Machine Learning of Emotional Expressions In the Wild from Acoustic Signals and Text

Sina Ali Samir

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique
Spécialité : Informatique
Unité de recherche : Laboratoire d'Informatique de Grenoble

Apprentissage automatique des expressions émotionnelles à partir de signaux acoustiques et de textes

Machine Learning of Emotional Expressions In the Wild from Acoustic Signals and Text

Présentée par :

SINA ALI SAMIR

Direction de thèse :

François PORTET PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Directeur de thèse
Fabien RINGEVAL MAITRE DE CONFERENCES, Université Grenoble Alpes	Co-encadrant de thèse

Rapporteurs :

Emily MOWER PROVOST
PROFESSEUR, University of Michigan
Mohamed CHETOUANI
PROFESSEUR DES UNIVERSITES, Sorbonnes Université

Thèse soutenue publiquement le **3 octobre 2023**, devant le jury composé de :

Emily MOWER PROVOST PROFESSEUR, University of Michigan	Rapporteuse
Mohamed CHETOUANI PROFESSEUR DES UNIVERSITES, Sorbonnes Université	Rapporteur
Hussein AL OSMAN ASSOCIATE PROFESSOR, University of Ottawa	Examinateur
Florian EYBEN INGENIEUR DOCTEUR, audEERING GmbH	Examinateur
Catherine PELACHAUD DIRECTEUR DE RECHERCHE, CNRS	Examinatrice
Martial MERMILLOD PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes	Président

Acknowledgements

I would like to express my deepest gratitude to all those who have supported me throughout my doctoral journey, without whom this thesis would not have been possible.

First and foremost, I extend my heartfelt appreciation to my academic supervisors, Fabien Ringeval, and François Portet, whose guidance, expertise, and unwavering support have been invaluable throughout this research endeavor. Your mentorships and insightful feedbacks have played a vital role in shaping the quality of this thesis.

I am also thankful to Atos as the main funder of my research, and my industry supervisors, Béatrice Bouchot, and Jean-Philippe Vigne, for their collaboration, and continuous encouragement. Working for Atos alongside Université Grenoble Alpes, has enriched my research with all the challenges involved towards building emotion recognition for a real-life application. I thus also like to extend my gratitude to ANRT (Agence Nationale de la Recherche et de la Technologie) for administering this CIFRE (Conventions industrielles de formation par la recherche) thesis, which facilitated the collaboration between academia and industry.

I am also deeply grateful to all the members of the GETALP (Groupe d'Étude en Traitement Automatisé des Langues et de la Parole) team. I am honoured to have been part of this team and to have met such wonderful people during my years as a PhD student, with whom I have shared so many lunches, so much laughter and love.

Last but not least, I would like to thank my dear parents, who have always given me unconditional support and love, as well as Reza, Pooneh and little Elena, who have motivated me throughout this journey.

Thank you all for being a part of this thesis.

Abstract

Automatic emotion recognition (AER) from text, or audio recordings of natural human-human or human-machine interactions, is a technology that can have an impact in areas as diverse as education, health and entertainment. Although existing AER systems can work well in specific scenarios, they are not yet robust enough to deal with different environments, speakers and microphones (i.e. in the wild). In this thesis, several contributions have been made to advance the research on AER in the wild.

State-of-the-art AER systems use data-driven machine learning methods to recognise emotion from numerical representations of acoustic signals or text. One contribution of this thesis is to investigate the fusion of speech representations and their corresponding textual transcriptions for AER on both acted and in-the-wild data. In addition, as human transcriptions are not always available, existing Automatic Speech Recognition (ASR) systems are further explored within the same paradigm. The results show that the use of fused acoustic-textual representations can achieve better AER performance for acted and in-the-wild data than using the representation of each modality alone. The acoustic-textual representations were further fused with speaker representations, resulting in additional improvement in AER performance for acted data.

Moreover, as emotion is a subjective concept with no universal definition, it is annotated and used in various ways across different AER systems. To address this issue, this thesis proposes a method for training a model on different datasets with different emotion annotations. The proposed method is composed of one model that is trained across multiple datasets, which computes the generic latent emotion representation, and several specific models, which can map the emotion representation to the set of emotion labels specific to each dataset. The results suggest that the proposed method can produce emotion representations that can relate the same or similar emotion labels across different datasets with different annotation schemes. Finally, by combining the proposed method with joint acoustic-textual representations, it was shown that this method can leverage acted data to improve the performance of AER in the wild.

Résumé

La reconnaissance automatique des émotions (RAE) à partir de textes ou d'enregistrements audio d'interactions naturelles entre humains ou entre humains et machines est une technologie qui peut avoir un impact dans des domaines aussi divers que l'éducation, la santé et le divertissement. Bien que les systèmes de RAE existants puissent fonctionner correctement dans des scénarios spécifiques, ils ne sont pas encore assez robustes pour être utilisés de manière fiable pour des enregistrements d'environnements, de locuteurs et de microphones différents (c.-à-d. les données naturelles). Dans cette thèse, plusieurs contributions ont été faites pour avancer la recherche sur la RAE pour les données naturelles.

Les systèmes de RAE les plus récents utilisent des méthodes d'apprentissage automatique basées sur les données pour prédire les annotations numériques des émotions à partir des représentations numériques des signaux acoustiques ou du texte. L'une des contributions de cette thèse est d'étudier la fusion des représentations vocales et de leurs transcriptions textuelles correspondantes pour la RAE sur des données actées et naturelles. En outre, comme les transcriptions humaines ne sont pas toujours disponibles, les systèmes de reconnaissance automatique de la parole (RAP) existants sont explorés dans le même paradigme. Les résultats montrent que l'utilisation de représentations acoustiques et textuelles fusionnées permet d'obtenir de meilleures performances en matière de reconnaissance automatique d'émotion pour des expressions actées et naturelles, comparé à l'utilisation séparée de chaque modalité. Les représentations acoustiques et textuelles ont également été fusionnées avec les représentations du locuteur, ce qui a permis d'améliorer les performances en RAE pour des expressions actées.

En outre, l'émotion étant un concept subjectif sans définition universelle, elle est annotée et utilisée de diverses manières dans les différents systèmes de RAE. Pour résoudre ce problème, cette thèse propose une méthode d'entraînement d'un modèle sur différents ensembles de données avec différentes annotations d'émotions. La méthode proposée est composée d'un modèle partagé entre plusieurs ensembles de données, qui calcule la représentation latente générique de l'émotion, et de plusieurs modèles spécifiques, qui peuvent faire correspondre la représentation de l'émotion à l'ensemble des étiquettes d'émotion spécifiques à chaque ensemble de données. Les résultats suggèrent que la méthode proposée peut produire

des représentations d'émotions qui peuvent relier des étiquettes d'émotions identiques ou similaires dans différents ensembles de données avec différents schémas d'annotation. Enfin, en combinant la méthode proposée avec des représentations acoustiques et textuelles conjointes, il a été démontré que cette méthode peut exploiter les expressions émotionnelles actées pour améliorer les performances de la RAE effectuées sur des expressions naturelles.

Contents

Acronyms

AAE Adversarial Auto-Encoder. 36, 37, 38, 45

AER Automatic Emotion Recognition. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 23, 25, 30, 31, 32, 33, 34, 35, 36, 37, 38, 40, 43, 44, 45, 46, 47, 48, 49, 50, 54, 55, 56, 58, 60, 61, 62, 63, 64, 65, 72, 73, 74, 75, 76, 77, 78, 79, 80, 82, 83, 84, 85, 86, 87, 88, 89, 90, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 123, 124, 125, 126, 128, 130, 131, 132, 133, 134, 135, 148

ANN Artificial Neural Network. 19, 23, 24, 33, 37, 38, 44, 51, 56, 63, 64, 65, 72

APC Autoregressive Predictive Coding. 40

ASR Automatic Speech Recognition. 4, 7, 8, 9, 10, 12, 25, 32, 45, 46, 61, 67, 70, 73, 74, 75, 76, 83, 87, 94, 95, 96, 97, 98, 101, 103, 105, 106, 108, 109, 118, 123, 124, 125, 131, 133, 148

BERT Bidirectional Encoder Representations from Transformers. 43, 44, 49, 54, 60, 63

BoAW Bag of Audio Words. 21, 41, 54

CBOW Continuous Bag-Of-Words. 42

CCC Concordance Correlation Coefficient. 64, 65, 69, 72, 79, 92

CPC Contrastive Predictive Coding. 38, 39, 40

DNN Deep Neural Network. 13, 23, 25, 26, 28, 30, 31, 38, 40, 41, 42, 43, 46, 54, 55, 60, 62, 108, 130

FV Fisher Vector. 21, 41

GA Gradient Accumulation. 64, 84, 85, 86, 87, 89, 95, 100, 104, 124

GAN Generative Adversarial Networks. 37

GeMAPS Geneva Minimalistic Acoustic Parameter Set. 21, 32, 41

GMM Gaussian Mixture Model. 21

GRU Gated Recurrent Unit. 26, 27, 28, 47, 51, 52, 53, 63, 64, 77, 81, 84, 85, 87, 90, 99, 113, 114, 115, 116, 122, 124, 128, 133

HCI Human-Computer Interaction. 49

LLD Low Level Descriptors. 21, 30, 45, 46, 74, 87, 90, 92, 94, 97, 109

LOSO Leave One Speaker Out. 118

LR Learning Rate. 84, 86, 89, 95, 100, 104, 124

LSA Latent Semantic Analysis. 22, 23, 41, 42

LSTM Long Short Term Memory. 26, 27, 28, 43, 51, 52, 63, 81, 90

MAE Mean Absolute Error. 65, 93

MFB Mel-scale Filter Bank. 20, 21, 38, 41, 51, 52, 54, 61, 62, 77, 78, 79, 81, 82, 84, 108, 113

MFCC Mel-Frequency Cepstral Coefficients. 20, 31, 81, 87

MHA Multi-Head Attention. 29

MSE Mean Square Error. 65, 92

MTL Multi-Task Learning. 5, 9, 10, 12, 19, 46, 47, 48, 49, 56, 62, 69, 111, 112, 114, 116, 117, 120, 122, 123, 124, 125, 126, 127, 128, 133, 134

NLP Natural Language Processing. 44

SGD Stochastic Gradient Descent. 24, 64, 72

SSL Self-Supervised Learning. 6, 7, 10, 38, 44

STL Single-Task Learning. 47, 48

SVM Support Vector Machine. 23, 51, 54

TF-IDF Term Frequency-Inverse Document Frequency. 22, 41, 42, 54

UAR Unweighted Average Recall. 64, 67, 72, 87, 93, 116

VAD Voice Activity Detection. 148

VAE Variational Auto-Encoder. 35, 36, 37, 38

WER Word Error Rate. 68, 95, 96, 105, 125

List of Figures

List of Tables

Chapter 1

Introduction

From pixabay.com

Emotion and reason have long been thought of as two independent rival parts in human brain. It is not until many years ago that the correlation between the two was made clear thanks to a brain lesion (Damasio, 1994). Recent psychological findings suggest that it is not only the case that reasoning is needed for an emotional response, but in fact, emotion is also needed for reasoning and filtering all the constant sensory information that our brains receive every second of our lives. It is emotion that shapes our perception of the world, decides which information should stay in memory and eventually what decision we would take in different events and situations (Brosch et al., 2013). Enabling machines to perceive human emotion, can be a revolutionising technology, effecting many different domains ranging from education and health, to entertainment. In what follows, some of the use-cases of

Automatic Emotion Recognition (AER) in different domains are introduced.

In education, AER can be used to provide a better learning experience for students. For example, prediction of the learners' engagement, which relies on the analysis of emotions (Monkaresi et al., 2016), can be used to keep the students engaged by automatically changing the contents during an online course. Similarly, in entertainment, AER can be used to create more immersive experiences for users, by estimating each user's engagement. For example, to keep each user entertained, AER can predict cognitive engagement or boredom from the user's vocal expressions to create a personalised experience that keeps a user engaged. AER has also been used in serious games to stimulate cognitive functions. For instance, children with autism spectrum conditions, can play a game to learn emotion recognition in an entertaining and motivating way, resulting in an improvement in their ability to adapt socially (Fridenson-Hayo et al., 2017). The use of AER is not limited to serious games, which are used as entertainment tools with health benefits, but can also be used as a technology solely in the health and well-being domain. For example, AER can be used to remotely monitor the emotions of at-risk patients. This provides the psychologists with crucial feed-backs, and automatic assessment of cognitive disorders of the patients (Low et al., 2020). The remote monitoring of patients' emotions can further be used to inform doctors when the patients need help. The remote evaluation of the emotions, can also be used in customer service. Namely, the estimation of frustration or satisfaction (Macary et al., 2020), can be used to improve the quality of customer services, by allowing agents to identify relevant strategies to avoid customer's frustrations. The data of customers' emotions (supposing that they are gathered ethically), can also further help companies for marketing purposes.

The above examples can only shed some light on the vast possibilities that AER can bring to technology in various fields, and thus to society at large. In order to bring such technological use-cases into existence, AER needs to perform well on data collected in the wild[1], which has several characteristics, namely different speakers having natural interactions, using different microphones and in different environments (Kossaifi et al., 2021). However, AER research to date has mainly focused on acted expressions of emotions, recorded in laboratory environments. Nevertheless, this paradigm is gradually shifting towards exploiting in-the-wild data. This paradigm shift is mainly due to novel deep learning techniques, such as deep representations, which can provide us with acoustic and textual representations that are more suitable for in-the-wild emotional expressions (see Section 2.2). The novel methods of deep learning, together with the vast technological possibilities of AER,

[1]Here the term "in the wild" in the context of AER refers to emotional expressions that are the result of natural interactions between humans, or humans and a machine, recorded in a variety of environments, such as in a classroom, in public, or at home, and by using various microphones. in-the-wild data is in contrast to acted (or induced) emotional expressions, which are usually collected in controlled laboratory environments (This is further explained in Section 3.1).

have attracted the interest of the industry in recent years. For example, Atos, a digital services company based in France and the industrial partner of this CIFRE thesis[1], is interested in developing AER technology for virtual assistant applications that work with acoustic and textual input. Based on such a target application, the focus of this thesis is mainly on acoustic and textual modalities, from different speakers, with data captured in the wild. Given the technological possibilities of AER and the scope of this thesis, the next section introduces the reader to the relevant state of the art and its shortcomings.

1.1 Automatic emotion recognition

Human emotion, which is mostly studied in psychology, is often seen as an evolutionary response to internal or external (to the brain) stimuli that would result in a subjective state of mind, which can then be expressed through verbal and non-verbal communication channels (see Section 2.1.1). AER can then be defined as the automatic process of predicting human emotion from different modalities such as acoustic signals or text[2]. To accomplish this task, state-of-the-art AER uses supervised machine learning techniques to predict numerical representations of emotion, from their corresponding numerical acoustic or textual representations (see Section 2.2). However, as emotion is a concept for which several psychological theories coexist, the numerical representations of emotion are defined in various subjective ways from one machine learning method to another. This subjectivity in the representation of emotions can occur at different levels, including the psychological model used to describe the emotion, the selection of a set of emotions for a method, and the subjectivity involved in human annotation (see Section 2.1.2). In Section 1.1.1, the problem of varied emotion annotation will be discussed in more detail, followed by the state-of-the-art solutions to this problem and its current challenges. Moreover, as the use of acoustic and textual modalities for AER also falls within the

[1]The industrial agreements for training through research or CIFRE (Conventions industrielles de formation par la recherche) is a mechanism that allows companies registered under French law to recruit a doctoral student whose research project is carried out in collaboration with a public laboratory (in this case, the University of Grenoble Alpes), where the French ministry of research would pay an annual subsidy to the company (in this case, Atos).

[2]It is hypothesised here that AER aims to predict an emotional state being experienced by a user, which leads to an emotional expression. And that the emotional expression can be used to predict an affective state from acoustic signals or text. Although these hypotheses may not be true in all circumstances, one cannot predict an affective state from its expressions in speech or text if the two are not correlated. Moreover, "utilitarian emotions" (such as anger, fear, happiness, and sadness), which are the target of this study, are often considered to be highly associated with "response synchronisation", which corresponds to the response of an appraised emotion appropriate to an event (Scherer, 2005). It should also be noted that the AER should not be used in violation of the privacy of any individual.

scope of this thesis (see the paragraph above), Section 1.1.2 presents the state-of-the-art AER from acoustic and textual representations and its current shortcomings.

1.1.1 Varied emotion annotations for machine learning

In order to build a machine learning model for AER, it is first necessary to define emotion as a numerical target space (see Section 2.1.2). However, emotion is a rather ambiguous concept with various coexisting theories in psychology (see Section 2.1.1), which makes it difficult to define as a numerical target. In other similar tasks, such as Automatic Speech Recognition (ASR)[1], it is generally accepted that only the linguistic part that can be written is the objective of the ASR task. This makes speech transcriptions more or less standardised for most spoken languages, whereas emotion has no universally accepted definition, despite having been studied in psychology for more than two centuries (see Section 2.1.1). Although the definition of emotion is not standardised, most AER research today follows specific psychological theories, namely the works of Ekman (1992) and Russell (1980). According to Ekman, emotions are considered to be independent states of mind, and different independent categories should be used to define them. In particular, Ekman argued that there are six 'basic' categories of emotion –fear, anger, happiness, sadness, disgust and surprise– that can be distinguished for all humans from facial expressions. On the other hand, Russel and studies such as (Scherer, 2009), view emotion not as independent mental states, but as the continuous response of several interconnected subsystems in the brain, forming multiple perpendicular axes[2] of emotion dimensions such as arousal –activation– and valence –intrinsic pleasantness–[3] (see Section 2.1.2).

The aforementioned views of emotion are used in state-of-the-art AER systems to annotate emotional expressions for training machine learning models. However, as there is no consensus on what should be considered as an emotion annotation, different datasets consider different schemes for annotating emotion. For example, the CaFE dataset (Gournay et al., 2018) includes seven emotion labels (Following Ek-

[1] ASR is a technology that can convert spoken language to written text. To train machine learning models for ASR, the targets are usually phonemes, letters, or words, which are standardised for most spoken languages.

[2] In theory, the dimensional axes of emotion are considered to be perpendicular, suggesting the independence of each axis from the others. However, several studies have shown that there is a statistical correlation between different emotional dimensions. For example, arousal tends to increase with positive or negative valence, forming a V-shaped relationship of arousal as a function of valence (Kuppens et al., 2013, 2017).

[3] Here, the term valence refers specifically to the notion of intrinsic pleasantness, i.e. how pleasant an event is appraised by an individual, regardless of the individual's affective state of mind prior to the appraisal process (see Section 2.1.1 for the appraisal process). Although intrinsic pleasantness is a more specific term than valence, which is a broad term with multiple definitions, in this thesis the term valence is used due to its current popularity in the field of affective computing.

man's six emotions, plus one to include "neutral" expressions) for each expression, whereas the RECOLA (Ringeval et al., 2013) dataset considers arousal and valence dimensions (Following Russel's theory of affect), providing continuous emotion annotations for each dimension. Most AER research today focuses on training specific machine learning models for each dataset. However, as each dataset uses a limited range of emotional expressions and a handful of human annotators who subjectively perceive the emotional annotations, training specific machine learning models for specific datasets results in AER models that only cover a specific range of emotional expressions and annotations. Therefore, it is important to exploit multiple datasets for training AER models, to generalise across a wide range of emotional expressions.

In order to exploit multiple datasets with varied annotation schemes, AER research either unifies different annotations to standardise them across different datasets, or considers a holistic view of emotion annotations, where all annotations are treated as they were originally intended for each dataset. Unifying different emotion annotations typically involves either mapping labels to a common subset of emotion categories, or ignoring a subset of emotion labels. However, the expression and perception of the same emotion category can vary between people, causing the same emotion annotation to refer to different emotional expressions across datasets. Using the same target to model different emotional expressions, would then lead to training problems when using machine learning methods, such as catastrophic information loss (Zhang et al., 2017b; Zhu and Sato, 2020). On the other hand, the holistic view of emotion annotations is typically achieved by Multi-Task Learning (MTL), which can consider different classifiers for different tasks, while sharing a main model across tasks. For example, by considering different emotion annotations of different datasets as different tasks, MTL can provide different views of annotating the same emotional expression. The use of MTL has been shown to improve the accuracy of AER from the acoustic signals in several works (Zhang et al., 2017b, 2022). Furthermore, using MTL with the original annotations of each used dataset, has been shown to significantly outperform using the unified view of different labels, even when the unified emotions refer to the same or similar psychological phenomenon (da Silva et al., 2020). However, MTL has not yet been evaluated for deep acoustic and textual representations, and is often considered as part of a holistic view of paralinguistic features, as opposed to being studied specifically in the context of MTL to holistically account for different emotion annotations.

1.1.2 Deep acoustic and textual representations

The previous section discussed the problem of having various emotion annotations. Since training machine learning models requires both the annotations and the representations of the data, this section explores recent methods for representing acoustic

signals and text for AER. In what follows, the reader is first introduced to deep representation learning and how it has dominated the AER's state of the art in recent years. This is followed by a discussion of how the current state of the art in AER attempts to jointly represent acoustic signals and text, and making the representations more aware of "verbal" and "speaker" idiosyncrasies.

Deep pre-trained representations of data

Before explaining how deep representations have come to dominate the state of the art in AER, we first need to understand how deep learning models work. Deep learning models usually refer to complex neural network models that can capture complex patterns in data through a series of interconnected layers (see Section 2.2.2 for more information). Each layer involved in a deep learning model applies a transformation to its input, in order to achieve a higher level of abstraction for a particular task compared to the previous layer (LeCun et al., 2015; Alisamir and Ringeval, 2021). Subsequently, as we analyse different layers from input to output, each layer provides us with a more abstract representation that is less sensitive to local changes in the input data (Bengio et al., 2013) (see Figure 2.4). This means that deep learning models can learn to represent acoustic and text data at different levels, from the input data to the prediction of emotion. Through this process, deep learning models can extract more sophisticated representations of acoustic signals and text than traditional features (Naseem et al., 2021), while also jointly modelling emotion (see Section 2.2.1 for traditional methods). This approach is usually referred to as end-to-end learning, and has been shown to better predict arousal and valence dimensions from raw acoustic signals than using traditional feature extraction methods (Trigeorgis et al., 2016). However, learning deep representations in a supervised manner is not an ideal approach for AER, because the emotion annotations are subjective, and follow different emotion annotation schemes (see Section 1.1.1). Furthermore, there is a much wider range of possible emotion expressions than what is available in existing datasets. This discrepancy can lead to the inability to generalise representations to capture a wide range of emotions. Additionally, the lack of available emotion labels and the potential for incorrect annotations could lead to inaccurate representations being formed.

On the other hand, effective deep representations for AER can be trained using only unlabelled data –unsupervised learning– (see Section 2.2.3). A famous example of unsupervised learning is auto-encoders, which first encode an acoustic signal or text into a dense abstract representation, and then decode the dense representation back to the original acoustic signal or text. Another example is Self-Supervised Learning (SSL), where instead of reproducing a given signal or text, a neural network is trained to predict randomly masked elements of the data. SSL has been shown to provide representations that are able to predict emotion annotations better than traditional feature engineering techniques by using less complex models (Latif

et al., 2020; Liu et al., 2021; Nandan and Vepa, 2020; Atmaja et al., 2022; Evain et al., 2021b). This can show that SSL can learn acoustic signal representation that are more useful for AER across different emotional expressions than traditional feature extraction methods (Alisamir and Ringeval, 2021).

Joint representations of acoustic signals and text

The application of self-supervised deep representations for AER is not limited to acoustic signals, but can also be extended to textual data. In addition, recent research has focused on joint acoustic-textual representations to address how emotion is conveyed by both what is said –verbal communication– and how something is said –non-verbal communication–[1]. For example, the prosodic information of the uttered phrase "I am fine" might reveal that the speaker is angry, even though the verbal information contains no indication of this. On the other hand, the phrase "I am angry" contains direct verbal information about the speaker's emotional state.

Moreover, It can be argued that acoustic signals contain both verbal and non-verbal communication, so one does not need to use textual representations. While this is theoretically true, in practice the verbal information in the textual representations is more informative in smaller dimensions than the acoustic representations. This is because the acoustic representations contain not only the verbal information, but also other information related to the speaker's tone, gender, ambient noise and even the characteristics of the microphone. Therefore, the use of both acoustic and textual representations provides a more holistic view of both verbal and non-verbal communication. This method has been evaluated in Li and Lee (2019); Siriwardhana et al. (2020); Ho et al. (2020), where human transcriptions of speech signals were used to jointly represent acoustic and textual modalities, demonstrating an improvement in the performance of AER compared to using acoustic representations alone.

