.27	55.49	55.27	54.84	95.63	55.15	57.91	55.15	50.85
	50	94.50	54.82	55.11	54.82	54.19	52.50	50.22	60.77	50.22	34.09
	60	96.25	51.02	54.60	51.02	39.19	96.67	54.84	25.00	50.00	33.33
dev_4	10	100.00	53.22	25.00	50.00	33.33	62.50	50.07	50.37	50.07	37.30
	20	98.75	50.92	25.00	50.00	33.33	61.25	51.13	53.13	51.13	41.82
	30	95.83	50.72	51.81	50.72	42.03	95.00	51.81	25.00	50.00	33.33
	40	93.12	51.33	53.37	51.33	42.66	93.12	50.07	50.22	50.07	40.25
	50	82.50	51.96	52.50	51.96	49.24	81.00	52.74	52.74	52.74	52.73
	60	56.25	50.54	57.14	50.54	35.68	72.50	53.23	54.66	53.23	49.34

Table B.9: Performance Metrics for RECOLA with noise - OHALLWAY

Id	K	Arousal Model					Valence Model				
		A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)	A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)
dev_1	10	52.50	49.66	24.91	49.66	33.18	80.00	51.65	54.51	51.65	42.52
	20	100.00	57.33	25.00	50.00	33.33	100.00	52.67	25.00	50.00	33.33
	30	99.17	54.62	55.34	54.62	53.04	100.00	52.89	51.75	50.58	40.64
	40	92.50	54.02	54.51	54.02	52.73	93.75	51.03	51.19	51.03	49.37
	50	98.00	58.02	25.00	50.00	33.33	87.50	51.90	51.99	51.90	51.30
	60	49.58	50.16	58.37	50.16	33.95	93.33	54.57	54.58	54.57	54.57
dev_2	10	77.50	50.69	57.99	50.69	36.08	95.00	58.21	59.80	58.21	56.43
	20	100.00	61.57	61.61	61.57	61.53	81.25	53.57	57.71	53.57	46.38
	30	100.00	61.43	57.65	53.84	47.26	75.00	54.86	57.14	54.86	50.93
	40	98.75	51.19	25.00	50.00	33.33	71.88	50.51	61.80	50.51	34.96
	50	92.50	57.12	58.15	57.12	55.72	96.00	57.62	58.75	57.62	56.20
	60	100.00	61.24	61.08	60.77	60.49	92.08	54.88	25.00	50.00	33.33
dev_3	10	50.00	50.48	54.19	50.48	36.40	82.50	54.51	54.54	54.51	54.45
	20	95.00	54.99	55.06	54.99	54.84	96.25	58.01	60.44	58.01	55.40
	30	99.17	53.82	25.00	50.00	33.33	92.50	58.74	65.35	58.74	53.76
	40	61.87	53.26	54.04	53.26	50.90	99.37	58.41	63.96	54.57	45.39
	50	99.50	51.83	25.00	50.00	33.33	99.00	54.98	56.69	51.66	40.47
	60	100.00	51.40	49.26	49.38	47.29	75.42	50.46	50.75	50.46	45.12
dev_4	10	100.00	52.61	48.44	48.77	45.95	100.00	53.77	55.49	53.77	49.84
	20	90.00	50.77	64.67	50.77	35.50	96.25	50.92	51.55	50.92	45.31
	30	83.33	51.16	51.66	51.16	47.17	73.33	51.37	52.89	51.37	44.05
	40	94.38	53.69	53.85	53.69	53.21	58.75	50.07	53.59	50.07	33.89
	50	99.00	52.13	25.00	50.00	33.33	83.00	51.97	51.97	51.97	51.97
	60	89.58	52.09	56.29	50.16	34.08	87.92	52.04	52.20	52.04	51.17

Table B.10: Performance Metrics for RECOLA with noise - OOFFICE

Id	K	Arousal Model					Valence Model				
		A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)	A_s(%)	A_p(%)	P(%)	R(%)	$F1$(%)
dev_1	10	87.50	56.79	57.03	56.79	56.42	95.00	52.95	53.05	52.95	52.55
	20	56.25	50.28	52.00	50.28	36.68	97.50	54.08	25.00	50.00	33.33
	30	94.17	50.00	25.00	50.00	33.33	88.33	51.81	48.91	48.99	48.07
	40	78.75	57.30	57.32	57.30	57.28	94.38	51.63	52.41	51.63	47.33
	50	86.50	55.18	55.29	55.18	54.96	82.00	52.20	54.78	52.20	44.75
	60	50.83	50.23	53.06	50.23	35.29	98.33	52.25	46.21	48.76	38.41
dev_2	10	97.50	50.41	50.00	50.00	33.70	92.50	54.65	54.67	54.65	54.61
	20	100.00	59.34	60.05	59.34	58.61	90.00	52.94	52.95	52.94	52.91
	30	98.33	58.26	58.29	58.26	58.22	67.50	51.36	57.45	51.36	38.86
	40	99.37	60.01	25.00	50.00	33.33	90.62	57.69	58.88	57.69	56.22
	50	99.00	59.27	25.00	50.00	33.33	79.50	57.51	58.07	57.51	56.75
	60	88.75	61.66	62.64	61.66	60.90	100.00	53.42	25.00	50.00	33.33
dev_3	10	97.50	51.92	46.63	47.47	44.00	100.00	56.90	57.96	56.90	55.42
	20	100.00	50.63	51.79	50.63	41.12	68.75	50.49	75.12	50.49	34.42
	30	70.83	50.22	50.30	50.22	46.63	78.33	50.51	53.70	50.51	36.88
	40	78.12	52.53	53.48	52.53	49.02	76.88	56.17	57.57	56.17	54.04
	50	84.00	51.22	56.38	51.22	38.85	100.00	58.55	25.00	50.00	33.33
	60	47.92	50.00	50.00	50.00	34.01	68.33	50.23	52.65	50.23	35.52
dev_4	10	97.50	52.06	52.81	52.06	48.62	100.00	50.75	51.31	50.55	42.17
	20	68.75	50.63	56.98	50.63	36.10	93.75	54.51	54.58	54.51	54.36
	30	94.17	53.46	53.90	53.46	52.10	96.67	54.12	54.31	54.12	53.59
	40	55.62	53.26	53.73	53.26	51.76	93.75	50.30	25.00	50.00	33.33
	50	85.00	50.15	55.04	50.15	34.20	82.50	52.98	52.98	52.98	52.98
	60	96.67	50.47	25.00	50.00	33.33	72.08	50.71	50.71	50.71	50.70