Although joint acoustic-textual representations have proven successful, in a realistic application we usually do not have access to the human transcriptions of a speech signal. Nevertheless, recent advances in ASR technologies have made it possible to obtain reliable automatic transcriptions that are comparable to human transcriptions in most circumstances (Kim et al., 2019). This has led affective computing research to focus on integrating ASR transcriptions with acoustic representations (Atmaja et al., 2022). The main trend observed in recent studies is to use ASR to provide us with transcriptions first, and then to use the joint acoustic-textual representations for AER, achieving comparable results to using human transcriptions for joint acoustic-textual representations (Heusser et al., 2019; Yoon et al., 2019;

[1] Visual cues can also be part of non-verbal communication. However, since in this thesis the focus is on acoustic and textual modalities, only the non-verbal communication from speech is considered.

Wu et al., 2021; Peng et al., 2021). Although the aforementioned studies have investigated joint representations of acoustic signals and ASR-based transcriptions, they have all focused only on acted emotional expressions, and thus the effect of such joint representations for emotional expressions captured in the wild has not yet been explored. Also, the use of joint acoustic-textual representations used for acted data to improve the performance of AER in the wild, has not yet been studied.

The previous paragraphs have discussed current research trends in adding verbal awareness to deep acoustic representations. Since different individuals express emotions in different ways, the next paragraph shifts the discussion to why and how current research attempts to personalise acoustic representations.

Personalised representations

Different people express emotions in different ways, depending on each individual's assessment of a particular situation, which depends on their psychobiological background (Sander et al., 2005; Scherer, 2009) (see Section 2.1.1). This means that a better understanding of each user's characteristics can provide insight into an individual's emotional state, and thus can be used to improve current AER models[1]. Therefore, taking the characteristics of different speakers into account, has attracted the interest of AER research, especially research on AER from speech.

There are different lines of research investigating personalisation of acoustic signals, from assignment of different personalised classifiers (Rudovic et al., 2018), to forcing disentanglement of speaker and emotion characteristics (Peri et al., 2021). However, the main line of research on this topic focuses on improving AER systems by exploiting the speaker-related representations, that are achieved by training speaker recognition models (Xi et al., 2019; Moine et al., 2021; Peng et al., 2021). For example, a recent study shows that the use of both verbal and speaker information to improve AER from the acoustic signals can outperform models that do not consider the verbal and speaker information (Ta et al., 2022). Despite recent advances made on this topic, a comparison between the effectiveness of the different modalities, and the use of ASR transcriptions, seems to be lacking from the state of the art.

1.2 The aims of the thesis

In order to advance the research on AER, the aim of this thesis is to address the aforementioned shortcomings in the state of the art. Namely the following[2]:

[1] Based on the assumption that AER targets the affective state of mind, and that the emotional expressions can be used to predict the affective state of mind.

[2] Admittedly, other issues such as limited computational resources, privacy and bias are also among the existing challenges for AER (Lee et al., 2021), which are not addressed in this thesis.

1. The state of the art shows promising results with ASR transcriptions for joint acoustic-textual representations and speaker-aware deep acoustic representations (See Section 1.1.2). Thus, in Chapter 4 the related research is advanced by investigating the use of such representations for acted and in-the-wild emotional expressions, where the text is transcribed either by humans or by an ASR system.

2. The state of the art shows the effectiveness of using MTL to train machine learning models on multiple corpora with different sets of emotion categories (See Section 1.1.1). On the other hand, deep pre-trained representations have been shown to generalise well across different emotional expressions (See Section 1.1.2). Therefore, this thesis in Chapter 5 proposes a method that uses deep acoustic and textual representations with MTL to predict a latent emotion representation that can recognise the same or similar emotion categories across different datasets.

In the following, the research questions and contributions of this thesis in relation to the aims described above are elaborated in more detail.

1.2.1 On the use of deep acoustic and textual representations

The use of deep pre-trained acoustic and textual representations has been shown to achieve unprecedented state-of-the-art performance (see Section 1.1.2). Despite their success, these models have yet to be fully explored and applied to AER, especially for emotional expressions in the wild. In order to advance the existing state of the art in AER regarding deep pre-trained acoustic and textual representations, the research questions and contributions related to this thesis are as follows:

- **Question 1**: What is the effect of different amounts and types of acoustic signals used to train deep representations, on the performance of such representations for the AER task?
  - **Contribution 1**: An exhaustive evaluation of AER performance on multiple datasets using different pre-trained deep acoustic representations trained with different types and amounts of data (see Section 4.1).
- **Question 2**: How can automatic transcriptions from an existing ASR model be exploited to improve the performance of AER models from speech signals, for both acted and in-the-wild emotional expressions?
  - **Contribution 2**: The exploitation of an ASR model to extract transcriptions for later use in joint acoustic-textual representations, and the evaluation of this method on AER performance for both acted and in-the-wild emotional expressions (see Section 4.3).

- **Question 3**: Given that speaker recognition models can provide us with latent speaker representations, how can we improve the performance of deep acoustic and textual representations for AER, by fusing them with such speaker representations?
  - **Contribution 3**: Investigation of the effect of fusing speaker representations with acoustic, textual, and joint acoustic-textual representations on AER, where the text may be a human transcription or generated by an ASR (see Section 4.4).

1.2.2 Generalisation beyond emotion schemes

Most current AER research focuses on experimenting with specific datasets. However, each dataset covers a specific range of emotional expressions with subjective and inconsistent annotations. To train machine learning models to understand a wide range of emotional expressions, the research points to multi-corpus training through MTL. On the other hand, deep representation learning methods, especially SSL, have been shown to effectively model a wide range of acoustic signals and texts in an unsupervised manner. Therefore, this thesis proposes a method that uses pre-trained self-supervised representations with MTL to predict a latent emotion representation that can generalise beyond specific emotion schemes used for each dataset. The research questions and contributions related to this topic are presented below (the numbering continues from the previous section to count all the research questions in the thesis):

- **Question 4**: How effective is the latent emotion representation, computed by using the MTL-based method using deep representations, in recognising the same or similar emotions across different corpora that might use different emotion annotation schemes?
  - **Contribution 4**: Evaluation of the proposed multi-corpus training method in within-corpus and cross-corpus settings on acted datasets, and by using deep pre-trained acoustic representations (see Section 5.1).
- **Question 5**: By using the proposed multi-corpus training method, can acted emotional expressions be useful in improving in-the-wild AER from acoustic signals and text?
  - **Contribution 5**: Evaluation of the performance of acoustic and textual representations in the proposed MTL-based method for acted and in-the-wild emotional expressions (see Section 5.2).

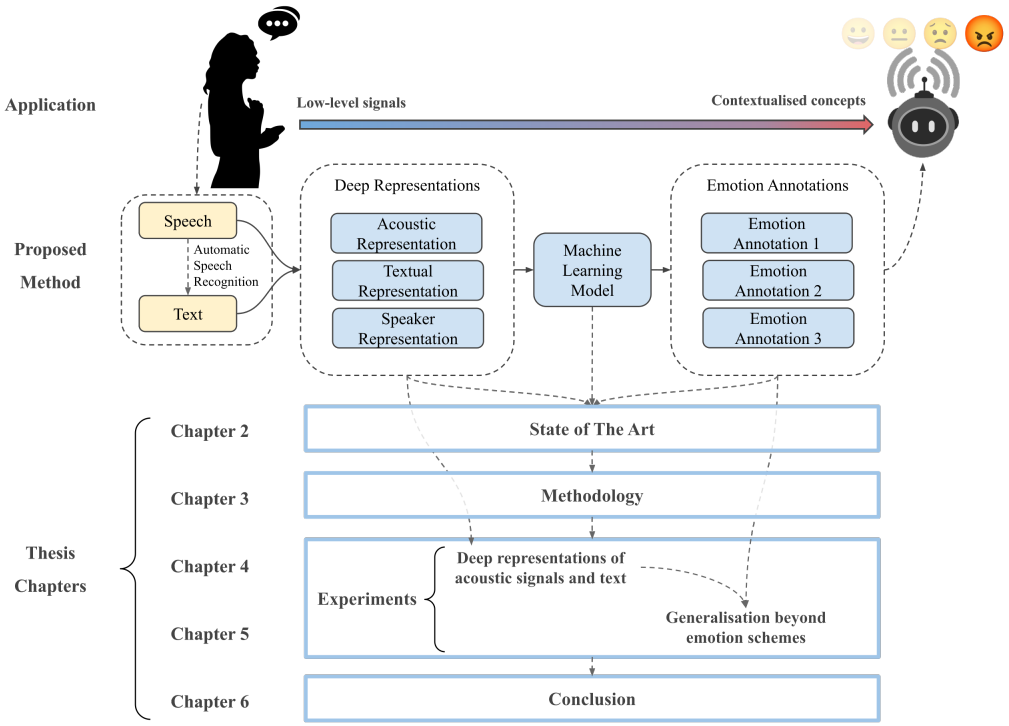


Figure 1.1: Overview of the thesis. The envisioned application of the thesis is shown at the top, where the user can interact with a machine through speech or text, based on which the AER model predicts the user's emotion. Below the application, the proposed method, is shown in more detail, where the input representations are used to train a machine learning model to predict an emotion representation, which is then used to predict a specific emotion annotation. At the bottom, the topics discussed in each chapter are clarified, with dashed lines connecting different parts of the proposed method to the topic of each chapter.

1.3 Thesis overview

The rest of this thesis is divided into four chapters, which are explained below. A visual overview of the structure of the thesis is also shown in Figure 1.1.

Chapter 2 provides the state of the art in AER from acoustic signals and text. The chapter begins by discussing how emotion is viewed in psychology and how it is used to define emotion annotations for machine learning. It then reviews several state-of-the-art AER methods, followed by a case study to quantitatively compare them.

Based on the study of the state of the art, chapter 3 then describes the methodology of the experiments in the following chapters, namely the representations, datasets, training methods, loss functions and metrics used to achieve the goals of the thesis, as well as the technical description of the implementation of the ex-

periments.

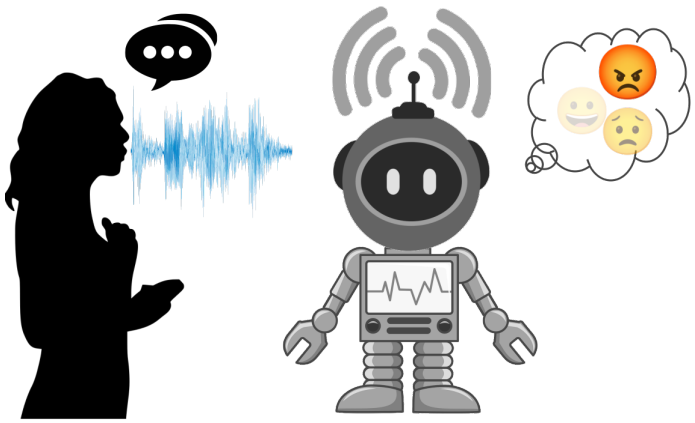
Later in chapter 4, pre-trained deep representations of acoustic signals and text, which have shown the best performances in the literature, are used for AER with both acted and in-the-wild emotional expressions. In addition, this chapter experiments with the use of personalised pre-trained joint acoustic-textual representations, where the text is either human transcriptions or generated by an ASR.

Then, in chapter 5, the proposed method using the pre-trained deep representations with MTL is explored. Also, the effectiveness of using acted emotional expressions for AER with in-the-wild data is experimented.

Finally, chapter 6 concludes the thesis and discusses possible future directions.

Chapter 2

Emotion recognition from acoustic signals and text



This chapter provides a comprehensive overview of the current state of the art in Automatic Emotion Recognition (AER). It begins by discussing the theoretical underpinnings of emotion in psychology and how they serve as a basis for the creation of numerical emotion targets. It then examines the various machine learning techniques for predicting emotion targets. Particular emphasis is placed on Deep Neural Network (DNN)s and exploring the deep acoustic and textual representations that have recently dominated the state of the art in AER. A case study is then conducted to quantitatively compare different existing methods for AER. The chapter concludes with a brief summary of the state of the art and its shortcomings.

2.1 Theoretical background of emotion

This section first briefly introduces the reader to the different views of emotion in psychology. It then explains how some of the psychological models of affect are used to define a numerical emotion space. Then, it is explained how a numerical space of emotion is used to annotate emotional expressions, in order to automatically supervise the training of an AER model (see Section 2.2). The reader is also introduced to the advantages and disadvantages of the different annotation schemes.

2.1.1 Emotion in psychology

The scientific study of emotion can go back to 19th century when emotion was first seen as certain mental states, which cause certain body expressions, to later on, viewing the bodily activity as the cause of emotion and not the result of it (Gendron and Feldman Barrett, 2009). Today, several strands of thought coexist in psychology, ranging from the strand that "emotions are universal" (Ekman et al., 1987) to the strand that "emotion cannot be detected at all" (Barrett, 2017). For example, Ekman argued that people can universally recognise six different emotions –fear, anger, happiness, sadness, disgust and surprise– from facial expressions (Ekman, 1992) (see Section 1.1.1). Many experiments have also shown that people around the world can match such emotion categories to different muscle movements performed by actors, who are not actually feeling those emotions. This has led to the argument that what was objectively detected in these experiments was the facial muscle movements, not the emotion itself, because the actors were not actually feeling the detected emotions. This argument goes on to explain that emotion cannot be objectively detected because it is a concept created by human agreement and does not have an existence outside the mind (Barrett, 2017).

Nevertheless, in recent years, emotion has been seen mostly as a complex human reactionary response to an event. For example, the American Psychological Association (APA) defines emotion as:

A complex reaction pattern, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event[1].

This is in line with the recently popular theory of "appraisal", which argues that humans are hardwired to continuously evaluate stimuli in their environment, and that the interactions between cognitive functions involved in appraisal mechanisms trigger observable expressions of affect. Emotion is therefore seen as an adaptive response to an event, produced as a result of the synchronisation of multiple subsystems, and is considered central to the well-being of the individual (Sander et al., 2005).

[1] https://dictionary.apa.org/emotion

Emotion caused by our constant subjective appraisal of what we see as important events (which depends on the psychobiology and culture of each individual), would lead to certain motor expressions that can be observed in the voice, face and body. An emotional experience can then be verbalised in abstract terms by associating words such as happiness or anger (Scherer, 2009). We can also accurately identify other people's emotional expressions by observing them in social interactions. For example, we can visually distinguish between smiling, crying or frowning, which are associated with different emotions (Gross, 2020). Similarly, when listening to someone, we can detect an expression of intrinsic pleasantness by observing an increase in low-frequency energy, and unpleasantness by detecting more high-frequency energy (Scherer, 2009). The existence of such correlations between measurable sensory inputs and the concept of emotion has inspired researchers to build machines that can automatically recognise emotions. Next, we discuss how different views of emotion in psychology are used to annotate emotional expressions for training data-driven machine learning models.

2.1.2 Emotion annotations

To build machines capable of recognising emotions, we first need to define a numerical space of affect. This numerical space is then used to annotate a recorded expression of emotion, which is later used to train a machine learning model in a supervised manner. However, as discussed earlier, emotion does not have a universally agreed definition and is not standardised for AER (see Section 1.1.1). Therefore, different emotional expressions are annotated rather subjectively, following different theories of affect. Nevertheless, there are two main ways in which affect models are used to annotate emotional expressions in the AER literature: 1) discrete classes of emotion following Ekman's basic emotions –anger, disgust, fear, happiness, sadness and surprise– (Ekman, 1992), and 2) continuous dimensions of emotion such as arousal (or activation) and valence (or intrinsic pleasantness), based on Russell's circumplex model of affect (Russell, 1980). In what follows, how these two main models of affect are used to annotate emotional expressions are explained in more detail.

Annotations based on categorical model of affect

Ekman suggested that emotions should be seen as distinct, independent states of mind, and that distinct categories should be used to define them. His argument was that, from an evolutionary perspective, different emotions have different functionalities and are inherently different (Ekman, 1992), which also means that emotions cannot be described as dimensions (which was contrary to what Russell had proposed). Assuming that emotions are distinct categories for machine learning,

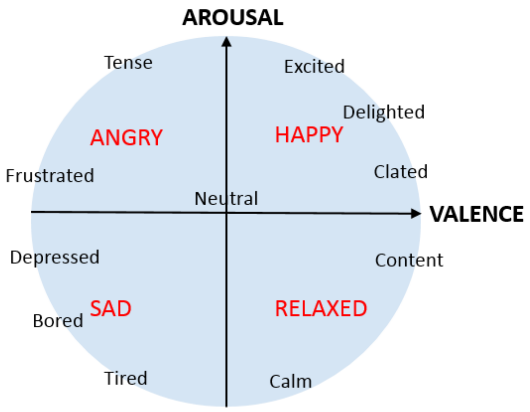


Figure 2.1: Russell's circumplex model of affect. According to this model, the intensity of an emotion (arousal) and its intrinsic pleasantness (valence) form a two-dimensional space for affect. Moreover, different categories of emotion can be identified in this two-dimensional space, depending on the arousal and valence values associated with an emotional expression.
The figure is from Siirtola et al. (2023)

implies that the output of an AER system, has to discriminate between a set of categories (i.e. a classification task). However, the set of categories used for AER systems can vary from one corpus to the other. For example, the CaFE corpus (Gournay et al., 2018) considers the six basic categories of Ekman –fear, anger, happiness, sadness, disgust and surprise–, plus one extra category for "neutral" expressions, whereas the GEMEP corpus (Bänziger et al., 2012) considers 12 "core" emotions –anger, despair, worry, irritation, fear, sadness, amusement, joy, pride, interest, pleasure, and relief–. Categorical annotations are a common practice in the creation of corpora for AER, mainly because they are often reliable, since the annotators are forced to choose from a limited set of categories. For example, Cohen's Kappa coefficients, a method for calculating inter-annotator agreement for categories, is calculated to be over 0.87 for four emotion categories –happy, relaxed, sad and angry– in the DEAP corpus, which is considered an excellent score (Juremi et al., 2017). On the other hand, considering only a limited list of categories for emotion cannot exhaustively represent people's varied emotions (Picard, 2003), as emotional expressions in the wild are much more nuanced than can be explained with only a limited set of categories. This has led many of the researchers in AER to follow Russell's dimensional view of emotion, which considers a much larger target space of affect, rather than focusing on a limited set of emotion categories.

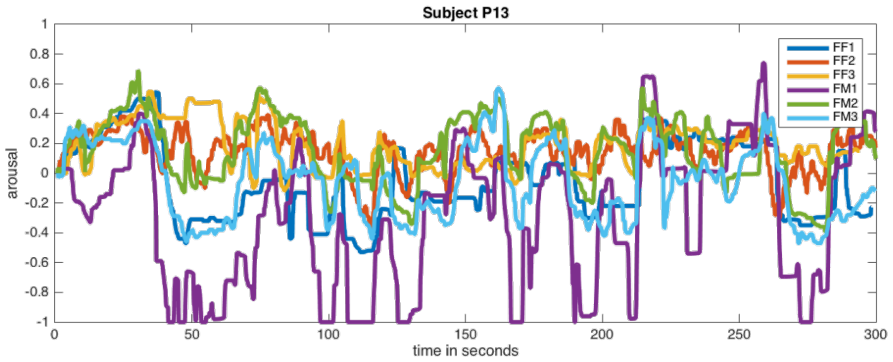


Figure 2.2: An example of continuous annotations of arousal dimension, performed by three French Female (FF1-3) annotators and three French Male (FM1-3) annotators on the audiovisual recording of the subject P13 from the RECOLA dataset (Ringeval et al., 2013). As can be seen in the figure, the continuous annotations do not agree with each other in time, and in value.
The figure is from diuf.unifr.ch/main/diva/recola

Annotations based on dimensional model of affect

Russell's circumplex model of affect, in its simplest form, considers a two-dimensional space for emotion, which are spanned by the basis vectors of arousal (or activation) and valence (or intrinsic pleasantness). According to this theory of affect, arousal refers to the perceived intensity of an event, ranging from calm to very excited. On the other hand, valence refers to the intrinsic pleasantness of a stimulus, ranging from negative to positive (Kensinger and Schacter, 2006; Costanzi et al., 2019). This two-dimensional arousal-valence space of emotion is illustrated in Figure 2.1. As can be seen in this figure, defining a two-dimensional space for affect can encompass different categories of emotion, depending on where an emotion label lands on the two-dimensional space. For example, happiness is associated with high arousal and high valence, whereas anger is associated with high arousal but low valence. Because the two-dimensional space of affect encompasses different categories of emotion, Russell's model provides a more comprehensive view of emotion than the Ekman's categorical model of affect. Hence, in recent years, many studies have started using the dimensional model of affect to annotate emotional expressions for AER (Ringeval et al., 2019; Kossaifi et al., 2021; AlBadawy and Kim, 2018; Khorram et al., 2021). However, there are several practical problems associated with dimensional annotation of emotion, which are discussed below.

The fact that arousal and valence are considered perpendicular to each other in Russell's circumplex model of affect, suggests that arousal and valence are assumed completely independent from each other. It has also been shown that the two dimensions are mostly processed by different parts of the brain (Lewis et al., 2006).

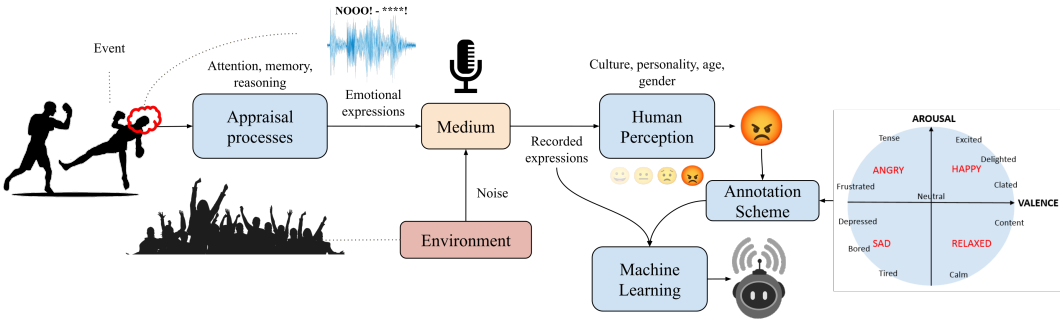


Figure 2.3: A visual example of the conceptual theories of affect involved in building an emotion recognition system. Based on the appraisal theory of affect, an event (here a boxer being punched in the face) first triggers a series of subjective appraisal processes involving attention, memory and reasoning that take place in the boxer's mind, triggering an expression of emotion. A microphone can then be used to record the emotional expression, along with any existing unwanted sounds, such as ambient noise. The recorded expression can then be used for annotation, following a specific annotation scheme. The emotion annotations are then used to train a machine learning model in a supervised manner.

However, multiple studies found out that in practice, when people are annotating emotional expressions according to this theory, there exists a correlation among the two axis of arousal, and valence (Nicolaou et al., 2011; Pantic et al., 2007; Nicolaou et al., 2012), with arousal forming a V-shaped curve as a function of valence (Kuppens et al., 2013, 2017). The reason of this, is not completely known, and might be related to how the annotators perceived the emotional expressions in terms of arousal and valence, rather than how these dimensions are actually processed in the brain.

Moreover, the dimensional annotations of emotion are typically done in a time-continuous format to capture the nuances of emotional experience as emotion may not be static, but rather evolve over time. The time-continuous annotations are often done in a continuous format, resulting in different annotations having different times and values, based on the judgement and time it takes for each annotator to process and act on an observed expression of emotion. An example of continuous dimensional annotations of the RECOLA dataset (Ringeval et al., 2013) is given in Figure 2.2. The continuous traces seen in this figure are registered by the annotators who watch a video (without pausing), and move the computer mouse to the left or right to continuously indicate a low or high value of the arousal or valence dimension. This process inherently introduces noise and delay into the annotations, which would require an additional step of dynamic modelling to mitigate (Khorram et al., 2021; Alisamir et al., 2022b).

Although this problem is not theoretically intrinsic to the dimensional model

of affect, in practice it is often the dimensional annotations that are evaluated with continuous values, rather than the emotion categories. This may be because by choosing the categorical emotion labels, we try to simplify the task and have more reliable labels for AER, whereas with the dimensional model the focus is more on having fine-grained details of emotional expressions.

A visual summary of what has been discussed in this section, from how emotion is viewed in psychology to how emotion is annotated for machine learning, is provided in Figure 2.3. Knowing how to annotate emotions for machine learning, the next section introduces the reader to state-of-the-art machine learning techniques for AER.

2.2 Automatic emotion recognition methods

The previous section discussed how emotion is a complex and contextual human concept. On the other hand, various state-of-the-art modeling techniques have different and often limited capabilities in different modeling tasks. Therefore, the state-of-the-art AER traditionally relies on several stages of data transformation to model the signal at different levels, from acoustic signals or text to emotion. At each stage, we expect an increase in the level of data abstraction, from the raw waveform or text to the emotion labels or dimensions. A visual summary of the most common techniques used at different stages is shown in Figure 2.4. The different models shown in the figure are explained further in this section. It should be noted that while traditional methods were designed to process each stage separately, with the advent of deep neural networks, the boundary between different stages is no longer clear. Nevertheless, in this thesis, the "feature extraction" and "emotion modelling" stages are considered separate in order to follow the traditional AER methodology. Therefore, in the following, traditional feature engineering methods as well as statistical modeling approaches are explored. Artificial Neural Network (ANN)s are then explained in more detail, followed by deep acoustic and textual representations, which have achieved significantly better results than traditional features. As this thesis investigates the personalised deep acoustic-textual representations (see Section 1.2.1), the state-of-the-art methods regarding both joint acoustic-textual representations and personalised representations are also discussed. Moreover, this thesis also investigates MTL for generalisation beyond emotion annotation schemes (see Section 1.2.2). Therefore, in Section 2.2.7, the work related to MTL for AER is further discussed. Finally, a quantitative performance comparison of the discussed methods is performed through a case study on the RECOLA dataset.

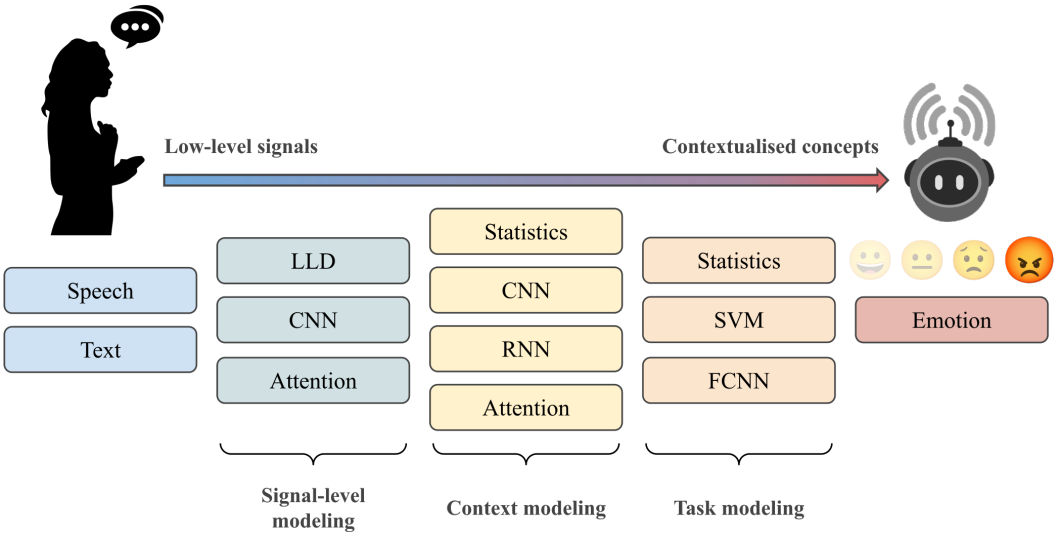


Figure 2.4: The modelling stages involved in emotion recognition from acoustic speech signals and text, based on current state-of-the-art methods. LLD: low level descriptors (expert-knowledge feature extractors like speech spectrograms); CNN: convolutional neural network; Attention: attention based neural network; Statistics: statistical summarisation approaches; RNN: recurrent neural network; FCNN: fully connected neural network; SVM: support vector machine.

2.2.1 Traditional feature extraction methods

In this thesis, the representations of both acoustic signals and text are investigated (see Section 1.2.1). Since this subsection discusses traditional feature extraction methods for both modalities, the reader is first introduced to traditional feature extraction methods for acoustic signals, and then traditional textual representations are covered.

Traditional acoustic representations

Over the last few decades, scientists studying the acoustic properties of speech have sought to create more descriptive representations of the acoustics that could be more easily exploited for various speech-related tasks, such as AER. Subsequently, several feature extraction techniques based on the scientific study of how speech is phonated, articulated and perceived were proposed, which are still widely used today. An example of such feature extraction techniques is Mel-scale Filter Bank (MFB), which is the result of expert knowledge design based on the human auditory system. MFBs exploit the short-term energy coefficients of the signal's frequency, similar to the perceptual characteristics of the human ear. Another famous feature extraction techniques is Mel-Frequency Cepstral Coefficients (MFCC), which

is based on MFB, with the additional step of applying a deconvolution to preserve the variability of the original speech signal. It is worth noting that MFB-based feature extraction techniques are still one of the most common choices as the input of state-of-the-art AER from acoustic signals (Latif et al., 2020) (cf. section 2.2.8). However, MFBs are the result of studying the acoustic properties of speech in general and without considering its relation to emotion.

As was discussed earlier in section 2.1.1, there are correlations between measurable sensory inputs, such as acoustic signals and different emotions. For example, an increase in low-frequency energy, is often associated with pleasantness, while more high-frequency energy is usually attributed to unpleasantness. The result of studying such correlations was Geneva Minimalistic Acoustic Parameter Set (GeMAPS), which contains a reduced set of acoustic and prosodic features, chosen to be more informative about affective physiological changes in speech production (Eyben et al., 2015).

The traditional methods mentioned above, such as MFB and GeMAPS, are considered to be rather Low Level Descriptors (LLD)s of acoustic signals, as they only describe speech at the signal level. Moreover, LLDs usually process an acoustic signal with frames that range from 20ms to 40ms, where the statistical properties of the signal are considered to be stationary. On the other hand, emotion is a complex and contextual human concept that depends on a speaker in a specific environment, and with highly variable durations ranging from a few seconds to several hours (Brans and Verduyn, 2014), and is thus difficult to model using only LLDs of acoustic signals. Therefore, in order to more accurately predict different emotions, LLDs are usually accompanied by statistical or machine learning models to take into account different lengths of emotional expressions, different environments, and different speakers (see Figure 2.4).

One of the most common statistical modeling of LLDs involves clustering approaches such as Bag of Audio Words (BoAW), Gaussian Mixture Model (GMM)s, and Fisher Vector (FV)s. For example, BoAW is a method that can cluster different audio features into a dictionary, whose distributions are then used to extract more contextual features, and has been shown to predict emotion dimensions better than LLDs (Schmitt et al., 2016; Ringeval et al., 2018a). Similarly, GMMs, which can cluster data into different Gaussian distributions with unknown parameters, can also be used as a statistical summary of LLDs. GMMs are clustering methods that can incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians. In addition, FV can be used to predict the likelihood of a probabilistic model, such as GMM, to provide contextual features for AER (Gosztolya, 2020).

So far, only the traditional methods of processing acoustic signals have been discussed. However, one of the aims of this thesis is also to investigate textual representations (see Section 1.2.1). Therefore, the following is a brief summary of

the traditional methods of representing text for machine learning purposes.

Traditional textual representations

The first step in processing a given text is to segment it into subunits, called tokens, which are encoded numerically so that a machine can process them. The subunits of a text, can be words, sub-words, or characters, while keeping the order of them. Then, depending on the task and the language, different pre-processes may be involved, such as removing punctuation, removing frequent words, or replacing each word by its root (i.e. stemming). Then, different methods can be used to compute a series of feature vectors from the pre-processed tokens. One popular approach is Vector Space Model (VSM), which represents a given sub-unit of text (a symbol) by a point in an N-dimensional space (a numerical vector) (Liang et al., 2017). The following is a discussion of various traditional approaches to numerical representation of text.

The simplest approach to numerically represent text, is to have a dictionary of all the tokens, and assign each token with a unique number. For example, we can assign a numerical id to each word used in a language. After assigning a unique number to each token, usually a one-hot vector is created. For instance, the one-hot vector of each word has the length of the vocabulary, and the value of one at the index corresponding to the position of the word in the vocabulary, and all other elements are zero. Although this approach results in numerically representing any given text, it would not encode any syntactic or semantic information of the text, to help solve a specific task. In order to include syntactic or semantic information, traditional methods for extracting features from text usually apply filtering, statistical, and mapping techniques (Liang et al., 2017). For example, one can simply filter out the words that are not repeated often in a text, in order to reduce the size of the vocabulary, and thus the dimension of the one-hot vectors (Singh et al., 2013). One can also compute the mutual information between a numerical word representation and its label for a classification task (Paninski, 2003), in order to select only the representations with high information gain for a target task. Different statistical modeling approaches have also been investigated in order to gain more contextual information, by putting similar textual subunits closer to each other in the representation space (Huang, 2008; Aggarwal and Zhai, 2012). A famous example of a statistical measure used for text processing is the Term Frequency-Inverse Document Frequency (TF-IDF), which encodes the statistical saliency of each word, by reflecting how often a word is used in a document –Term Frequency (TF)– and how rare it is in a given set of documents –Inverse Document Frequency (IDF)–. The mentioned filtering and statistical approaches still do not encode any semantic information of a given text. On the other hand, mapping methods involve approaches that map a high dimensional one-hot vector into a lower dimensional latent semantic space (Liang et al., 2017). One example of such approach is Latent Semantic

Analysis (LSA) (Evangelopoulos, 2013), which analyses the relationships of different terms within different documents to compute the existing similarity among the terms and the documents. This approach has also proved to be effective in classifying different emotions from text (Wang and Zheng, 2013).

As discussed in this section, statistical modelling methods have been effectively applied to AER on top of traditional acoustic and textual representations. These methods are often accompanied by machine learning models, such as Support Vector Machine (SVM)[1] (Schmitt et al., 2016; Ringeval et al., 2018a) and later with DNNs (Jianqiang et al., 2018; Ezz-Eldin et al., 2021) to achieve high performances in AER (see Section 2.2.2). In fact, the fast growing advances made for DNNs, has completely changed the way of processing acoustic signals, as well as, textual data, by improving the performance in affective computing by a large margin (Evain et al., 2021b; Latif et al., 2020; Alisamir and Ringeval, 2021; Siriwardhana et al., 2020). The quick rise of DNNs to state-of-the-art results in many domains, including affective computing, has not stopped at effectively mapping acoustic or textual features to emotion (Trigeorgis et al., 2016). But it has continued to completely replace all the steps involved in a traditional AER methodology, making traditional feature extraction techniques and statistical modelling increasingly rare. In what follows, DNNs are explained in more detail.

2.2.2 Deep neural networks

ANN is a type of machine learning technique, which is loosely based on the concept of biological neural networks in the human brain. Each artificial neuron, similar to the synapses and axons of a biological neuron, can be connected to other neurons to send or receive information. Artificial neurons are usually put together as groups, which are called neural layers. Multiple layers of ANN can then be cascaded together in different ways to model more complex tasks, in which case they are referred to as DNNs (Liu et al., 2017). This fact, in theory, gives DNNs universal approximator abilities, which means that with enough layers, DNNs can represent any function with high precision (Goodfellow et al., 2016). DNN can do this because they consist of multiple layers, and each layer is able to transform its input to a higher level of abstraction, which is more important for prediction or discriminative tasks (LeCun et al., 2015). The higher level of abstraction achieved by the layers further from the input is also considered to be less sensitive to local changes in the input (Bengio et al., 2013). The following is a brief description of the most

[1] SVM is a machine learning model that learns to draw a hyperplane to separate the data into different classes. Unlike DNNs, which are randomly initialised and trained to find a local minimum solution, SVMs use a convex optimisation technique that ensures a unique global minimum solution. However, this also means that, unlike DNNs, SVMs are not well suited to training on large amounts of data, as all training points need to be stored and computed together.

common layers used in neural networks today in most domains, including affective computing.

Fully connected feed-forward layers

The most basic form of artificial neural layers are fully connected feed-forward layers, where all neurons in the first layer are connected to all neurons in the next layer (sometimes referred to as dense or linear layers). To describe how fully connected layers work through mathematical notations, we can consider the input of each layer to be a numerical vector, that is transformed to a different vector, through a matrix multiplication, and usually followed by a non-linear function. This process can be written as followed:

$$y = h(Wx + b) \tag{2.1}$$

where x is the input vector, W is the weight matrix, b is the "bias" vector, which is there to off-set the linear matrix multiplication, and $h(.)$ is usually a non-linear function such as tangent hyperbolic or sigmoid, and y is the output vector. The weight, and bias elements here are trainable through the backpropagation process, by different algorithms, such as Stochastic Gradient Descent (SGD) (Ruder, 2016). In what follows, the training of an ANN is explained in more details.

Training neural networks

Backpropagation with SGD-based algorithms first calculates the gradient of a given loss function with respect to each of the weights in a neural network. It then uses the chain rule of calculus to iteratively compute the gradients of each layer in the network. The gradients can then be used to update the weights of the network to minimise a given loss function. This process for each layer can be described as follows:

$$W = W - \eta \frac{\partial \mathcal{L}}{\partial W} \tag{2.2}$$

where $\mathcal{L}$ is the loss function, $\frac{\partial \mathcal{L}}{\partial W}$ is the derivative of the loss with respect to a weight matrix W, and η is the learning rate, which lets us control how fast or slow the weights are updated for each iteration. Through back-propagation, a fully connected layer can theoretically be iteratively trained to recognise the spatial structure of data. In practice, however, training fully connected layers can suffer from the curse of dimensionality because they inherently have a large number of trainable parameters. The curse of dimensionality occurs when the volume of space increases relative to the amount of data available, causing the data to become sparse in that space. This makes it more difficult for backpropagation to train a fully connected layer, as there would not be enough data in the search space from which to

approximate a function. This has led to the introduction of convolutional layers, which suffer less from the curse of dimensionality, because they limit the number of trainable parameters by orders of magnitude compared to fully connected neural networks through sharing a sliding weight (Abend, 2022). The convolutional layers are described in more details below.

Convolutional layers

Convolutional layers are another type of neural layers that are commonly used in DNN architectures. Unlike fully connected layers (see above), where the value of each input neuron has an independent weight to be transformed to its output, the input neurons in a convolutional layer share a sliding set of weights. This mimics the mathematical operation of convolution, and to calculate the output value of each neuron n, one can write as follows:

$$y[n] = \sum_{k=0}^{N} w \times x[n-k] \tag{2.3}$$

where x and y are the input and output vectors (i.e. layers), N is the total number of elements (i.e. neurons) in the input vector x (i.e. input layer), and w is a shared *trainable* weight vector, which has the same size as the input vector $x[n-k]$. The formula above is only for one trainable weight vector w, however, it is common practice to use multiple set of weights to better model the structure of data. In this way, convolutional layers can model the temporal structure of an acoustic signal by sharing the parameters across the temporal dimension. This process is similar to traditional filters used in signal processing, which is why convolutional weights are often referred to as filters. However, unlike traditional filters, which are designed by experts, convolutional filters are designed by learning from the data (Palaz et al., 2015). In addition, convolutional filters can be combined with fully connected and recurrent layers (see below) to perform a specific task, such as AER or ASR from acoustic signals, using only raw speech. This technique, also known as end-to-end learning, has been shown to outperform traditional feature extraction techniques using the same fully connected and recurrent layers for continuous prediction of arousal and valence dimensions (Trigeorgis et al., 2016) (see Section 2.2.3).

The fully connected and convolutional layers are among the most commonly used algorithms and are known to be particularly effective in computer vision tasks due to their ability to detect different features in images. However, for sequential tasks such as natural language or speech processing, this is often not enough, as sequential data depends on the relationship of each part of the sequence to the next. This has led to the idea of recurrent layers, where the neurons are assumed to be in a sequence. Recurrent layers are described in more detail below.

Recurrent layers

Recurrent layers are neural layers that are known for their ability to process sequences of data. This is possible because each neuron in an output vector of a recurrent layer influences the computation of subsequent neurons in the same output vector. In its simplest form, recurrent layers can be described mathematically as follows (Goodfellow et al., 2016):

$$\begin{aligned} a_n &= W_1 y_{n-1} + W_2 x_n + b_1 \\ y_n &= h(a_n) \end{aligned} \tag{2.4}$$

where x_n, and y_n are input and output vectors at step n. W_1 and W_2 are trainable weight matrices. b_1 and b_2 are trainable bias vectors, and h(.) is the activation function, which is usually considered to be a non linear function such as tangent hyperbolic. Also, a_n is an auxiliary notation to better understand the equation. Note that in the first step ($n = 0$), y_{n-1} cannot be calculated from x_{n-1} (as x_{-1} does not exist), and is usually initialised randomly, or to all zeros.

Recurrent layers differ from other types of neural layers in that they are considered to have a "memory" component, as they use previous inputs to inform the current output (see equation above). This makes recurrent layers in theory effective methods for modeling long-term dependencies of sequential data, such as audio, video and text. In practice, however, training the basic recurrent layers with backpropagation would run into the problem of either vanishing or exploding gradients. The vanishing gradients problem in recurrent layers means that the gradient information becomes too small, when they are propagated from the end of a sequence to its beginning. On the other hand, exploding gradients problem happens when instead of decaying, gradient information grow exponentially during backpropagation. To alleviate these problems, special recurrent layer architectures have been introduced, namely Long Short Term Memory (LSTM) cells (Hochreiter and Schmidhuber, 1997), and more recently a simplified version called Gated Recurrent Unit (GRU) (Cho et al., 2014).

LSTMs and GRUs are both recurrent layers that are capable of capturing complex, long-term dependencies in a sequence. They can solve the vanishing and exploding gradients problem of recurrent layers, by using a series of "gates" to regulate the flow of information into and out of each cell. The gates act as a kind of filter, allowing the network to retain only relevant information. This helps the network to focus on the most important data and ignore unnecessary details related to each task, thereby preventing the gradients from vanishing or exploding. This has made LSTMs and GRUs popular choices for DNN models for emotion recognition, especially from acoustic signals (He et al., 2015; Chao et al., 2015; Weninger et al., 2016; Trigeorgis et al., 2016; Le et al., 2017; Evain et al., 2021b; Alisamir et al., 2022c) .

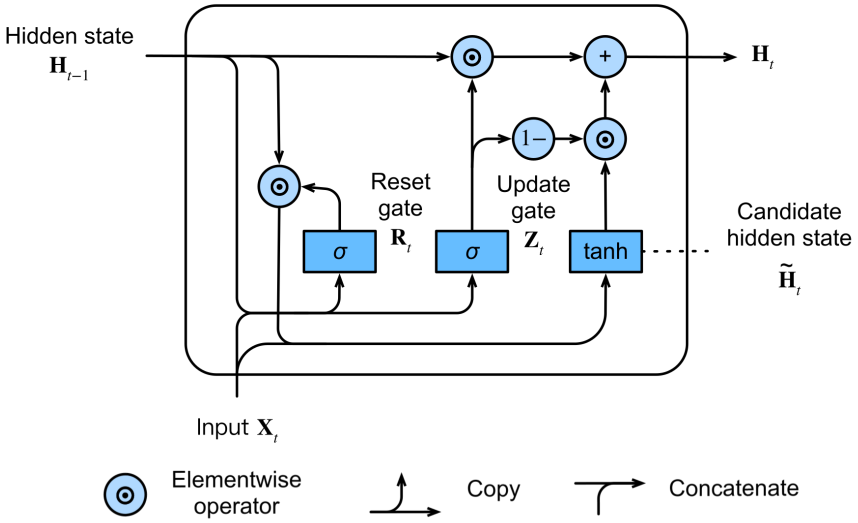


Figure 2.5: A visual representation of a Gated Recurrent Unit (GRU). This figure depicts a middle step of a GRU in computing the hidden state H_t, given the previous hidden state H_{t-1}. Also, σ and *tanh* refer to the mathematical functions of sigmoid, and tangent hyperbolic respectively.
The figure is from d2l.ai

While both LSTMs and GRUs share the same basic architecture, they differ slightly in the number of gates they use. While LSTMs use an input, an output and a forget gate to store and discard information, GRUs only use an update gate and a reset gate. This makes GRUs a simplified version of LSTMs, allowing them to be trained slightly faster, easier and sometimes with slightly better performance than LSTMs (see Section 2.2.8). Thus, many of the experiments in this thesis are performed with GRUs, and in what follows, the inner workings of them are explained in more detail.

Figure 2.5 shows the architecture of a GRU, which calculates the following:

$$\begin{aligned} r_t &= \sigma(W_{ir}x_t + b_{ir} + W_{hr}h_{t-1} + b_{hr}) \\ z_t &= \sigma(W_{iz}x_t + b_{iz} + W_{hz}h_{t-1} + b_{hz}) \\ \tilde{h}_t &= Tanh(W_{in}x_t + b_{in} + r_t \odot (W_{hn}h_{t-1} + b_{hn})) \\ h_t &= (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{(t-1)} \end{aligned} \tag{2.5}$$

where x_t is the input at time t, h_t and $h_{(t-1)}$ are the hidden states at times t and $t-1$. r_t, z_t are the reset, and update gates respectively, and $\tilde{h}_t$ computes the candidate hidden state. Various W and b tensors are trainable weights and biases similar to what was defined for a linear layer (see "Fully connected feed-forward layers" in this section). σ represents the sigmoid function and $\odot$ is used for the Hadamard

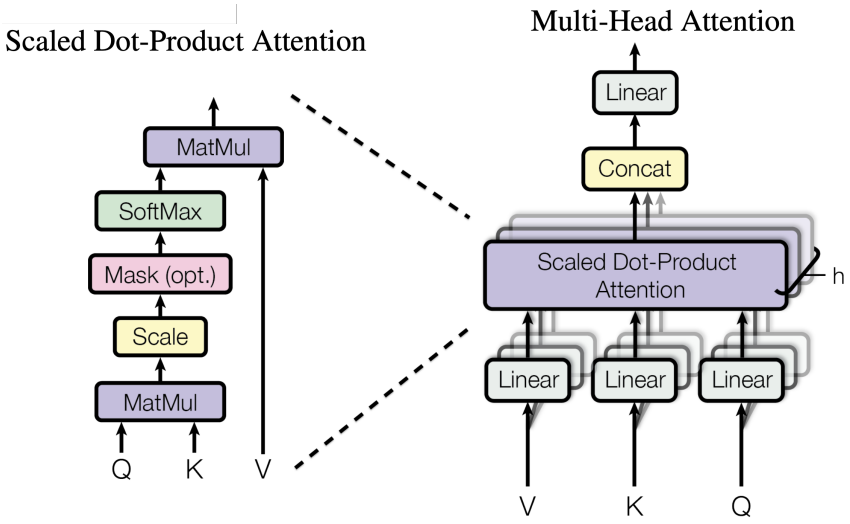


Figure 2.6: Depiction of Multi-Head Attention (MHA), and scaled dot-product attention. In the figure above, Q, K, and V refer to query, key, and value in attention mechanism. Also, "MatMul" calculates dot-product between two different vectors, and "Linear" refers to a fully connect layer. Moreover, "Masking" is related to encoder-decoder architectures for training a DNN in a self-supervised manner (see Section 2.2.3 and section 2.2.4 to know more about self-supervised learning for acoustic and textual representation respectively).

The figure is from Vaswani et al. (2017a)

product [1].

Although LSTMs and GRUs have proved capable of modelling the context of sequential data, they are time-consuming to compute, because each value in a sequence must be computed in turn. Moreover, in practice many different neural layers use parallel computing to run faster on modern computers, which is not possible for recurrent layers, due to their inherent sequential nature (Vaswani et al., 2017a). In addition, the recurrence and non-linearity used in recurrent layers makes them difficult to interpret. Although these problems do not directly affect the performance of these models, they have partly led many studies to use the more recent attention mechanisms to model long-term dependencies in sequential data instead of recurrent layers.

Attention mechanism

The attention mechanism can be thought of as a machine learning method that can focus on the most important parts of the input. This is achieved by learning the weight of the input components and assigning greater importance to certain parts of

[1] https://pytorch.org/docs/stable/generated/torch.nn.GRU.html

the input. This allows the attention mechanism to focus on the relevant information and ignore the irrelevant parts, which can significantly improve the accuracy of sequential tasks (Vaswani et al., 2017a). It has also shown capable of achieving state-of-the-art results in affective computing (Vazquez-Rodriguez, 2021; Wagner et al., 2022). Knowing the effectiveness of the attention mechanism in the state of the art, what follows is a more detailed explanation of how it works.

Although the attention mechanism can be defined in different ways (Chaudhari et al., 2021), it is usually implemented as dot-product attention following the work of Vaswani et al. (2017a), and it can be mathematically written as follows:

$$A(Q,K) = QK^T \tag{2.6}$$

where A is the attention vector, calculated by the dot-product of Q and K vectors. Traditionally, the attention concept comes from retrieval systems, where a query (Q) is first compared to a set of keys (K), which are associated with a set of values (V). This way, the attention mechanism can map a query and a set of key-value pairs to an output, which can be computed as a weighted sum of the values. In order to compute the weighted sum of the values, a softmax function can be used on the attention vector A, in order to make the values describe a probability distribution. Furthermore, it is often assumed that Q and K vectors have d_k dimensions, and that they have a random distribution with zero mean and d_k variance. Thus, a common practice to divide them by d_k, in order for them to have zero mean and unit variance (this is known as "scaling"). Thus, a scaled dot-product attention can be calculated as follows:

$$A(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.7}$$

Moreover, in order to allow the attention mechanism to learn more complex representations of the data, multiple scaled dot-product attentions can also be used. This approach is called Multi-Head Attention (MHA), and is depicted in Figure 2.6.

Furthermore, to better model sequential data, Vaswani et al. (2017b) proposes the idea of positional encoding, which encodes the position of each item in a sequence. This is a necessary step for modelling sequential data when using only the attention mechanism, as this mechanism does not inherently consider the order of a sequence. The positional encoding may be defined as follows:

$$\begin{aligned} PE_{(pos,2i)} &= sin(\frac{pos}{10000^{2i/d_{seq}}}) \\ PE_{(pos,2i+1)} &= cos(\frac{pos}{10000^{2i/d_{seq}}}) \end{aligned} \tag{2.8}$$

where pos is the position, i is the dimension, and d_{seq} is the dimension of the input sequence. The idea is to encode the position of each element in a sequence on a circle to keep the numerical values of the encoding between -1 and 1. This idea

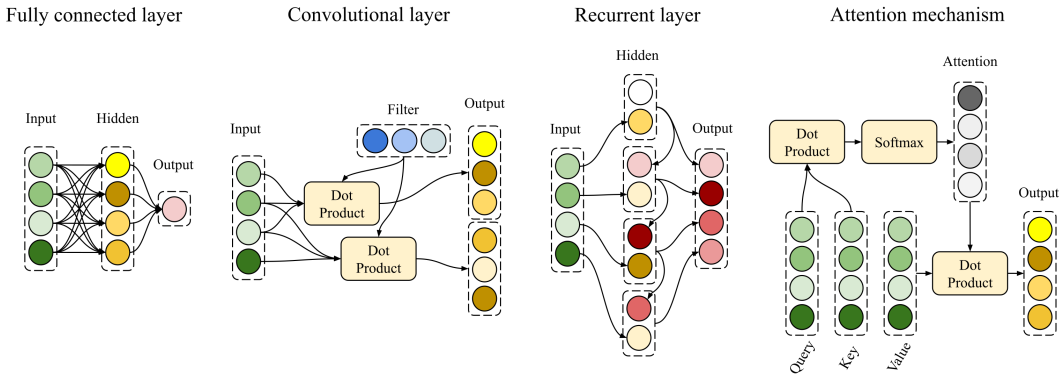


Figure 2.7: A simplified depiction of the most common neural layers used in most domains today, including affective computing.

helps in practice to have a better positional encoding than just assigning incremental integers. This is specially important as the positional encodings would get summed with the input sequence[1]. Also, the value 10000 in the formula is chosen rather arbitrarily, and it can be any high value to ensure the uniqueness of the embeddings.

This section gave a brief introduction to DNNs. The focus was on the neural layers and methods used in the experiments within this thesis. Furthermore, a simplified visual summary of the most common neural layers is shown in Figure 2.7. In recent years, DNNs composed of the aforementioned neural layers have dominated the state of the art in the field of AER from acoustic signals and text. This is achieved in particular by deep representation learning techniques used for acoustic signals and text using only unlabelled data. The following section explains how deep representations of acoustic signals work and how they have come to prominence.

2.2.3 Deep acoustic representations

Over the last decades, Low Level Descriptors (LLDs) of the acoustic signal, built by studying the acoustic properties of speech, have been used exhaustively for AER (see Section 2.2.1). However, the usage of the LLDs as the first step in acoustic processing, is fading away year by year, being replaced by deep representations. Although "hand-crafted" LLDs are more explainable because they are designed according to the human view of speech perception, it also makes them limited to human knowledge. On the other hand, deep representations learn to adapt to an objective function, optimising for the best representation that fits a given data distribution. This has led to deep representations being much more effective than traditional

[1]The choice of summation over the more intuitive concatenation seems to be avoiding using extra trainable parameters.

"hand-crafted" feature sets, in many tasks, including AER from acoustic signals (Ringeval et al., 2018b; Latif et al., 2020; Evain et al., 2021a). To train such effective deep representations for AER, there are different methods, among which we can name supervised learning –namely, end-to-end and transfer learning– and unsupervised learning –namely, (traditional) auto-encoders, variational and adversarial auto-encoders, and self-supervised learning–. In the following, the aforementioned representation learning techniques are explained in more detail.

End-to-end learning

End-to-end learning refers to the type of deep learning method that learns different representations of data at different levels of abstraction, from input to output. This is achieved in particular through the use of convolutional layers, which can be trained on a given set of data to model its structure. It has also been shown that acoustic representations learnt in an end-to-end manner, can achieve better performance than using hand-crafted representations to predict arousal and valence dimensions of emotion (Trigeorgis et al., 2016) (see Section 2.2.2). Interestingly, analysis of the convolutional layers trained in an end-to-end manner, has shown that some layers learn a smooth spectral envelope in the average frequency response over time, which is similar to traditional features such as MFCCs (Palaz et al., 2015). In addition, the data-driven convolutional filters have been shown to learn fundamental frequency correlates of emotion, as well as other related prosodic features that were not previously apparent to researchers studying the acoustic properties associated with emotion (Bertero and Fung, 2017). This is largely due to the properties of DNNs, which allow multiple convolutional layers to be cascaded and trained together on large amounts of data, resulting in complex and contextualised representations suitable for specific tasks.

End-to-end learning has enabled the feature extraction and task modelling stages to be merged for the first time in the history of speech processing and other related domains. However, this has come at the cost of requiring more and more labelled data, in order to increase the ability of the deep representations trained in an end-to-end manner to generalise well across different contexts. This problem is compounded for AER, since emotion annotations are also subjective and expensive to collect (see Section 2.1.2). Moreover, a labeled dataset gathered for a specific tasks, represents only a specific distribution of emotional expressions (Tagliasacchi et al., 2019). As a result, the performance of end-to-end trained convolutional layers may not be as good for data in the wild, where there is a wide variety of emotional expressions. For example, Deschamps-Berger et al. (2021) shows that using acted emotional expressions to train convolutional layers in an end-to-end manner has limited application for using such deep convolutional representations for AER in the wild.

Therefore, representations based on end-to-end learning have limited generali-

sation capabilities for tasks with small amounts of labelled data, such as AER. This has led to the exploration of deep convolutional layers trained for other similar tasks where large amounts of labelled data are available (e.g. ASR). This paradigm in deep learning is usually called transfer learning, because we would like to transfer the data structure captured by deep representations for one task to another. Transfer learning and its applications in AER are explained in more detail below.

Transfer learning

Transfer learning usually refers to a machine learning paradigm that allows the knowledge gained from solving one task to be applied to a different but related task. For the task of AER from acoustic signals, transfer learning is usually related to first training the convolutional filters to model the acoustic signals for a task related to AER, then evaluating the effectiveness of such filters for the task of AER. For example, Tits et al. (2018) shows that convolutional layers trained on ASR can learn more effective representations of acoustic signals for predicting arousal and valence dimensions than "hand-crafted" GeMAPS feature sets. Moreover, since the spectrogram of an acoustic signal can be represented as an image, the convolutional filters trained for snore sound classification from spectrograms, or even for classification of different objects from images, have been successfully used for AER, outperforming "hand-crafted" features (Amiriparian et al., 2017; Ringeval et al., 2018b, 2019). Describing acoustic signals as the probability of classifying different objects in an image may seem absurd, but the regularity of certain patterns in a spectrogram can be similar to the regularity of different objects in images. Nevertheless, research has shown that representations learned for a similar task with a similar distribution are more effective for transfer learning (Zhang et al., 2017a; Triantafyllopoulos and Schuller, 2021). This point brings us to the caveats of transfer learning, which is that it is not always easy to predict how much of knowledge gained in a task, such as object classification in computer vision, can be transferred to recognising emotion from the acoustic signals. Furthermore, acoustic signal representations learnt for similar tasks to AER, such as ASR, may not be able to detect specific patterns of emotional expressions. This is because ASR's objective is to transcribe speech, and thus representations trained for ASR have learnt to abstract acoustic signals in a way to only contain verbal information. However, abstracting only verbal information is not sufficient for AER, as emotional expressions are conveyed through both verbal and non-verbal communication (see 1.1.2).

On the other hand, training deep representations in an end-to-end learning fashion for AER is not straightforward due to the lack of diverse emotionally labelled data. And as discussed earlier, data-driven deep learning models require large amounts of diverse data to generalise well across different contexts (see Section 1.1.1). On the other hand, today there is an abundance of unlabelled recordings of emotional expressions from a variety of speakers available online and under Cre-

ative Commons licence, which can be used to abstract representations of acoustic signals in an unsupervised manner. Thus, in recent years, the focus of research has shifted dramatically from training models with supervised approaches to exploiting unsupervised representation learning techniques. These techniques are described below.

Unsupervised learning

Unsupervised learning refers to a type of machine learning strategy that can train a deep learning model without the need for labelled data. Instead of using labelled data, it uses various techniques to find patterns and insights from unlabelled data. The goal is usually to provide a deep representation of the data by capturing its structure through different methods. Below, the reader is first introduced to auto-encoders, which form the basis of many of the unsupervised learning methods used today. It then explores a line of research that focuses on the use of both auto-encoders and end-to-end learning (i.e. semi-supervised learning) to provide deep acoustic representations. The reader is then introduced to the latest state-of-the-art unsupervised learning approaches, such as adversarial and variational auto-encoders, as well as self-supervised representation learning.

Auto-encoders

Auto-encoders are one of the most famous ANN architectures for unsupervised learning of representations. Auto-encoders used for acoustic signals, first encode the acoustics into an abstract representation, and then decode the representation back to the original acoustic signal. The encoder and the decoder are separated neural networks that are trained together in a tandem to reconstruct an input signal, while learning to map such signal to an intermediate representation, which densifies the information relevant for reproducing the signal, by the decoder (see Figure 2.8). As a result, huge amounts of unlabelled data can be used to obtain a dense abstract representation of a wide range of acoustic signals, which would then require less complex models to predict emotion labels than using traditional features. Moreover, the fact that auto-encoders can be trained on a data distribution without the need for labels, means that one can train effective deep representations for a target data distribution. For example, Deng et al. (2014) shows that auto-encoders can adapt to a target domain in this way, and achieve good performance in cross-domain AER with acoustic signals. Their method involves training auto-encoders with feed-forward layers on statistical features of acoustic signals. However, with the advent of recurrent layers and then attention mechanism, various studies have been carried out to train more contextual auto-encoders for AER from acoustic signals.

As recurrent layers are particularly effective in modelling the context of acoustic signals, they have been investigated in an auto-encoder paradigm and have been

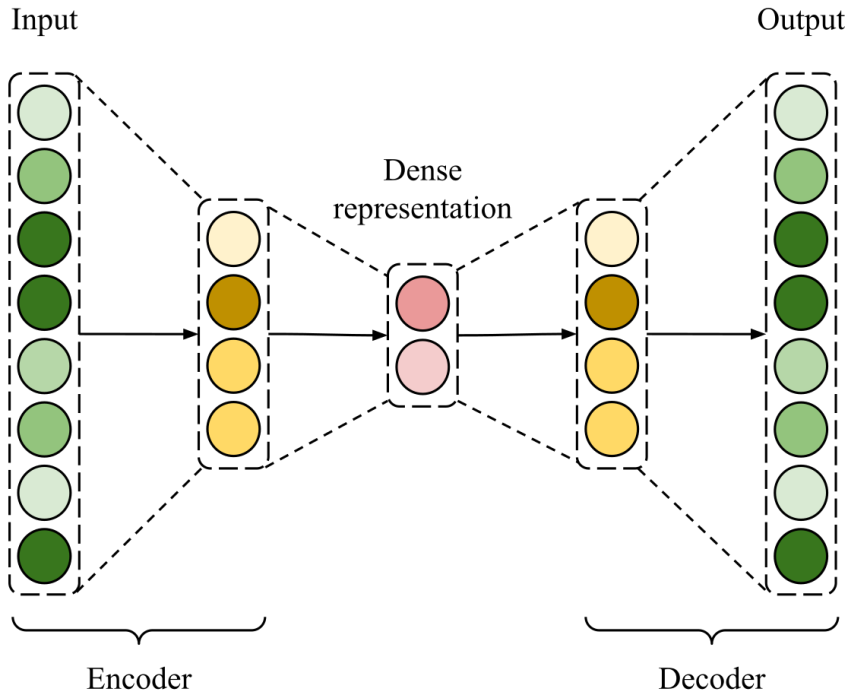


Figure 2.8: A simplified depiction of an auto-encoder neural network. The encoder and the decoder are trained together to reproduce the input in the output. This process forms a dense representation as the output of the encoder, which can then be used to model emotional expressions from different modalities, such as text or acoustic signals.

shown to produce good results both within and across corpus AER (Neumann and Vu, 2019). In another study, the use of convolutional layers first to model low-level changes in the signal, followed by recurrent layers to model the context, has shown further improvements for AER in cross-corpus settings (Dissanayake et al., 2020). In addition, attention-based auto-encoders have also been investigated and have shown good performance for speech translation, sound event detection and AER from speech (Zhang et al., 2020b). However, studies in this area are rather limited, and no fair comparison of different methods using attention and recurrent layers for auto-encoders used in AER has been found.

As there is no supervision involved in the training of auto-encoders, the representations are not learned to be effective in modelling a specific task. This has led to some studies that first use auto-encoders for large amounts of unlabelled data, and then try to make them more effective for a specific task with supervised end-to-end learning (i.e. semi-supervised learning). For example, Huang et al. (2014) proposed a method where they first trained dense representations with auto-encoders on large amounts of unlabelled data, and then extracted the salient features for AER from

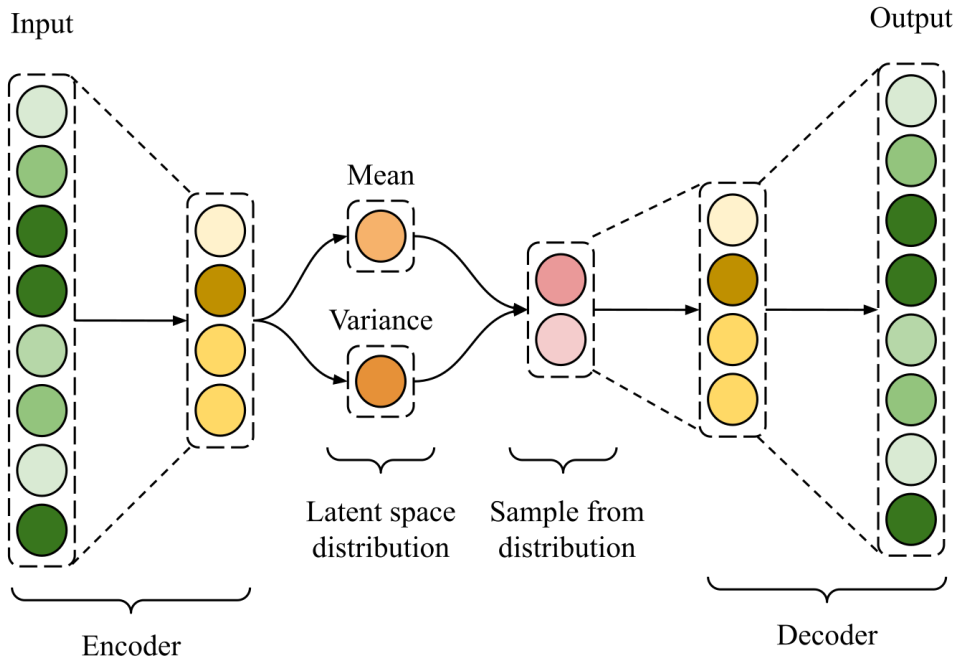


Figure 2.9: A simplified depiction of a VAE. Similar to traditional auto-encoders, the encoder and the decoder of VAEs are trained together to reproduce the input in the output. However, unlike traditional auto-encoders, VAEs add a regularisation term to ensure the latent representation follows a probabilistic distribution, which in most works is a Gaussian distribution.

speech in an end-to-end manner. Using auto-encoders as an auxiliary task to reconstruct intermediate acoustic representations while training an end-to-end AER model has also proven effective in learning good acoustic representations for AER (Parthasarathy and Busso, 2018; Deng et al., 2017).

The semi-supervised learning methods mentioned above were aimed at modelling the data distributions more efficiently than traditional auto-encoders by also shaping the representations through supervised training. However, semi-supervised learning still requires task-specific annotations. On the other hand, in recent years, another type of auto-encoder has been introduced, called Variational Auto-Encoder (VAE), which is similar to traditional auto-encoders in that it does not require labelled data, but can model the data distribution better than traditional auto-encoders. In the next paragraph, VAEs and its application for AER from acoustic signals are explained in more detail.

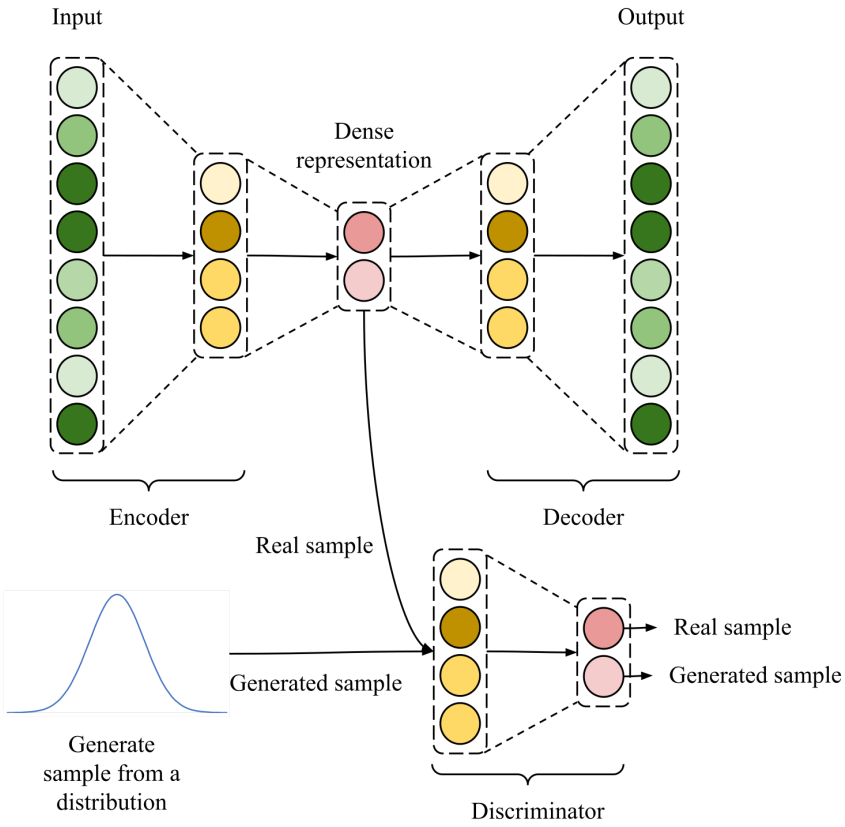


Figure 2.10: A simplified depiction of an AAE. Similar to traditional auto-encoders, the encoder and the decoder of AAEs are trained together to reproduce the input in the output. However, unlike traditional auto-encoders, AAEs also train a discriminator model in an extra step, for each time the auto-encoder is updated to distinguish the latent representation from another sample randomly generated from an arbitrary distribution.

Variational and Adversarial Auto-Encoders

VAEs aim to enrich the dense representations achieved by traditional auto-encoders by also encoding the distribution of the unlabelled data used to train such models. More specifically, compared to traditional auto-encoders, VAEs also add a regularisation term in the loss function to ensure that the representations are in an arbitrary probabilistic distribution (see Figure 2.9). The arbitrary distribution is usually considered to be Gaussian, as it has been shown to be useful for unsupervised and semi-supervised training strategies (Kingma et al., 2014; Dilokthanakul et al., 2016). The Gaussian distribution objective in VAEs allows different inputs to be represented by their variations in the latent representation space, and this is why this approach is known as VAE. An example of the application of VAEs for AER is the work of

Latif et al. (2018), where VAEs were used with recurrent layers to learn deep representations, which at the time were able to achieve state-of-the-art results for both categorical and dimensional AER.

The ability of VAEs to model the data distribution in the dense latent space also allows them to capture the statistical uncertainty among input samples. This ability of VAEs is used not only to have a good dense representation provided by the encoder, but also allows VAEs to generate new data samples with the decoder. Thus, VAEs are also can also be used as generative models. Another famous generative model is Generative Adversarial Networks (GAN), which consists of a discriminative and a generative model instead of an encoder and a decoder. GANs are trained in such a way that the discriminative model tries to distinguish the generated samples from real data. And by training the generative and discriminative models alternately for each training epoch, GANs learn to generate more realistic samples that resemble the training data. In affective computing, GANs have also been used to generate data for underrepresented emotional expressions, resulting in state-of-the-art AER performance (Chatziagapi et al., 2019).

By analysing GANs, the researchers realised that the adversary process, which focuses on discriminating between different patterns, can also provide good pattern encoding. This has led to the introduction of Adversarial Auto-Encoder (AAE), which add the adversary process of GANs to traditional auto-encoders. More specifically, AAEs use the same structure as traditional auto-encoders, with an additional training step alternately at each training epoch, where a discriminator model is also trained to distinguish between the real dense representation and a false representation sample, generated randomly from an arbitrary distribution (see Figure 2.10). For AER from acoustic signals, it means that the discriminatory process of AAEs enables them to obtain low-dimensional representations of acoustic signals, which can have the same discriminative power across different emotion categories as traditional higher-dimensional representation spaces (Sahu et al., 2018). In other words, the adversarial process in AAEs can capture the underlying patterns associated with different emotional expressions in a lower-dimensional space compared to traditional auto-encoders, i.e. a more abstract acoustic representation.

Both VAEs and AAEs provide better representation learning approaches than traditional auto-encoders because they also model the data distribution. As mentioned above, the modelling of the data distribution is achieved in VAEs by adding a regularisation term to the loss function of traditional auto-encoders, and in AAEs by an extra step of training for another discriminator model. These forms of regularisation are usually referred to as "explicit" regularisation, because they impose explicit constraints on the training, as opposed to the training constraint being an implicit result of the properties of an ANN architecture (i.e. implicit regularisation) (Hernández-García and König, 2018). Although additional explicit regularisation terms in the loss function can provide effective representations, it also adds com-

plexity to the training of the model. For example, VAEs require a further step to find optimal coefficients for different terms used in the loss function, and AAEs often have difficulty converging as a result of using multiple loss functions to alternatively train the same model[1] Salimans et al. (2016). On the other hand, the recent advent of Self-Supervised Learning (SSL) methods allows the data structure to be modelled through one loss function with no additional regularisation term. This is achieved by SSL methods predicting adjacent input samples, rather than reconstructing the input signal, which is what traditional auto-encoders, VAEs and AAEs do [2]. Moreover, the representations in SSL are not taken from a prior distribution, similar to VAEs. This fact, combined with the ability of SSLs to model the data structure in an unsupervised way, has made SSLs the mainstream representation learning technique in recent years. In the following, SSLs are explained in more detail.

Self-Supervised Learning

SSL is a type of unsupervised learning where DNN models are trained on a *pretext* task, such as predicting or recognising masked elements in the data, typically used to provide effective contextual representations of the data. Various SSL methods have been introduced in the past years that rely on various models, training strategies and inputs to learn the representation of data (see Table 2.1). One of the popular SSL training strategies for acoustic signals is Contrastive Predictive Coding (CPC), which distinguishes the representation of a masked frame of an acoustic signal from the representation of another frame, usually randomly chosen from other audio frames in the same acoustic signal (Oord et al., 2018). For example, Wav2Vec2 architecture (Baevski et al., 2020) is a recently introduced DNN based on attention mechanism that can be trained for large amounts of data and uses CPC as its training loss (see Figure 2.11). For AER from acoustic signals, Evain et al. (2021a) have shown that Wav2Vec2 is capable of learning representations that can later be used with simple feed-forward layers to predict different emotion dimensions, which was not possible with traditional features such as MFBs. This shows that Wav2Vec2 can provide us with higher-level representations of acoustic signals than signal-level features like MFBs (Alisamir and Ringeval, 2021).

[1]The problem of adversarial training of a generator and a discriminator model is usually discussed in the context of game theory. Thus, the problem is seen as finding a "Nash equilibrium" where each "player" would learn the equilibrium strategy. However, as ANN weights are randomly initialised and often solve a non-convex task, training often fails to converge in many cases.

[2]Here, SSL methods are compared to VAEs and AAEs. It should be noted, however, that SSL is mostly considered a training strategy for DNNs, while VAEs and AAEs are specific architectures of DNNs and a specific way of training that comes with that structure. Therefore, SSL and the various auto-encoders presented here are not exact substitutes for each other, and can even be combined (Kim et al., 2020b; Gatopoulos and Tomczak, 2021)

Table 2.1: Some of the most cited papers on training self-supervised representations of speech signals from 2018 to 2020.

Reference	Task and Approach	Loss	Model	Input
Oord et al. (2018)	Introducing Contrastive Predictive Coding (CPC)	InfoNCE	CNN-GRU	Raw
Chung and Glass (2018)	Speech2Vec: CBoW and Skip-gram	MSE	LSTM	MFCC
Tagliasacchi et al. (2019)	CBoW and Skip-gram, temporal gap	MSE	CNN	MFCC
Schneider et al. (2019)	wav2vec: Binary classification task for identifying the true log-mel filter-bank features using CPC loss	InfoNCE	CNN	MFB
Chung et al. (2019)	Introducing a novel autoregressive approach called APC, which predicts future log-mel spectrograms and show better performance compared to CPC	L1	LSTM	MFB
Baevski et al. (2019)	vq-wav2vec: Learning discrete representations of speech using Gumbel-Softmax and K-means. This allows for using NLP algorithms like BERT on top of the discrete representations.	Contrastive-MSE, Gumbel-Softmax	CNN	MFB
Quitry et al. (2019)	Prediction of the phase of a STFT of an audio signal from its magnitude	Cosine loss	CNN	STFT
Jiang et al. (2019)	Reconstructing masked input	L1	Transformer	MFB
Pascual et al. (2019)	PASE: Jointly solving different self-supervised tasks	L1, MSE, binary cross-entropy	SincNet, CNN	MFB, Raw, Prosody, LPS, LIM, GIM, SPC
Song et al. (2020)	Speech-xlnet: Predicting next frame	Huber loss	Transformer	MFB
Chung and Glass (2020)	A novel model using APC approach and showing better performance on ASR, speech translation and speaker identification compared to CPC and PASE	L1	GRU-Transformer	MFB
Ling et al. (2020)	DeCoAR: APC	L1	B-LSTM	MFB
Baevski et al. (2020)	wav2vec 2.0: Masking the quantised latent raw speech input using both a loss consisting of contrastive and diversity losses	Contrastive and diversity	CNN-Transformer	Raw
Chung et al. (2020)	Prediction of vector quantised log-mel spectrograms (VQAPC), showing better performance than APC without quantisation for phoneme speaker classification	Gumbel-Softmax	GRU	MFB
Wang et al. (2020)	MPC	L1	B-LSTM	MFB
Liu et al. (2020)	Mockingjay: Reconstructing masked frames	L1	Transformer	MFB

As mentioned above, the CPC loss function used in models such as Wav2Vec2 is computed on the latent representations of audio frames, not on the audio frames themselves. This is because acoustic signals are inherently continuous and therefore

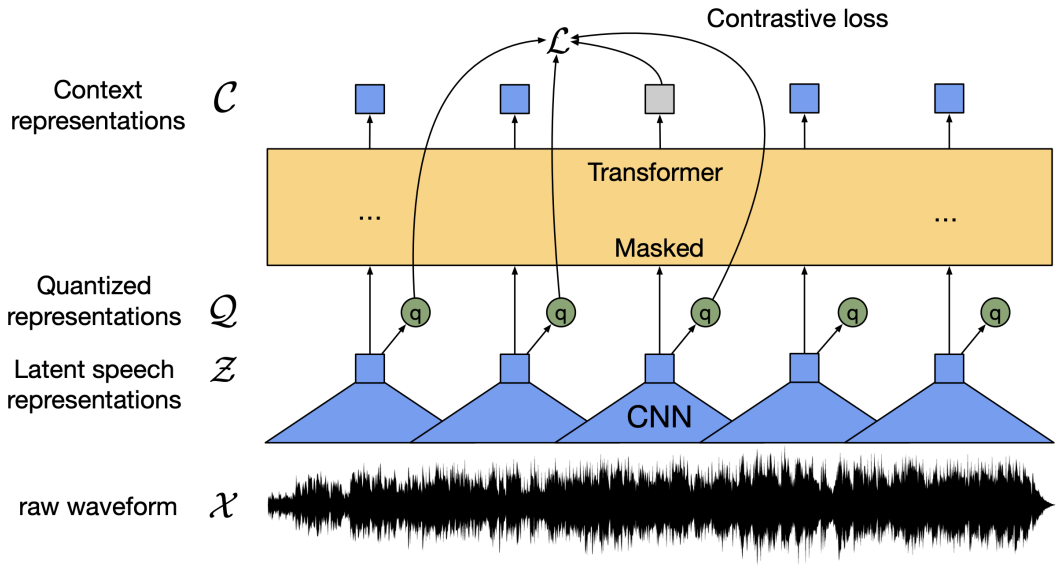


Figure 2.11: The Wav2vec2 architecture that can learn contextual acoustic representations. This model uses Convolutional Neural Networks (CNNs) to predict the latent representations, and then transformers, which are models consisting of multi-head attention and feed-forward layers, to predict the masked representations. For training, this model uses contrastive loss to train the model in a self-supervised manner to predict the masked quantised latent acoustic representations based on the context.

The figure is taken from Baevski et al. (2020)

cannot be used in a binary classification function such as CPC. This has led to the introduction of Autoregressive Predictive Coding (APC), which solves a regression task by minimising an L1 loss. A DNN model trained with APC can thus learn to predict masked audio frames as they are, rather than discriminating correlates of the latent representations of the signal (Chung et al., 2019). Self-supervised representations based on APC have also shown state-of-the-art results for AER from speech (Zhang et al., 2021), even outperforming representations trained with CPC (Chung and Glass, 2020).

The deep acoustic representations discussed above have completely changed machine learning paradigms in many domains, including AER, replacing traditional feature extraction steps with the computation of pre-trained deep representations (see Figure 2.12). Such paradigm shift has its origin in the text processing domain, where deep textual representations have completely overtaken the state of the art. Deep textual representations are discussed below.

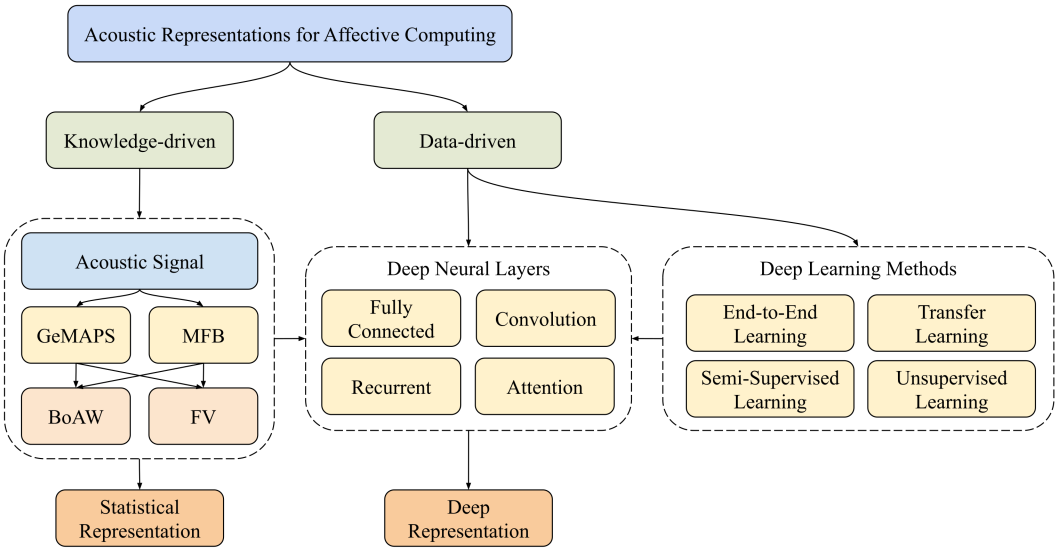


Figure 2.12: Common speech representations used for affective computing. Traditional features such as MFB and GeMAPS can compute the low-level descriptions of the signal, then rudimentary statistical methods such as BoAW and FV can be used to model the signal more contextually. Novel techniques all involve deep learning methods to train different combinations of deep neural layers to compute deep representations of the acoustic signals.

2.2.4 Deep textual representations

Any given text is first tokenised and numerically encoded in order to be represented for machines. Earlier in this chapter, "Traditional textual representations" were discussed in Section 2.2.1. It was discussed that the simplest approaches assign independent one-hot vectors to each token (e.g. word), which would not encode any semantic or syntactic information. Statistical approaches such as TF-IDF were then introduced, which could encode the statistical saliency of each token in different documents. However, TF-IDF is a rather rudimentary statistical measure and does not take any semantics into account. Later on, relational mapping methods such as LSA were introduced, which can encode semantic similarities between different words but suffer from a "syntactic blindness" problem (Suleman and Korkontzelos, 2021). On the other hand, the advent of data-driven DNNs has led to the modelling of a language using such techniques, and mostly abandoning the aforementioned traditional techniques. In what follows, a brief description of such methods is provided to the reader.

In order to model the semantic similarity using the context in which the word was placed, word embedding methods have been introduced. Word embedding can embed the knowledge conveyed by different words into low dimensional vectors.

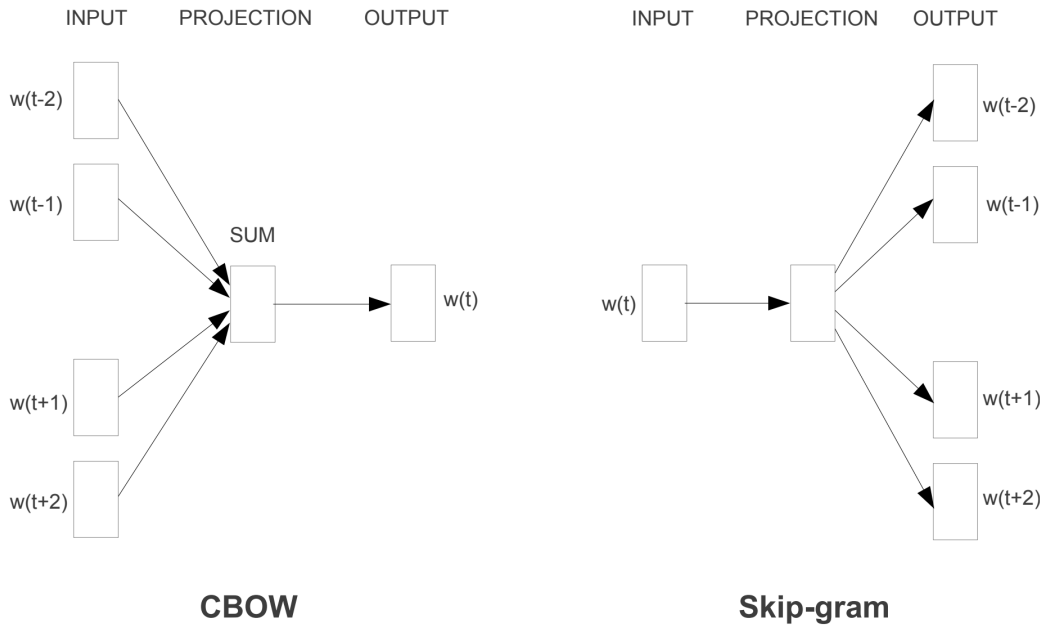


Figure 2.13: The Continuous Bag-Of-Words (CBOW) and skip-gram model architectures used to learn word2vec embeddings. The CBOW model learns to predict the representation of a word from the its neighbours. On the other hand, the skip-gram model uses the representation of a word to predict its neighbours.
The figure is from Mikolov et al. (2013)

One of the most famous word embedding methods is word2vec (Mikolov et al., 2013), where a DNN model is trained to produce deep representations of each word. To produce such word representations, word2vec uses two different model architectures: 1) Continuous Bag-Of-Words (CBOW), and 2) continuous skip-gram (see Figure). In both cases, a window slides over the entire corpus, taking a set of adjacent words as input. In CBOW, the model is trained to predict the representation of a word from the representations of its neighbours. skip-gram follows the opposite strategy to CBOW, where the model learns to predict the representations of a word's neighbours from the representation of that word. word2vec has shown good performance for many text-based tasks, such as document categorisation (Lilleberg et al., 2015) and predicting arousal and valence dimensions of emotion (Povolny et al., 2016). However, it has been shown that word2vec does not always outperform traditional TF-IDF measures (Cahyani and Patasik, 2021). One of the shortcomings of word2vec is that it is trained on the "local context" of each word and does not have a "global" description of the occurrences of each word, similar to TF-IDF or LSA. This has led to the introduction of Global Vectors for Word Representation (GloVe), an unsupervised method for learning global word-word co-occurrence matrices (Pennington et al., 2014). GloVe has also been used

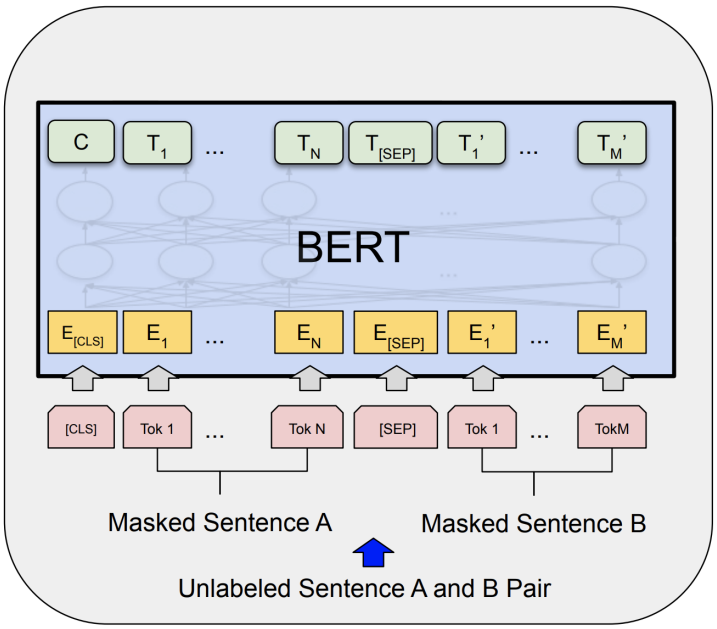


Figure 2.14: The overview of the BERT model, which uses transformers to predict masked tokens. Each embedding ("E") is computed as the sum of the token embeddings, its positional embedding within a sentence, and its sentence positional embedding across a set of sentences. Also, "[SEP]" is used as a separator token, and "[CLS]" is a special token intended for sentence level classification tasks.
The figure is from Devlin et al. (2019)

for AER from text. For example, Krishna and Patil (2020) and Xu et al. (2019) trained LSTM and attention-based DNNs on Glove word embeddings and achieved state-of-the-art emotion recognition from text.

Word embedding methods such as word2vec and GloVe are considered static representations of text because they represent each word with a fixed vector, regardless of the context (e.g. sentence or paragraph) in which the word occurs. This is a problem because a word can have different meanings depending on its context. For example, the word "right" in the phrase "this is right" means "correct", but the same word in "keep right" means the right side (as opposed to the left). In order to represent each word or token contextually, several techniques have been proposed in recent years. For example, Embeddings from Language Model (ELMo) is an architecture that uses bidirectional LSTMs to learn the language model to account for the syntax, semantics, and also the variation of each word in different contexts (Peters et al., 2018).

The advent of attention-based architectures, such as transformers (Vaswani et al., 2017b), which are models consisting of multi-head attention and feed-forward layers, has further allowed the training of more accurate and complex language

models that can better capture the context and thus have a better "understanding" of each input sentence as a whole. For example, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) uses transformers to further improve the representation of a word by taking into account its context more effectively (see Figure 2.14). It has been shown that using BERT's representations can achieve state-of-the-art results in several Natural Language Processing (NLP) tasks (Devlin et al., 2019) as well as in sentiment analysis (Sun et al., 2019) and AER (Yang et al., 2019). The good performance of BERT is mainly attributed to the use of the attention mechanism and training the model with huge amounts of data using SSL (see "Self-Supervised Learning" in Section 2.2.3). Later, a robustly optimised BERT pre-training approach (RoBERTa) was introduced, which, by training on more textual data and dynamically masking different sentences during training (BERT only does the masking once in pre-processing), proved to be more robust and outperformed traditional BERT pre-training for several downstream tasks, including AER (Liu et al., 2019; Siriwardhana et al., 2020).

So far, this section has explored the superior performance of deep representations pre-trained with SSL techniques for both acoustic and textual data for AER. On the other hand, we know that emotion can be conveyed by both verbal and non-verbal information (see "Joint Representations of Acoustic Signals and Text" in section 1.1.2). This has recently led to a new line of research investigating joint acoustic-textual representations for AER. In the following, current techniques for producing such representations are explored in more details.

2.2.5 Joint representations of acoustics and text

In recent years, many researchers have investigated joint acoustic-textual representations to improve the state of the art in AER. However, the joint learning of acoustic and textual representations is not straightforward because text and acoustic signals are inherently different. For example, textual representations are computed on the basis of tokens (i.e. subunits of text), whereas acoustic signals are continuous in time and are usually first broken down into stationary parts (typically around 25 ms). This rather technical difference means that the length of textual representations and acoustic signals are not the same, making it difficult to fuse acoustic and textual information at the signal level.

The state-of-the-art solution to this problem is to align representations of an acoustic signal and its corresponding transcription in a shared latent space. For example, Denisov and Vu (2020) first uses two separate ANN models to compute the latent space of an acoustic representation and its transcription separately. A loss function is then defined to reduce the distance between the textual and acoustic representations. In this way, the latent spaces of the acoustic and textual representations computed by the ANN models learn to be close to each other. Another example is

the work of Huang and Epps (2017), where the LLDs of acoustic signals are trained to represent the probabilities of uttered phonemes. They show that this idea can introduce "phonetic awareness" into acoustic representations, leading to a significant improvement in the performance of continuous dimensional emotion prediction compared to just using LLDs. This shows that even a phonetic level of knowledge about acoustic signals can help us to obtain verbal-aware representations, which in turn increase the accuracy of AER[1].

Different lengths of the acoustic signals and text are not the only challenge associated with making joint acoustic-textual representations. Another challenge is that textual tokens are inherently discrete both in time and value, while acoustic signals are inherently continuous in both aspects. The discreteness of text in particular implies that textual representations of a word is always unique, whereas acoustic representations of an uttered word can be different as acoustic signals are also affected by speakers, microphones and environments (Chung and Glass, 2018). Moreover, representing acoustic signals in terms of their corresponding verbal message would remove the non-verbal information needed for AER. This has led to different methods of predicting a joint representation of acoustic signals and their corresponding transcriptions, instead of aligning them to each other.

The joint acoustic-textual representations are usually achieved by concatenating the latent space of such representations. For example, in Kim et al. (2020a), acoustic and textual latent representations are first trained using AAEs and then concatenated to each other to predict different categories of emotion. In another work, self-supervised representations of acoustic signals and text are first concatenated to each other and then a fully connected layer is used to map the joint representation to different emotion categories, yielding better results than using each modality alone (Siriwardhana et al., 2020; Macary et al., 2021).

The joint acoustic-textual deep representations have been shown to be able to improve the performance of AER from acoustic signals alone. In practice, however, we do not always have access to textual transcriptions to compute the joint acoustic-textual deep representations. On the other hand, recent advances in ASR technologies have shown capable of producing reliable transcriptions in most circumstances (Kim et al., 2019). This has led to multiple recent studies on joint acoustic-textual representations for AER, where the textual transcriptions are acquired by an ASR (Heusser et al., 2019; Yoon et al., 2019; Wu et al., 2021; Peng et al., 2021). Such studies show that although the use of ASR is not as powerful as the use of human transcriptions when training AER on joint acoustic-textual representations, it still outperforms the use of acoustic representations alone. Moreover,

[1]The "phonetic-aware" acoustic representations also divide the acoustic space based on phonemes, as opposed to dividing the time domain into fixed segments, which is traditionally done to ensure stationarity. In a later study, this type of partitioning based on phonemes was shown to be a stronger contributor to the good results of predicting the Valence dimension of emotion than using the phonetic information (Huang and Epps, 2020)

these studies mainly focus on the use of acted emotional expressions and mostly pre-transcribed scenarios. Therefore, the effect of joint acoustic-textual representations with ASR transcriptions for emotional expressions in the wild has not yet been explored. Another recently emerging area that still needs to be explored is the integration of speaker information into joint acoustic-textual representations, which is discussed below.

2.2.6 Integration of speaker information

As also mentioned earlier, speaker information can be used in various ways to further improve the accuracy of AER models (see "Personalised representations" in section 1.1.2). For example, Rudovic et al. (2018) uses behavioural assessment scores of children with autism to better predict their emotional state. This is done by assigning a personalised classifier to each individual's data, while sharing a main model for all data. However, this approach is not popular in DNNs because each personal classifier would have fewer training examples, and the similarities between different individuals are ignored. Rather than assigning personal classifiers to achieve a more generalised AER model across different speakers, Peri et al. (2021) attempt to disentangle latent emotion and speaker representations. This is achieved by using a Multi-Task Learning (MTL) paradigm (see section 1.1.1), where one task is assigned to AER and the other to speaker recognition. The model is then trained using an adversarial training strategy, where an auxiliary loss function is set to discourage similarities between the emotion and speaker latent representations. However, this approach may also not be the best solution for "speaker-awareness", as it inherently considers emotional and speaker representations independent of each other, which does not fit well with the fact that emotional expressions depend on psychological idiosyncrasies (see section 2.1.1).

Given the shortcomings of the aforementioned works, the main line of research in this area focuses on representing the speaker by a "speaker style" vector, where it can be used to make acoustic representations "speaker-aware". For example, the latent speaker representations computed from pre-trained speaker recognition models have been used for AER showing a better performance than traditional LLD's (Assunção et al., 2020; Pappagari et al., 2020). Joint training of a model for both speaker recognition and AER was also investigated and outperformed using a pre-trained speaker recognition model to compute speaker representations (Moine et al., 2021). A more recent study shows that state-of-the-art AER can further be improved by exploiting speaker information into acoustic representations, which are computed by further training an ASR model to categorise emotions (Ta et al., 2022). Their work shows that AER from acoustic signals is effected by both speaker and verbal information. Although the use of speaker representations and text representations separately has been shown to improve the performance of AER from

acoustic signals, the investigation of a joint representation of acoustic signals, text and speakers seems to be missing from the state of the art.

Moreover, state-of-the-art Automatic Emotion Recognition (AER) techniques use the acoustic or textual representations with supervised data-driven machine learning methods to predict emotion annotations of a given dataset. However, each dataset represents a limited range of the vast possibilities of all the emotional expressions that can be observed in the wild. Therefore, it is important to use multiple datasets to train AER models in order to generalise across a wide range of emotional expressions. However, as numerical representations of emotion are defined in different subjective ways from one dataset to another, it is challenging to consider multiple datasets to train AER models. To address this challenge, state of the art often exploits MTL in order to consider different classifiers for different annotation schemes of each dataset, while sharing a main model across all used datasets (see Section 1.1.1). This is discussed further below.

2.2.7 Multi-task learning across various emotion annotations

MTL refers to any machine learning paradigm in which we attempt to train a common model for multiple tasks, in order to exploit the related information between different tasks, and thus improve the generalisation of the common model across all the tasks (Caruana, 1998). This is further elaborated below with concrete examples of how MTL has been used for AER.

MTL for AER from acoustic signals initially began as a means of exploiting different emotion annotations for a given dataset in order to improve overall recognition performance. For example, using arousal and valence dimensions as an auxiliary task to predict emotion labels has shown improvements compared to using only emotion categories as targets (Xia and Liu, 2015; Kim et al., 2017). Moreover, Akhtar et al. (2019) uses a GRU-based system with MTL to predict multiple emotion categories in addition to a sentiment dimension, by taking advantage of other modalities such as video and text in addition to audio, achieving state-of-the-art performance at the time. However, these works only focused on MTL of different emotion annotations for the same dataset. As there are usually several domain mismatches between different datasets, several studies have evaluated the performance of MTL for cross-corpus emotion prediction, where the model is trained on one corpus and tested on another. In particular, (Parthasarathy and Busso, 2017) focused on comparing MTL on arousal, valence, and dominance emotion dimensions in cross-corpus settings with Single-Task Learning (STL), where the target is only one task, and showed that MTL can provide models with better generalisability across different datasets when there is a correlation between the emotion dimensions across the datasets.

Zhang et al. (2017b) was the first work that investigated training on multiple

datasets at the same time in an MTL framework for AER from acoustic signals. They used different set of emotion labels from nine different corpora, as they were originally defined, and significantly improved AER performance over STL. They deliberately did not map different emotion categories from different corpora into the same subspace, as this has been shown to result in a loss of information, even when the unified emotions refer to similar affective state of mind (da Silva et al., 2020). In a later work, Zhang et al. (2022), other paralinguistic tasks were considered in addition to emotion, including 18 different classification and regression tasks. A task relatedness matrix was also introduced in order for the model to benefit more efficiently from related tasks. They also showed that their MTL approach significantly improved performance over several different tasks compared to STL. However, the focus of their study was not specific to AER from acoustic signals or to find a representation of emotion that generalises well across different annotation schemes, but rather to exploit a holistic view of different language-related tasks. Furthermore, a recent study on six different corpora showed that multi-corpus training can improve the performance of cross-corpus AER, as this approach is better suited to deal with incongruent conditions (Braunschweiler et al., 2021).

The mentioned works above mostly focused on using one language and specifically English, because using multiple corpora with different languages and cultures to train a model has shown to reduce AER performance, as emotions may be expressed differently depending on the language and culture (Ringeval et al., 2019). For example, the work of Lee (2019) has investigated multilingual MTL on gender, emotion and language tasks for two different Japanese and English datasets, and reported that multilingual models did not perform better than monolingual models. Furthermore, some emotion dimensions, such as valence, are more sensitive to language (Neumann et al., 2018), especially when only the audio modality is used (Ringeval et al., 2019). Despite the reported drop in performance in some studies, it has been shown in many others that using multiple corpora with different languages can still not only achieve reasonable performance, but can also be beneficial in dealing with rare events that occur frequently in real life Zhang et al. (2017b); Neumann et al. (2018); Zhang et al. (2022); Ringeval et al. (2019). Thus, we can benefit from using multiple emotion corpora, even if they contain emotional expressions from different languages.

The shared model in the MTL paradigm is hypothesised to learn a latent emotion representation that is more general than the numerical representation of each dataset's emotion annotation. The idea of a generic latent emotion representation, which is also called "emotion embedding" is further investigated in Zhu and Sato (2020), where an emotion encoder based on convolutional and recurrent layers are used to compute the emotion embedding. Two corpus-dependent classifiers are then used to map the emotion embeddings to corpus-specific emotion labels. They further show that by using an adversarial process to remove corpus-specific non-

emotional information, they can obtain an emotion embedding that contains cross-corpus emotional information. However, their results were only obtained on two acted datasets that had almost exactly the same set of emotion labels. Also in other similar works, the use of MTL for AER has mainly focused on acoustic signals and acted emotional expressions, and further research on both acoustic signals and text in the context of MTL for both acted and in-the-wild emotional expressions is lacking in the state of the art. Furthermore, the use of MTL with deep pre-trained acoustic and textual representations could also be considered in order to advance the state of the art in this area, due to the superior performance of deep acoustic and textual representations mentioned in Section 2.2.3 and Section 2.2.4 respectively.

So far this section has introduced different representations and AER models for predicting emotions. In the following, the different methods are quantitatively compared to each other, in a case study in order to better choose the experimental methodology of this thesis.

2.2.8 Performance comparison of methods

There are several studies that have attempted to quantitatively compare different methods for AER. For example, Wani et al. (2021) analyses different approaches and concludes that although deep learning methods represent a turning point in AER research, current algorithms are still not robust enough for use in today's Human-Computer Interaction (HCI). In another paper, Atmaja et al. (2022) examines different techniques for the joint acoustic-textual representation, showing no significant difference between different levels of merging information between the two modalities, as well as the effectiveness of BERT-based models for AER, identifying the room for research in bimodal multi-corpus AER, which is studied in this thesis (see Section 5.2). In another paper, Wang et al. (2022) systematically reviews multi-modal AER with a focus on physiological signals, then provides a taxonomy of current trends, and concludes that more datasets, as well as more research on deep pre-trained representation methods, are needed in affective computing. Although deep representations have shown the best performance for multimodal AER, a comparison between different methods is difficult by studying recent works. This is because different papers use different methods on different datasets, so different methods cannot be compared in a fair way (Zhao et al., 2021). In a similar study, Sharma and Dhall (2021) compares different labelled datasets, showing that the choice of data can significantly affect the performance of AER models, and thus comparing different techniques across different datasets may not be fair. In our recent work (Alisamir and Ringeval, 2021), we have also tried to quantitatively compare different acoustic representations and their fusion with linguistic information by analysing different results on recent challenges over several years. However, we could not draw any firm conclusions based on the results, as the challenges have

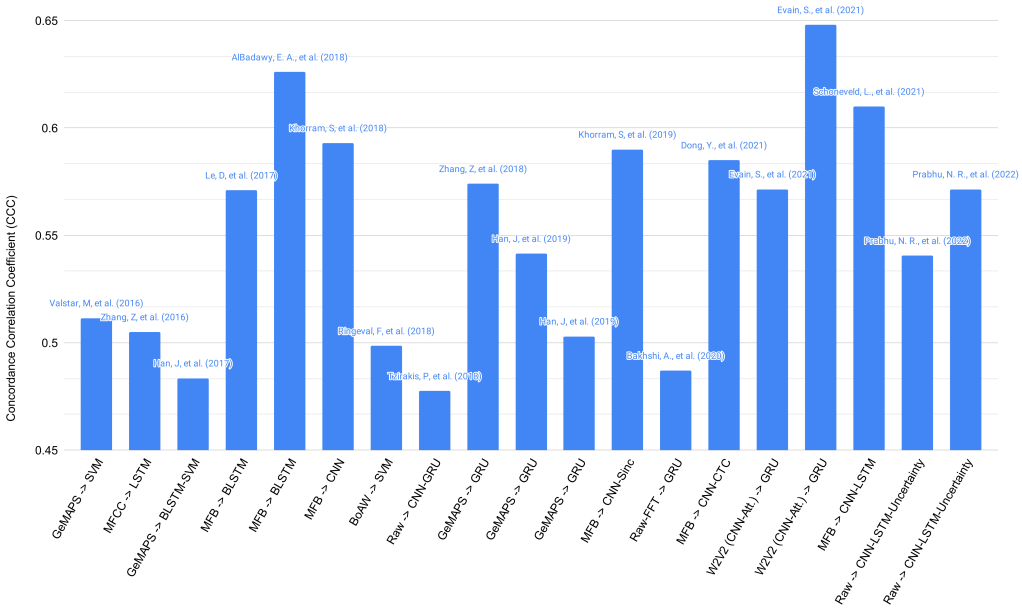


Figure 2.15: A quantitative comparison of the methods used for the RECOLA dataset, over the years 2016 to 2022. The results here are the average of the arousal and valence dimensions of emotion, based on the gold standards of the test sets.

different tasks, metrics, models and datasets.

To avoid the aforementioned problems, here we conduct a study on the results of the RECOLA dataset (Ringeval et al., 2013), which is a dataset for predicting arousal and valence dimensions of emotion (see Table 3.1 for a statistical summary of the various datasets used here). RECOLA contains 27 spontaneous and naturalistic recordings from french-speaking subjects that each are 5 minutes long and annotated with a sampling rate of 25Hz. This dataset divides the 27 subjects into 3 equal groups of training, development and test sets. There are six annotations done by french speakers. Also, a gold-standard rating as a consensus emotion was calculated from the annotations, which is not publicly available for the test set. Thus, researchers had to submit their predictions of the test set, which are all computed using the same metric of Concordance Correlation Coefficient (CCC) (see "Concordance Correlation Coefficient" in Section 3.4). Moreover, this dataset has been worked on for several years by different researchers across the world using different models and features, in order to obtain a comprehensive view of the trends in AER. The unified evaluation for the test set, together with the existence of several works over the past years, means that RECOLA has the potential for a statistical study over several years to find the impact of different representations and models on a unified metric and dataset. Therefore, this case study is carried out on the RECOLA

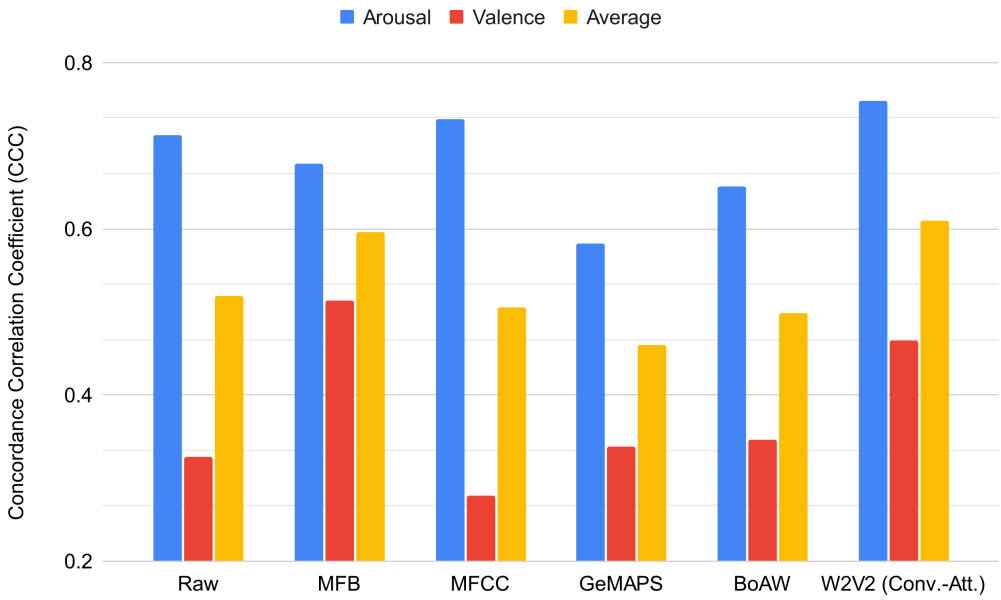


Figure 2.16: A performance comparison of the average results for each acoustic representation across different studies using different models to predict the arousal and valence dimensions of the RECOLA dataset, between the years 2016 and 2022.

dataset and the results are presented in the following paragraph. However, it should be noted that although the focus here is on representations and models, other details such as model parameters or training strategy may also affect the results. It should also be noted that since this is a case study on the RECOLA dataset, the effectiveness of the different methods cannot necessarily be extended to other datasets. Nevertheless, the results of this study can still be used as a starting point to define the experimental methodology for this thesis.

Figure 2.15, shows a summary of the different methods used for RECOLA, over the years 2016 to 2022. As can be seen, all methods involve machine learning techniques, in particular ANNs and SVMs. Moreover, the best results are obtained with Wav2vec2 representations, followed by a GRU (Evain et al., 2021a,a). Similar state-of-the-art results have been achieved previously, notably for the valence dimension by using MFBs with LSTMs (AlBadawy and Kim, 2018), and more recently for the arousal dimension by using raw input with CNN-LSTM modelling, in an end-to-end manner (Prabhu et al., 2022). As Wav2vec2 representations also contain convolutional and attention layers, we can see that all the best results involve convolutional layers, either learned through end-to-end or transfer learning, to effectively model the acoustic signals at the signal level. Similarly, recurrent layers used to model context and emotion are responsible for the best reported results. To fur-

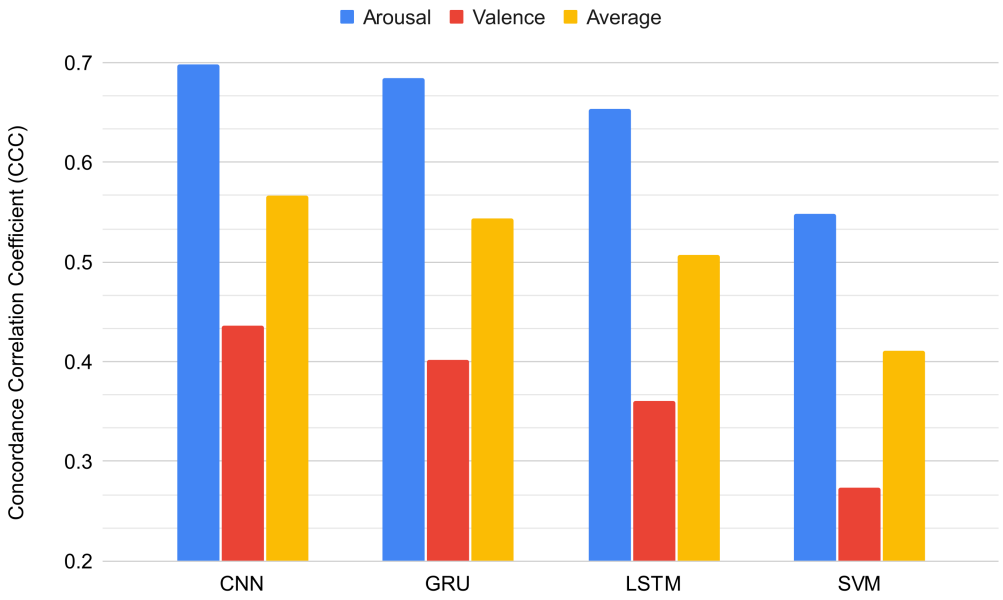


Figure 2.17: A performance comparison of the average results for each machine learning model across different studies using different acoustic representations to predict the arousal and valence dimensions of the RECOLA dataset, between the years 2016 and 2022.

ther compare the results, they are statistically summarised across different acoustic representations and models in Figure 2.16 and Figure 2.17 respectively. The summary of features shows that traditional MFB features are still able to achieve state-of-the-art results for both arousal and valence dimensions of emotion for RECOLA dataset, however deep representations seem to challenge the status quo on different datasets as seen in other work (see section 2.2.3). Figure 2.17 further shows the effectiveness of convolutional layers and recurrent units, especially GRU. However, it should be noted that the convolutional layers, which perform best in the figure, are often used for early stages of modelling the acoustic signal and are then accompanied by recurrent layers in most studies. And regarding the recurrent layers, the better performance of GRUs compared to LSTMs may be due to the fact that GRUs have a smaller number of parameters, which can be an advantage when training a relatively small amount of annotated data, as is the case for the RECOLA dataset (see "Training neural networks" in section 2.2.2). Other studies have also shown that GRUs can be trained faster than LSTMs when dealing with small amounts of labelled data, due to the smaller number of parameters, and can achieve comparable or even better results than LSTMs (Khandelwal et al., 2016; Rana, 2016; Yang et al., 2020). Since most of the experiments in this thesis are performed on rather

small annotated datasets (see Table 3.1), GRU models are the preferred choice. Further details of the experimental methodology and resources used in this thesis are explained in Chapter 3.

2.3 Summary

AER from acoustic signals and text has been an area of study for the past decades. Although emotion has no standard definition in psychology, AER usually targets either categorical view of emotion such as –fear, anger, happiness, sadness, disgust, and surprise– based on the work of Ekman (1992), or arousal and valence dimensions based on Russell (1980). In order to predict categorical or dimensional emotion annotations, traditional AER methods involve several stages of data transformation to model the acoustic signal or text at different levels. These stages are 1) low-level feature extraction, which involves signal-level modelling techniques like MFBs 2) statistical approaches to achieve more contextual modelling, which involves methods such as BoAW for acoustic signals and TF-IDF for text, and 3) mapping the statistical features onto numerical representations of an emotion annotation by using different statistical and later machine learning methods such as SVMs. However, the advent of DNNs has seriously challenged this paradigm, and made traditional low-level feature extraction methods and statistical modelling techniques less popular in recent years. This is largely due to the ability of DNNs to approximate complex functions using only data (data-driven), as opposed to traditional techniques which are mostly "knowledge-driven". Moreover, as deep neural layers can be cascaded together, they can be trained to effectively replace all of the aforementioned data transformation stages, blurring the boundaries between each stage. Furthermore, the use of DNNs pre-trained on large amounts of unlabelled data, such as Wav2vec2 model for acoustic signals and BERT for text, has dominated best performances in many domains, including AER. Such pre-trained models are particularly effective for AER because they are trained in an unsupervised manner and are not formed by subjective and often noisy emotion annotations. Moreover, the use of joint acoustic-textual representations, where the text can be either human transcriptions or produced by an ASR, has been shown to be more effective than using either acoustic or textual modality alone. Recent studies also suggest that this improvement can be further enhanced by exploiting speaker information. However, no study could be found on the use of deep pre-trained acoustic-textual representations augmented with speaker information. The use of acoustic-textual representations for emotional expressions in the wild also seems to be a gap in the state of the art.

Chapter 3

Experimental methodology and resources

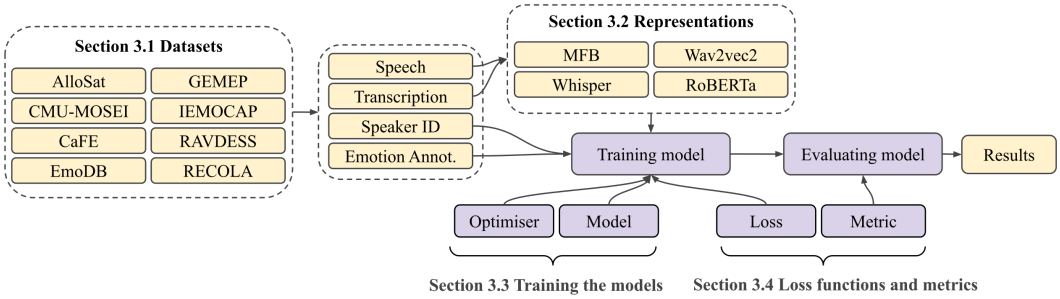


Figure 3.1: Overview of experimental methodology and resources. The datasets and representations are presented in Section 3.1 and Section 3.2 respectively. Also, the training strategy and optimisation of the different models used in this thesis are discussed in Section 3.3. The loss functions and metrics used to evaluate the model's predictions for training and testing the model are then described in Section 3.4.

At the time of writing this thesis, the best state-of-the-art performances for AER from acoustic signals and text, as well as other related domains, is attributed to the use of large DNN models, which are pre-trained on large amounts of data to extract high-level abstractions from the data (see Chapter 2). In addition, fusing acoustic and textual representations for AER to take advantage of both verbal and non-verbal communication has been shown to be more effective than using the representation of each modality alone (see Section 2.2.5). Recent studies have also shown that the use of speaker information can further improve the performance of AER from acoustic signals (see Section 2.2.6). Therefore, in Chapter 4 of this thesis, one of the main goals is to investigate the use of joint acoustic-textual representations for acted and

in-the-wild emotional expressions, and how joint acoustic-textual representations can be further fused with speaker information (see Section 1.2.1).

Most AER research today uses the aforementioned acoustic and textual representations to train models for different datasets separately, resulting in a trained model that is specific to its trained data distribution. And because each dataset covers a specific range of emotional expressions, the trained model for one dataset often cannot generalise to unseen emotional expressions. However, training a model on multiple datasets for AER to increase the generalisation ability of the trained model is challenging, because each dataset uses a specific set of emotional annotations (see Section 1.1.1). To generalise beyond the specific emotion schemes used for each dataset, in Chapter 5, this thesis investigates using MTL and joint acoustic-textual representations (see Section 1.2.2).

To achieve the aforementioned goals, the methodology used in the thesis involves training different ANN models for multiple datasets (see Section 3.1) and with deep representations of acoustic signals and text (see Section 3.1). Then, the optimisation method to train AER models based on the deep representations and different datasets is explained in Section 3.3, followed by the loss functions and metrics used to train and evaluate different models used in different experiments in Section 3.4. The details of the technical implementation of the experiments are also given in Section 3.5. Finally, a brief summary of this chapter is given in Section 3.6. Also, in Figure 3.1 the overview of the experimental methodology and resources used in this thesis is depicted.

3.1 Datasets

Different datasets typically use three types of contexts, in which an emotional expression is observed and collected (Kossaifi et al., 2021). These are 1) *Acted* emotional expressions, where a person, when prompted, attempts to utter a phrase that conveys a particular emotion, 2) *Induced* expressions, where different people's expressions are responses to a controlled setting, which is designed to elicit a particular emotion, and 3) *Natural* expressions, where different people express their emotions in natural interactions with other people or a machine. When natural expressions are captured from different speakers, microphones, and in different environments, they are usually referred to as emotional expressions in the wild.

One of the main goals of this thesis is to evaluate different deep pre-trained representations over a wide range of acted and in-the-wild emotional expressions, annotated with different annotation schemes (see section 1.2). Therefore, the list of datasets used in this thesis vary in terms of emotional context, recording environment, speakers and emotion annotations (see Table 3.1). In what follows, each dataset is explained in more detail.

Table 3.1: Summary of the data used in this thesis. Note that only the size of the annotated or labelled utterances are given as duration in the table below, and not the entire available recording files. Also, the emotional annotations here represent the annotations used in this thesis and do not represent all the annotations provided by the dataset.

Dataset	Language	Condition	Number of utterances	Number of speakers	Duration (hh:mm:ss)	Emotion annotation
AlloSat	French	In the wild, Call-center	29,704	308	20:59:27	Dimensional: Frustration-satisfaction
CMU-MOSEI	English	In the wild, youtube videos	23,259	1000	49:07:58	Categorical: Negative, positive
CaFE	French	Acted, controlled env.	936	12	01:09:16	Categorical: Anger, disgust, fear, joy, neutral, sadness, surprise
EmoDB	German	Acted, controlled env.	535	10	00:24:47	Categorical: Anger, anxiety, boredom, disgust, happiness, neutral, sadness
GEMEP	Pseudo-french	Acted, controlled env.	1080	10	00:43:20	Categorical: Anger, despair, fear, fun, interest, irritation, joy, pleasure, pride, relief, sadness, worry
IEMOCAP	English	Acted, controlled env.	5531	10	06:59:20	Categorical: Anger, happiness (including excited), neutral, sadness
RAVDESS	English	Acted, controlled env.	1440	24	01:28:48	Categorical: Anger, calmness, disgust, fear, happiness, neutral, sadness, surprise
RECOLA	French	Induced, controlled env.	1578	27	00:58:16	Dimensional: Arousal, valence

3.1.1 AlloSat

AlloSat is a recent corpus containing 37 h of continuous real-life call center recordings in French (Macary et al., 2020). It contains 29,704 utterances (21 h) in total, which is divided into 20,785 utterances (15 h) as training partition, 4272 utterances (3 h) for development and 4643 utterances (3 h) as the test partition. In total, this dataset contains 308 speakers, of whom 191 are women and 117 are men. The annotation of this dataset is done by three annotators to describe a time-continuous dimension of emotion, which ranges from frustration to satisfaction, with a sampling rate of 4 Hz. The annotations are made for each audio file, which consists of several utterances of variable length, ranging from 32 seconds to 41 minutes. The three annotations for each audio file are then averaged to define a *gold-standard* frustration/satisfaction dimension.

3.1.2 CMU-MOSEI

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) is currently the largest labelled dataset for AER (Zadeh et al., 2018). It contains 49 hours of emotion and sentiment annotation for 23,259 utterances, of which 18,542 utterances (38 h) are allocated for training, 1377 (3 h) for development and 3340 (8 h) for testing. The recordings are selected from 1000 online YouTube speakers, balanced by gender, and across a range of topics purposefully selected to cover a wide range of different emotional expressions. The annotations consist of Ekman's six basic emotion categories –anger, disgust, fear, happiness, sadness, surprise– (and a neutral expression indicating the absence of the six basic emotions), as well as a sentiment dimension. The sentiment dimension is annotated within the range of $[-3, +3]$, representing a range from unpleasant to pleasant emotions, similar to the valence dimension.

3.1.3 CaFE

The Canadian French Emotional (CaFE) dataset contains 12 actors (six male, and six female) reading six different French phrases expressing the basic emotions of anger, disgust, happiness, neutral, fear, surprise and sadness (Gournay et al., 2018). The CaFE dataset is in total about an hour long, with 936 utterances. As there was no standard partitioning allocated for this dataset, here to have both male and female speakers in all partitions, actors 9 (male) and 10 (female) are used for development (156 utterances, 12 minutes), the utterances of actors 11 (male) and 12 (female) are used for testing (156 utterances, 13 minutes), and the rest for training (624 utterances, 44 minutes).

3.1.4 EmoDB

The Berlin Database of Emotional speech (EmoDB), is the smallest dataset used in this thesis, by containing about half an hour of 535 acted German utterances (ten actors, five male, and five female), expressing happy, angry, anxious, fearful, bored, disgusted, sadness, and neutral emotions (Burkhardt et al., 2005). Due to the lack of a standard partitioning and in order to be gender balanced, the utterances of actors 11 (male) and 13 (female) are used for development (116 utterances, 5 minutes), 15 (male) and 16 (female) for testing (127 utterances, 6 minutes), and the rest is used as training set (292 utterances, 13 minutes).

3.1.5 GEMEP

The GEneva Multimodal Emotion Portrayals (GEMEP), is the only dataset used in this thesis that is deliberately designed to contain no verbal information, using

real French syllables in an order that does not produce meaningful words (Bänziger et al., 2012). The GEMEP corpus is based on ten actors who have expressed 18 different emotional states. However, to follow other works like Bänziger et al. (2012); Xu et al. (2018), only 12 core emotions of anger, despair, worry, irritation, fear, sadness, amusement, joy, pride, interest, pleasure and relief are used here, spanning 1080 utterances, 43 minutes in duration. As there is no standard partitioning for GEMEP, and to be gender balanced, actors 5 (male) and 9 (female) are used for development (216 utterances, 9 minutes), 8 (male) and 10 (female) for testing (216 utterances, 8 minutes) and the rest for training (648 utterances, 26 minutes).

3.1.6 IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP), is another acted dataset, containing about 12 hours of scripted and improvised audiovisual recordings (Busso et al., 2008). It contains 5531 utterances (7h) of expressed emotions, annotated by several people (varying between 5 and 7) for both categories and dimensions. To follow the state of the art (Siriwardhana et al., 2020), here also only the expressions labelled as anger, happiness (+ excited), sadness and neutral are included. The "excited" labelled utterances are also merged with the "happiness" labelled utterances to better balance the number of examples across the different categories. As no standard partitioning is found for this dataset, here, unless otherwise stated, the first three sessions are used as training data (3259 utterances, 4 hours), the fourth session as development set (1031 utterances, 1 hour), and the fifth session for testing (1241 utterances, 2 hours). As each session has one male and one female speaker, this results in gender balanced training-development-test partitions.

3.1.7 RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains both singing and normal speech utterances (Livingstone and Russo, 2018). The normal speech utterances is recorded by 24 professional actors who try to express different phrases while conveying eight categories of emotion: anger, calm, disgust, fear, happiness, neutral, sadness and surprise. Only the 1440 normal speech utterances (1h29m) are used here, with the utterances of actors 19, 20 and 21 (two males, one female) used for development (180 utterances, 12 minutes), actors 22, 23 and 24 (two females, one male) for testing (180 utterances, 11 minutes) and the rest (nine males, nine females) for training (1080 utterances, 1h6m).

3.1.8 RECOLA

REmote COLlaborative and Affective interactions (RECOLA) is a well-known corpus (Ringeval et al., 2013) to benchmark different models for predicting arousal and valence emotion dimensions from acoustic signals (Valstar et al., 2016; Ringeval et al., 2018a) (see section 2.2.8). It originally contains 46 audiovisual recordings of spontaneous interactions between French-speaking subjects solving a collaborative task under remote conditions. The recordings were made under noiseless conditions with the same recording equipment. Moreover, six people provided the annotations at a sampling rate of 25 Hz for both arousal and valence dimensions. The dataset provides a *gold-standard* calculated as the consensus between the annotations (by averaging them), which is used as the target for training AER models. Later, and due to the AVEC challenge (Valstar et al., 2016), a smaller version of this dataset with 27 subjects was popularised. Since most of the state-of-the-art results are based on this subset of the dataset, most of the experiments in this thesis are also performed on this shorter version of the dataset, in order to allow fair comparisons of the results with the state of the art. On the shorter version with 27 files, 9 files of 5 minutes each are assigned as training, development and test sets. As the recordings contain many silent parts without any utterances, we also provide statistics only for the utterances in this dataset, in order to be coherent with other datasets described above. There are 575 utterances (20 minutes) in the training set, 474 utterances in the development set (17 minutes) and 529 utterances (20 minutes) in the test set.

3.2 Representations

The review of the state of the art in chapter 2 shows that the use of deep pre-trained representations of acoustic signals and text has recently become popular. In particular, DNNs pre-trained on large amounts of unlabelled data in a self-supervised manner, such as the Wav2vec2 model for acoustic signals and BERT-based models for text, have dominated the state of the art in many domains, including AER. Therefore, Wav2vec2 and BERT architectures are used for the studies in this thesis. Also, as the state of the art suggests that traditional MFB features are still capable of achieving comparable performance to newer Wav2vec2 representations, they are also used as a baseline in some of the experiments in this thesis. In addition, a more recent "general-purpose" DNN model trained in an end-to-end fashion, called Whisper, has also become popular (Radford et al., 2022). Whisper is trained on various tasks such as language identification, phrase-level timestamps, multilingual speech transcription, and to-English speech translation. As such tasks are relatively close to AER from acoustic signals, this model is also used as a representation for transfer learning. Moreover, Whisper has not yet been tested for AER from acoustic signals, so a comparison of this model with Wav2vec2 and MFB representations can

Table 3.2: Summary of different acoustic and textual representations used in this thesis.

Representation	Language	Training data
MFB	-	Not trained with data (knowledge-driven design)
W2V2-En large	English	960 hours of read speech
W2V2-Fr-1k base	French	1k hours of read speech
W2V2-Fr-1k large	French	1k hours of read speech
W2V2-Fr-2.7k base	French	2.7k hours of read speech
W2V2-Fr-3k base	French	3k hours of read, spontaneous, emotional speech
W2V2-Fr-3k large	French	3k hours of read, spontaneous, emotional speech
W2V2-Fr-7k base	French	7k hours of read, spontaneous, emotional speech
W2V2-Fr-7k large	French	7k hours of read, spontaneous, emotional speech
W2V2-XLSR-53 large	Multi-lingual	56k hours of read speech
W2V2-XLSR-56 large	Multi-lingual	56k hours of read speech (fine-tuned for ASR)
Whisper	Multi-lingual	680k hours of mixed speech (web-crawled)
RoBERTa	English	160GB of text (books, stories, news, social media)

also be considered a novel contribution of this thesis. Furthermore, as BERT-based models, in particular RoBERTa, have shown state-of-the-art performance for AER, they are used as textual representations in this thesis. The MFB, Wav2vec2, Whisper, and RoBERTa models are described in more details below, and a summary of them is provided in Table 3.2.

3.2.1 Mel-scale filter bank

As one of the oldest traditional acoustic feature extraction techniques, MFBs have been used for the last decades and are still able to achieve state-of-the-art performance (see Section 2.2.3). MFB features are the result of first taking the Fast Fourier Transform (FFT) of the signal, an optimised algorithm for computing the Discrete Fourier Transform (DFT) of the signal, which is typically computed using a 25 ms window shifted forward in time every 10 ms. The choice of 25 ms is to have a window large enough to extract useful acoustic information, but small enough that the statistical properties of the acoustic signal do not change with time (i.e. the signal is considered stationary). The power spectrum of the frequencies $|FFT|^2/N$ (where N is the number of FFT points used) is then computed for each FFT in a 25 ms window. The power spectrums are then mapped to the mel scale, which is a non-linear perceptual scale based on the human auditory system. This process is in practice done with filter banks (hence the name mel-scale filter banks), which partition the frequencies into several bins, where each bin uses overlapping triangular filters that corresponds to the Mel Scale. Unless otherwise stated, in this thesis, the number of mel scale filters for each window is 80 to include a large enough detail of the frequencies. Moreover, in this thesis, the MFBs are also standardised using the mean and variance obtained on the training partition of each data set, in order

to have zero mean and unit variance on that partition. The standardisation is a common practice in signal processing when using signal-level features because it helps to transform the features from different sources to become statistically similar.

However, with the introduction of a normalisation layer for DNNs (Ba et al., 2016), the deep representations no longer need an extra standardisation step. This is because the normalisation layer can be trained over a wide range of data from different sources to learn more effective normalisation of the deep representations. The deep representations used in this thesis are presented below.

3.2.2 Wav2vec2

One of the recently popular self-supervised representation learning architectures for acoustic signals is Wav2vec2 (Baevski et al., 2020). Wav2vec2 uses convolutional and transformer layers (multi-head attention + feed-forward layers) with raw acoustic signals as input. It should be noted that Wav2vec2 is not a single model, but a model architecture, which means that one can have different Wav2vec2 models depending on the data used to train it. The Wav2vec2 models used in this thesis are trained on English (Baevski et al., 2020), French (Evain et al., 2021a) and multilingual data (Conneau et al., 2021), depending on the language of the dataset used and the specificity of each experiment (see Table 3.2). Wav2vec2 models, once trained, would be able to extract more high-level representations than traditional features such as MFBs, meaning that they would be less affected by the low-level changes in the signal that are uncorrelated with high-level tasks such as AER (see Section 2.2.3).

3.2.3 Whisper

Whisper is a transformer-based model that is trained in a MTL framework for language identification, phrase-level timestamps, multilingual speech transcription, and speech translation from multiple languages to English. Unlike Wav2vec2, which uses raw acoustic signals as input, Whisper uses MFBs as input. In addition, Whisper is trained on 30-second audio files, rather than on speech utterances of variable length as in Wav2vec2 training. Furthermore, unlike Wav2vec2, Whisper is not trained for representation learning per se, but in an end-to-end manner using an encoder-decoder architecture, where the encoder provides the dense representation and the decoder is used to solve the different tasks mentioned above. Thus, in this thesis, the encoder part is used to extract deep representations that can be used for AER. Furthermore, since Whisper is partly trained for multilingual speech transcription, it is hypothesised that its representations can provide us with useful "verbal-aware" acoustic representations. As the study of "verbal-aware" acoustic representations is one of the aims of this thesis (see Section 1.2), the use of Whisper

for AER is of interest in this thesis.

3.2.4 RoBERTa

Transformer-based deep representations of text, have recently become popular thanks to BERT models (Devlin et al., 2019), achieving state-of-the-art AER from text (see Section 2.2.4). This is because BERT can provide us with an effective representation of each word in its specific context. Moreover, BERT is pre-trained in a self-supervised manner to provide effective textual representations without the use of labelled data. More recently, RoBERTa (Liu et al., 2019) has introduced a more optimised pre-training approach, where the masking of frames necessary for self-supervised learning is done dynamically during training. The masking of frames for BERT models was previously done only in the pre-training phase. In addition, the tokenisation in BERT was previously done at the subword level, and RoBERTa takes this approach a step further by using Byte Pair Encoding (BPE), where the most frequent subword pairs are replaced by different tokens in order to use a smaller number of tokens for training. The optimised pre-training approach of RoBERTa, together with the use of 10 times larger amounts of unlabelled text data (contents of English Wikipedia and books) for training, resulted in this model significantly outperforming BERT for AER from text (Siriwardhana et al., 2020; Adoma et al., 2020). Given the impressive results of this model in the state of the art, in this thesis RoBERTa is used to extract deep representations of text.

This section has introduced the reader to the representations used in this thesis, in order to train different AER models. The following section discusses the ANN models used in this thesis and how they are trained.

3.3 Training the models

The ANN models used in this thesis mostly use GRU or linear (fully connected) layers. The reason for choosing GRU is that it can effectively model sequential data, such as acoustic signals and text, and compared to similar neural layers like LSTMs, it has been shown to be faster to train and often with better results for rather small datasets (see Section 2.2.8). The linear layers are also used here as they are often needed as the last layer of an ANN, in order to transform the latent representations so that they have the same size as the target –emotion classes or dimensions– that are to be predicted. Moreover, in some experiments linear layers are used as the main AER model on top of deep pre-trained representations such as Wav2vec2 or RoBERTa presented in Section 3.2, to show that such deep representations can achieve good results without sequential layers such as GRU, as they are already contextual representations. It should also be noted that, unless otherwise stated, the deep pre-trained representations are used exclusively for feature extraction (their

weights are "frozen" and cannot be trained further). However, in some experiments the deep pre-trained representations are allowed to be further trained alongside the GRU and linear layers as part of the AER models, in which case the deep representations are usually referred to as "fine-tuned" as opposed to "pre-trained".

To train the ANN models described above, the Adam optimiser (Kingma and Ba, 2015) is used in this thesis. Adam is a SGD-based optimisation method for updating the weights of neural layers based on a given loss function (see "Training neural networks" in section 2.2.2). Put simply, Adam expands SGD by using a weighted average of the gradients to converge faster, and also decays the gradients during training so that convergence moves towards the global minimum in the early stages of training, and then slows down the oscillations as it approaches it. Although it is still debated whether SGD generalises better in the long run than Adam, it has been shown that Adam can have comparable or better results than SGD while converging faster (Zhang et al., 2020a; Zhou et al., 2020). Therefore, in this thesis, the training of the ANN models is done with the Adam optimiser.

Furthermore, the batch size is considered to be one to avoid memory problems caused by having an extra dimension to the neural weights during training. Instead, in most experiments, Gradient Accumulation (GA) (Hermans et al., 2017) is used, where the gradients of several forward passes are used to update the weights of the used model in the backward pass. Thus, having a higher GA is similar to having a higher batch size during training, in the sense that several training examples are considered for each weight update, without causing memory problems associated with high batch sizes.

This section has explained the training strategy of the different ANN models used in this thesis. As mentioned above, the Adam optimiser used here is based on SGD, which updates the weights of an ANN model to minimise a given loss function. The loss functions used in the various experiments are explained below, as well as the metrics used to evaluate the trained models.

3.4 Loss functions and metrics

The AER models used in this thesis aim at either time-continuous prediction of arousal and valence emotion dimensions, or categorisation of different emotion labels at the utterance level. The loss function and metric used for the time-continuous prediction of emotion dimensions is the Concordance Correlation Coefficient (CCC). On the other hand, the loss function and metric used for the categorisation of different emotion labels is cross entropy and Unweighted Average Recall (UAR) respectively. These measures are explained in more detail below.

3.4.1 Concordance correlation coefficient

CCC is an agreement measure used to compute the similarity between two vectors[1]. Given $\hat{y}$ and y as the ground-truth (or gold-standard) and the prediction of a model respectively, CCC can be computed as follows (Li, 1989):

$$CCC(\hat{y}, y) = \frac{2\rho\sigma_{\hat{y}}\sigma_y}{\sigma_{\hat{y}}^2 + \sigma_y^2 + (\mu_{\hat{y}} - \mu_y)^2} \tag{3.1}$$

where $\mu_{\hat{y}}$ and μ_y are the means of the vectors $\hat{y}$ and y respectively, and their variances are defined as $\sigma_{\hat{y}}$ and σ_y. ρ is the (Pearson's) correlation coefficient, which can be written as follows:

$$\rho = \frac{\sigma_{\hat{y}y}}{\sigma_{\hat{y}}\sigma_y} \tag{3.2}$$

where $\sigma_{\hat{y}y}$ is the covariance of $\hat{y}$ and y vectors, which is mathematically written as $\mathbb{E}[(\hat{y} - \mu_{\hat{y}})(y - \mu_y)]$, where $\mathbb{E}$ is expectation. Therefore, as can be seen from the $CCC(\hat{y}, y)$ formula above, the CCC measure takes into account both the covariance and the difference of the means of $\hat{y}$ and y vectors. Furthermore, this difference of the means of the vectors $\hat{y}$ and y is similar to another famous measure called Mean Square Error (MSE). MSE is mathematically described as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{3.3}$$

where n is the total number of elements in the vectors $\hat{y}$ and y. Also, in some cases, instead of MSE, Mean Absolute Error (MAE) is used, which uses the absolute difference $|\hat{y}_i - y_i|$ instead of $(\hat{y}_i - y_i)^2$. The formula of MSE shows that it does not take into account the correlation between the two vectors, and on the other hand correlation alone does not account for the differences between the vectors $\hat{y}$ and y (neither point-wise like in MSE nor averaged like in CCC). Taking into account both the difference of the vectors and their correlation makes CCC a superior measure to other commonly used methods such as MSE and correlation[2]. This has led to CCC being widely used to measure the performance of state-of-the-art AER models for predicting continuous dimensional emotion (see Section 2.2.8). It can also be used as a loss function to train ANN models to be optimised for CCC. This loss function can be written as follows:

$$\mathcal{L}_{CCC}(\hat{y}, y) = 1 - CCC(\hat{y}, y) \tag{3.4}$$

[1]The term vector is used here to refer to a list of values as defined in Goodfellow et al. (2016).

[2]For more information on the the comparison between MSE or Pearson's correlation and CCC measures as a metric and a loss function, the reader is referred to the work of Pandit and Schuller (2019), where this has been extensively studied.

where $\hat{y}$ and y can be the predictions of the model and the *gold-standard* target respectively. In other words, the $\mathcal{L}_{CCC}$ loss function aims to bring the time-continuous prediction of the arousal and valence emotion dimensions closer to its *gold-standard* target. Although $\mathcal{L}_{CCC}$ is a useful loss function when working with datasets annotated for time-continuous emotion dimensions, not all datasets are annotated in this way (see Section 3.1). Below is a description of the loss function used for datasets with utterance-level emotion labels.

3.4.2 Cross entropy

To solve a classification task (e.g. emotion label recognition), the problem is usually seen as increasing the mutual information between the set of predicted and target probabilities for the same set of events (e.g. acoustic or textual data). In information theory, how much information there is in a random variable (here y) is usually calculated using the notion of entropy, which is defined as follows:

$$E(y) = -\sum_{y \in \mathcal{Y}} p(y) \log p(y) \tag{3.5}$$

where Σ is the sum over all possible values of y, denoted by $\mathcal{Y}$, and $p(y)$ is the probability of the event y. In classification tasks, however, we are mostly interested in measuring the difference between the entropy of the model's predictions $\hat{y}$ and the target y. This is often done by taking the "cross-entropy" between the target probabilities $p(y)$ and the probabilities of a model's predictions $p(\hat{y})$ as follows:

$$CrossEntropy(\hat{y}, y) = -\sum_{i=1}^{C} p(y_i) \log(p(\hat{y}_i)) \tag{3.6}$$

where C is the number of classes, i.e. the number of elements in $\hat{y}$ and y vectors. For example, suppose we have a model that classifies an acoustic signal into negative ($i = 1$) or positive ($i = 2$) sentiment classes. In this case, when $i = 1$, $p(\hat{y}_1)$ describes the probability of the acoustic signal to be classified as negative, based on the model. Moreover, the $p(\hat{y}_i)$ is usually calculated by Softmax, which is described as follows:

$$Softmax(\hat{y}_i) = \frac{e^{\hat{y}_i}}{\sum_{j=1}^{C} e^{\hat{y}_j}} \tag{3.7}$$

Interestingly, part of the reason why the use of cross-entropy loss is common for classification tasks, is that using Softmax to obtain prediction probabilities gives us a simple equation ($\frac{\partial \mathcal{L}}{\partial p(\hat{y}_i)} = p(y_i) - p(\hat{y}_i)$) when calculating the derivatives of the loss function for backpropagation (see "Training neural networks" in section 2.2.2).

Although cross entropy can also be used as a metric, in many cases the goal is not to evaluate how accurately a model can predict a probability distribution. The ultimate goal in many classification tasks is to automatically assign an accurate label to a given datum. Therefore, metrics such as accuracy or Unweighted Average Recall (UAR) are usually used to evaluate models built for classification. These metrics are described below.

3.4.3 Unweighted average recall

One of the most common metrics used to evaluate classification models is accuracy, which can be calculated following the formula below:

$$Accuracy = N_c/N \tag{3.8}$$

where N_c is the number of instances correctly classified by the model over the total number of instances denoted as N. Although accuracy is a simple but effective metric in most cases, it has a limitation. Namely, accuracy does not give a good indication of a model's performance if the number of instances is not balanced across different classes. For example, if a model predicts all testing samples as class one in a binary classification task, and the test instances are 95 % class one, then the accuracy of that model would be 95 %. However, the high accuracy of the model is misleading because we know that the model does not actually work and predicts everything as class one. In order to account for unbalanced class labels, the use of UAR has been proposed, which first calculates the accuracy within each class and then takes the average of the accuracies obtained for all classes. This can be written mathematically as (Eyben, 2015):

$$UAR = \frac{1}{k}\sum_{i=1}^{k}\frac{N_c^i}{N^i} \tag{3.9}$$

where k is the total number of classes, N_c^i and N^i are the number of correctly identified instances and the total number of instances for class i respectively. Furthermore, if the number of instances across different classes is balanced, then UAR is the same as accuracy. In this thesis, the number of instances across different classes is not always balanced for all the datasets used in different experiments. Therefore, UAR is used here as the metric to evaluate the models for classification tasks.

3.4.4 Word error rate

In the experiments of Section 4.3, Section 4.4.2, and Section 5.2, an ASR is used to obtain automatic transcriptions for the IEMOCAP and CMU-MOSEI datasets. To

evaluate the performance of automatic transcriptions, they are compared with human transcriptions using the Word Error Rate (WER) metric, which can be written mathematically as follows:

$$WER = \frac{S + D + I}{N} \tag{3.10}$$

Where S, D, and I are the number of substitutions, deletions, and insertions respectively. And N is the number of words in the human transcriptions.

So far, this section has explained the different metrics used to measure the performance of different models. However, in order to compare the performance of two different models, certain statistical tests are required. The statistical significance tests used in this thesis are explained below.

3.4.5 Statistical significance

In this thesis, a two-tailed p-test is used to calculate the statistical significance between two values across different experiments. A two-tailed p-test, also known as a two-tailed hypothesis test, is a statistical method used to determine whether there is a significant difference between an observed sample statistic and a population parameter. It is called "two-tailed" because it considers differences in both directions from the expected value, rather than just one direction.

It is also assumed here that the results of different experiments, whether using UAR or CCC, are the mean values in a normal distribution with a standard deviation of one. Then, to calculate the z-value, which represents the difference between the calculated metrics across different experiments, we can write:

$$z = \frac{v_1 - v_2}{\sqrt{2/n}} \tag{3.11}$$

where n is the number of the population (here utterances). Then, assuming that the z-value follows a standard normal distribution, we can calculate its cumulative distribution function (CDF) according to the equation below:

$$\frac{1}{2\pi} e^{-x^2/2} \tag{3.12}$$

Then, to calculate the deviations from the z-value of zero in both directions (i.e., in both tails of the distribution), we calculate the area under the curve of the normal distribution of z. In this way, we can account for the possibility of observing extreme results in both tails of the distribution. Thus, to calculate the area under the curve we can write:

$$\Phi(z) = \frac{1}{2\pi} \int_{-\infty}^{z} e^{-x^2/2} dx \tag{3.13}$$

and since the standard normal distribution is symmetric and the total area under it is one, to calculate the p-value we can write:

$$\Phi(-z) = 1 - \Phi(z) \tag{3.14}$$

$$p = 2.\Phi(-|z|) \tag{3.15}$$

Then, if the p-value is above 0.05, we assume that the null hypothesis is true and the difference between the two values, which can be the results of different experiments, is not significant. However, if the p-value is less than 0.05, we assume that the null hypothesis is false and there is a significant difference between the results of different experiments.

Furthermore, in the experiments where CCC was used as the evaluation metric, the Fisher r-to-z transformation is first used, before calculating the z and p values. The reason behind this is that algorithms that compute correlation coefficients, like CCC, are not directly commutable from separate individual values, which means that it is not possible to directly add or subtract the different averages of correlation values across different experiments. The Fisher r-to-z transformation is thus commonly used to assess the significance of the difference between two correlation coefficients, and can be written as follows:

$$z_r = \frac{1}{2}\ln(\frac{1+\rho}{1-\rho}) = artanh(\rho) \tag{3.16}$$

where ρ is the correlation coefficient (in this case it is CCC) and *artanh* is the inverse hyperbolic tangent function. Therefore, it is argued that the Fisher transformation approximates a variance-stabilising transformation for ρ, when ρ follows a normal distribution. Thus, with the Fisher transformation, the variance of r grows faster as ρ gets closer to 1. Then, z_r in the equation above can be calculated for two different ρ values across different experiments, and then the two z_r can be used in the equation 3.11 to calculate the z-value, which can then be used in the equation 3.15 to calculate the two-tailed p-value.

3.5 The technical pipeline of the experiments

Figure 3.2 depicts the technical pipeline used to carry out the experiments in this thesis. In the figure, the "experiment parameters" indicates the datasets, feature extraction method, model, optimiser, the loss function, and metrics to use. Then, the following steps are performed:

1. The video/audio files of all the datasets are converted to mono wav files with PCM signed 16-bit little-endian format and 16 KHz sampling rate using "ffmpeg". This is done to have a coherent audio format across different datasets, which is particularly important for the MTL experiments in Chapter 5.

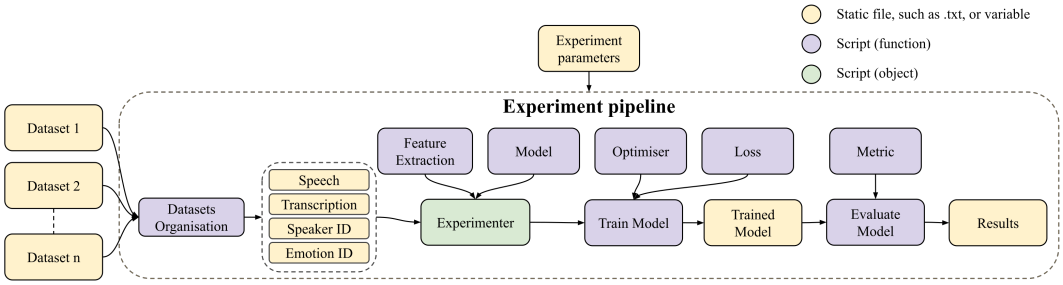


Figure 3.2: Overview of the technical pipeline for running experiments.

2. The paths associated with each converted wav file, the human transcriptions (if available), off-the-shelf ASR's transcriptions, and the emotion annotations for each file are stored in a static file (usually as a ".json" file). Similar to the first step above, this step helps to have a coherent audio loading paradigm when using multiple datasets.

3. The information about the input data and the targets, the used optimiser, the loss function, the model and hyper-parameters, are used to train the models dynamically after extracting features for each batch of training instances. Dynamic training here means that for each batch of audio files, they are first loaded into memory, and then features are extracted based on the representations mentioned in Section 3.2. Then, the model is trained based on the extracted acoustic or textual features. This is a common technical trick to save memory, as opposed to extracting features for all the files first, which would simply not be possible if we have limited memory, large amounts of data, and are using large pre-trained models such as Wav2vec2 for feature extraction. In the Figure 3.2, this process is encapsulated in the object "Experimenter". Again, this is a common practice in the implementation of the experiments for deep learning. For example, SpeechBrain[1] toolkit (Ravanelli et al., 2021), defines a "Brain" object which is responsible for the same processes as the "Experimenter" method, Keras[2] toolkit achieves the same objective with their "Model" object, and Huggingface[3] toolkit implements the same idea within their "Trainer" object (see Section 3.3 for training the models).

4. Once training is complete, the trained model is evaluated using a metric and the test partition of each dataset (see Section 3.1 for partitioning information, and Section 3.4 for metrics).

Lastly, all the experiments in this thesis were carried out using Pytorch[4] (Paszke

[1] https://speechbrain.github.io
[2] https://keras.io
[3] https://huggingface.co/
[4] https://pytorch.org

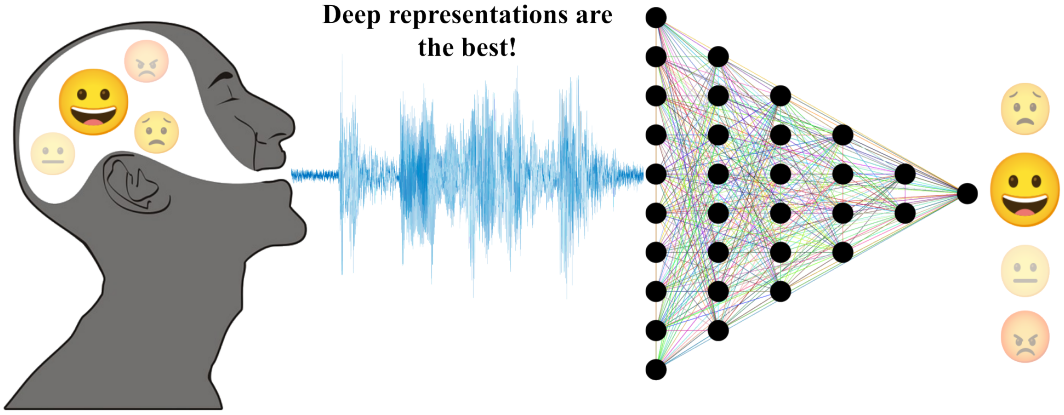
et al., 2019), and by using the "Brain" object of the SpeechBrain toolkit as the "Experimenter", with random seeds manually set to zero. The computer's operating system was Debian GNU/Linux 10, and the GPU used to train the models was an NVIDIA Quadro RTX 6000 with 23 gigabytes of memory, CUDA version 11.3.

3.6 Summary

The review of the state of the art in the previous chapter showed that deep pre-trained representations of acoustic signals and text are significantly more performant than traditional techniques for AER. The aim of this thesis is to extend the related state of the art on using deep pre-trained acoustic and textual representations for a wide range of emotional expressions, whether acted or in the wild. Therefore, several deep pre-trained representations are used in different experiments of this thesis, as well as a wide range of datasets. Namely, the pre-trained deep representations used here are Wav2vec2, and Whisper for acoustic signals and RoBERTa for text, chosen particularly for their superior performance in the state of the art. Also, the datasets used in this thesis are AlloSat, CMU-MOSEI, CaFE, EmoDB, GEMEP, IEMOCAP, RAVDESS, and RECOLA, which vary in terms of recording environment, speakers, emotion annotations, and the context in which the emotional expressions are collected (acted, induced, and natural). Furthermore, in various experiments throughout this thesis, the Adam optimiser is chosen to train the ANN models for different loss functions because it converges faster and can achieve comparable or better results than the basic SGD algorithm. The loss functions and metrics used to train and evaluate the models vary depending on the task. For time-continuous prediction of emotion dimensions such as arousal and valence, $1 - CCC$ is used as the loss function and CCC as the metric, because it can measure both the covariance of predictions and targets and the distance between their means. And to categorise emotion labels such as happiness, anger, sadness and neutral, cross entropy is used as the loss function and UAR as the evaluation metric. The choice of UAR over the more common accuracy metric is because accuracy does not give a good indication of a model's performance if the test data is not balanced. Lastly, all the experiments are performed using Pytorch and the SpeechBrain toolkit.

Chapter 4

On the use of deep acoustic and textual representations



Deep pre-trained representations have recently dominated state-of-the-art benchmarks in many areas of study, including AER (see Section 2.2). In particular, deep pre-trained acoustic representations have been shown to outperform traditional feature extraction methods for a wide range of speech-based tasks, such as ASR, AER, and speaker recognition (Latif et al., 2020). The good performance of deep representations is usually attributed to the fact that such methods are trained on large amounts of data to *brute-force* an approximation of an effective acoustic filter, rather than being hand-crafted from our limited acoustic knowledge. However, the fact that deep representations are trained in a data-driven fashion, rather than designed, makes them difficult to interpret, study and control. Nevertheless, it is clear that the functionalities of deep representations are influenced by their training data. Therefore, as part of this thesis, in Evain et al. (2021b) we studied the effect of different training data for deep acoustic representations on several tasks, including AER from speech signals, which is brought to the reader in Section 4.1.

Moreover, emotion from speech signals is conveyed both by how something is said, which can be observed from changes in the acoustic signals, and by what is said, which is the verbal message present in the transcription (see "Joint representations of acoustic signals and text" in Section 1.1.2). This has prompted a recent line of research to consider both acoustic and textual modalities of the same speech signal for AER. Also, the state of the art for AER shows that using joint deep acoustic-textual representations has better performance than using each modality alone (see Section 2.2.5). To this end, in Section 4.2, deep pre-trained acoustic and textual representations are used both separately and jointly in different AER experiments, for both acted and in-the-wild emotional expressions. Although these experiments do not advance the state of the art, they serve as a first step for the other experiments in this thesis, such as the use of ASR transcriptions for joint acoustic-textual representations, which is discussed below.

Since human transcriptions are not always available for a given speech signal, off-the-shelf ASR models are often exploited to provide us with automatic transcriptions, so that they can later be used to compute joint acoustic-textual representations. This paradigm has been shown in several studies to provide better performance than using acoustic signals alone for the recognition of acted emotional expressions (Heusser et al., 2019; Yoon et al., 2019; Wu et al., 2021; Peng et al., 2021). However, the use of ASR transcriptions for joint deep acoustic-textual representations has not yet been studied for recognition of emotional expressions in the wild. Therefore, in order to advance the state of the art, Section 4.3 investigates the effect of using ASR transcriptions in the context of joint acoustic-textual representations for the recognition of emotional expressions in the wild, as well as acted emotional expressions.

It was mentioned above that emotion is conveyed from speech signals both by how something is said and by what is said, which has led to the aforementioned studies on the use of transcriptions to improve the state-of-the-art AER. In addition, recent studies suggest that different speakers can vary greatly in how they express the same or similar emotional expressions (see Section 2.2.6). For example, Pappagari et al. (2020) has shown that the latent representations of a pre-trained speaker recognition model (i.e. speaker representations) can be used (as input features) to train AER models and achieve better performance than using LLDs of acoustic signals. Furthermore, Ta et al. (2022) showed that fusing pre-trained speaker representations with acoustic representations extracted from pre-trained ASR models can lead to better AER performance compared to using acoustic representations alone. To extend these studies, and to combine them with recent trends discussed above, Section 4.4 investigates the integration of deep speaker representations for joint acoustic-textual-speaker representations (where the text is human or ASR transcriptions) for the recognition of acted emotional expressions[1].

[1]We focused on acted emotions in order to more clearly study the interplay of the proposed

4.1 The effect of training data for deep acoustic representations

Self-supervised learning techniques, especially acoustic Wav2vec2 representations, have been shown to achieve state-of-the-art performance in speech-related domains such as ASR, and AER (see "Self-Supervised Learning" in Section 2.2.3). However, most state-of-the-art studies focus only on different architectures of deep representation learning models (see the models in Table 2.1) and their performance on downstream tasks. As a result, this trend of studies ignores the large effect of the training data in shaping the behaviour of the trained deep representations. Therefore, to advance the state of the art, in the studies published in Evain et al. (2021a) and Evain et al. (2021b), we evaluated the effect of different training data on the Wav2vec2 representations[1]. Moreover, the focus of these studies was on French data, since English and multilingual deep representations have been the subject of various state-of-the-art studies. This section presents the study of the performance of the French Wav2vec2 models (see Table 3.2) on AER, published as part of this thesis in Evain et al. (2021b). In what follows, the training of the French Wav2vec2 models are explained in more details.

4.1.1 Training self-supervised representation of French speech

Wav2vec2, first introduced in Baevski et al. (2020) as a framework for self-supervised representation learning from acoustic signals, provides two architectures, base and large. Wav2vec2-base models use 12 transformer blocks with 8 heads each, while the larger Wav2vec2-large models have twice the number of transformer blocks and heads. In Evain et al. (2021a), **we trained four Wav2vec2-large, and three Wav2vec2-base models using different amounts of training data, ranging from 1000 (1 k) hours of read speech to 7 k hours of read, spontaneous, and acted emotional speech** (see Table 3.2). This choice was made in order to study the effect of read, spontaneous and emotional speech on a series of speech-based tasks. The training of Wav2vec2 models were then performed until the loss on the development set no longer decreased significantly, which was 200k epochs for the W2V2-Fr-1K-base and W2V2-Fr-1K-large models, and 500k epochs for the rest. Figure 4.1 shows a visual summary of the type of speech used to train the W2V2-Fr-1k, W2V2-Fr-2.7k, W2V2-Fr-3k and W2V2-Fr-7k models. It should be noted that using the Wav2vec2 base or large architecture does not change the

model as a first step, before moving on to using data recorded in the wild, which would be more difficult to analyse.

[1]This study was a collaborative effort with more than a dozen researchers. My contribution to this study was explicitly to process the data for training different Wav2vec2 models and to evaluate their performance for the downstream emotion recognition task.

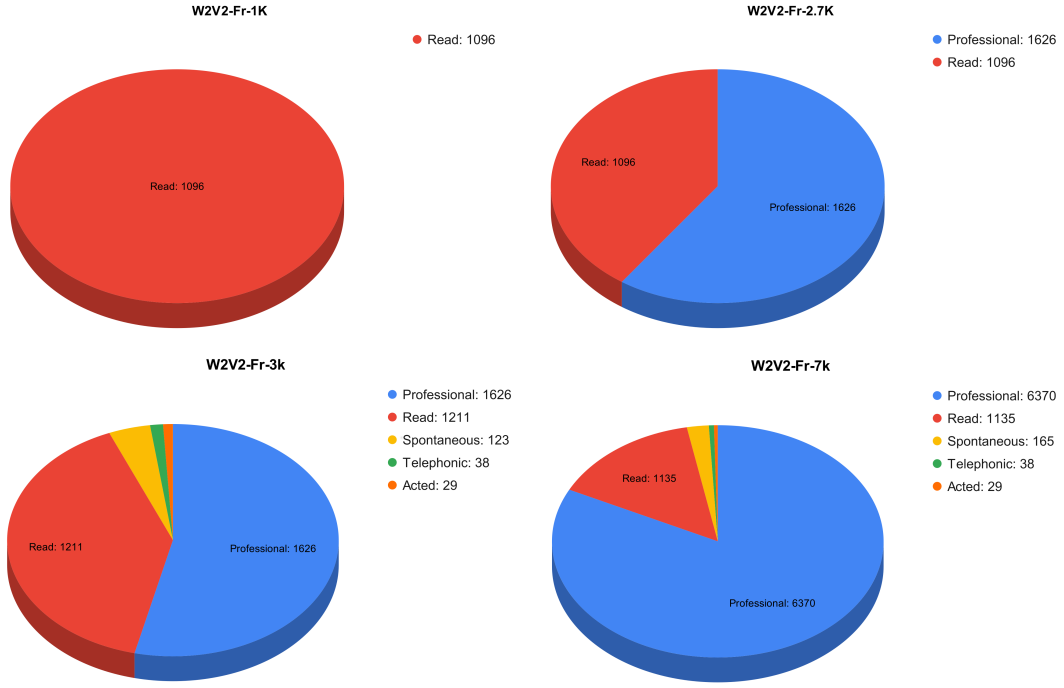


Figure 4.1: The number of hours per speech type used to train Wav2vec2 models for French speech.

data used for training the Wav2vec2 models.

The trained Wav2vec2 models were then used as speech representations, and were evaluated on a series of speech-related tasks, namely ASR, AER, natural speech understanding, and speech translation. In what follows, the AER experiments performed on the W2V2-Fr representations (with frozen weights) are explained in further detail.

4.1.2 Experiments on prediction of continuous emotion annotations

To explain the experiments used here to evaluate the W2V2-Fr representations, we first present the datasets and representations used here, and then the method used to evaluate the AER performance of the representations on the used datasets.

Datasets and representations

To evaluate the performance of the W2V2-Fr representations for AER of French speech, here two French datasets of **RECOLA** (Ringeval et al., 2013) and **AlloSat** (Macary et al., 2020) datasets were used. The **RECOLA** dataset has been used for years to benchmark the **arousal and valence** dimensions of emotion in spontaneous

Table 4.1: Statistics related to the training, development and test partitions of the AlloSat and RECOLA datasets.

Dataset	Number of utterances			Duration		
	Train	Dev	Test	Train	Dev	Test
AlloSat	20,785	4272	4643	15 hours	3 hours	3 hours
RECOLA	575	474	529	45 minutes	45 minutes	45 minutes

French speech (see Section 2.2.8). However, all RECOLA speakers are recorded with the same microphone and in the same noiseless environment. Therefore, a more recent dataset, AlloSat was also used, which consists of real-life call center conversations in French, and thus can be exploited for AER in the wild. Moreover, the **AlloSat** dataset targets one emotion dimension ranging from **frustration to satisfaction** (see Section 3.1 for more information on the RECOLA and AlloSat datasets). A summary of the RECOLA and AlloSat datasets is also provided in Table 4.1.

The RECOLA, and AlloSat datasets used for the experiments on W2V2-Fr representations contain long conversational audio files. However, the W2V2-Fr representations are trained on single utterances, lasting from one to 30 seconds. Therefore, in order to be in consistent with the training strategy of the Wav2vec2 representations, and to avoid memory issues related to loading long audio files, the long audio files of the RECOLA and AlloSat datasets are chunked into 30 seconds files for training and development. For testing however, the emotion predictions for all the audio files were concatenated before computing the CCC (see Section 3.4.1) to make the results comparable to the state of the art.

Moreover, MFB features were also used here, since they are still widely used in state-of-the-art AER models (see Section 2.2.8). In addition to using the MFB features as a baseline, two other Wav2vec2 representations were evaluated, one trained only on English –W2V2-En large– and the other trained on speech signals from 53 languages, including French –W2V2-XLSR-53 large– (see Table 3.2). The reason behind using representations not explicitly trained for French is to see how much a language other than French used to train Wav2vec2 representations can play a role in the performance of such representations for the AER on French speech. The method used to evaluate the AER performance of the Wav2vec2 representations is explained below.

Method

In order to better evaluate the performance of different Wav2vec2 representations, a number of AER models of different complexity have been considered. Namely, a simple **Linear-Tanh (LT)** model consisting only of a linear (feed-forward) layer followed by a tangent hyperbolic function, and **two one-layer GRU models, one**

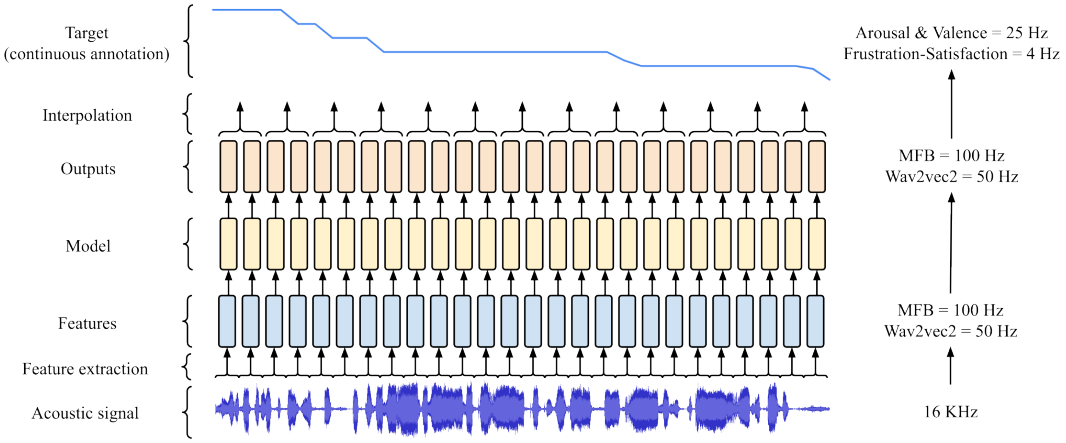


Figure 4.2: The pipeline of the method employed to evaluate continuous emotion prediction on the RECOLA and AlloSat datasets.

with 32 nodes and the other with 64 nodes, which are both followed by Linear-Tanh layers (GLT-32 and GLT-64). The GLT models were chosen because GRU has demonstrated its ability to achieve state-of-the-art performance for AER, due to its ability to effectively model sequential data, especially when having a small number of labelled examples (see Section 2.2.8). And the LT model was chosen mainly for its simplicity and inability to model data contextually, due to its time-linear nature. Therefore, by also using traditional signal-level MFB features as a baseline, the LT model would then quantitatively indicate at what point Wav2vec2 representations are more contextual than the MFB features.

To further illustrate the pipeline of the method used here, it is shown in Figure 4.2. As can be seen from the figure, the emotion predictions of the different models used here (see below) would have the same sampling rate as their input features. And the sampling rates of the different features, which were 100 Hz for MFB and 50 Hz for the Wav2vec2 representations, differ from the sampling rates of the annotations: arousal and valence annotations from the RECOLA dataset are annotated at a rate of 25 Hz, and frustration-satisfaction annotations from the AlloSat dataset are annotated at a rate of 4 Hz. This means that the sampling rate of the emotion predictions and their targets are different, which creates a problem for both training and evaluation of the models. To overcome this problem, during training, the annotations are resampled to match the sampling rate of the features (to allow backpropagation), while during testing, the output of the models is resampled to match the target annotation (to avoid altering the targets for a fair evaluation). More specifically, this is achieved by the "interpolation" part of the method depicted in Figure 4.2.

Regarding the training details of the method, the LT and GLT models were trained by Adam with 250 as the maximum number of epochs, with an early stop-

ping of 15 epochs, i.e. stopping the training if no improvement over the development set was observed (see Section 3.3). The loss (and evaluation metric) used here was the Concordance Correlation Coefficient (CCC) between the model predictions and the human annotations, as it provides a good measure for evaluating the agreement between time-continuous traces (see Section 3.4.1). The results of these experiments are presented below.

4.1.3 Results

The results of the experiments described above are shown in Figure 4.3. At first glance, there is a large variation in the results of the Wav2vec2 architectures trained on different training sets, showing the effect of different training data on the Wav2vec2 representations used for AER. For example, the GLT-64 model trained to predict the arousal dimension of emotion has the best performance (CCC= .741) with the W2V2-Fr-2.7k base representation and the worst performance (CCC= .078) with the W2V2-Fr-7k large. Also, the results for the prediction of the satisfaction dimension of the dataset corpus seem to be more stable than the prediction of the arousal or valence dimension of the RECOLA dataset, which is not surprising since the AlloSat dataset contains about 30 times more utterances than the RECOLA dataset for training, and more than eight time testing utterances (see Table 4.1).

Moreover, it is observed that the **larger amount of data used to train the Wav2vec2 representations does not necessarily lead to their better performance in predicting emotion dimensions** on the RECOLA and AlloSat corpora. For example, the best performance for arousal and valence on the RECOLA dataset is achieved when the representations are trained on 1k or 2.7k read speech (see 3.2). This is interesting because the RECOLA dataset used for AER contains spontaneous conversations, which is different from the read speech used to train the W2V2-Fr-1k representations and the radio broadcasts for the W2V2-Fr-2.7k representations. This may be because the "read speech" used to train the representations were actually readings from audio books, which are not devoid of emotional expressions, nor are radio broadcasts. Nonetheless, it is not easy to pinpoint the reason for the behaviour of deep representations, as they are slowly shaped by large amounts of data, making their behaviour difficult to interpret.

Another noteworthy result is that, on average, Wav2vec2 representations perform better than traditional MFB features for the LT model. However, for the more complex GLT model, the advantage of using Wav2vec2 representations over MFB features is less clear. This shows that the **self-supervised representations require less complex models to achieve better results than traditional MFB features**. Furthermore, the LT models are time-linear, meaning that they provide a one-to-one correspondence of features to predictions in time. Therefore, the better performance of Wav2vec2 representations with time-linear models, compared to MFB features,

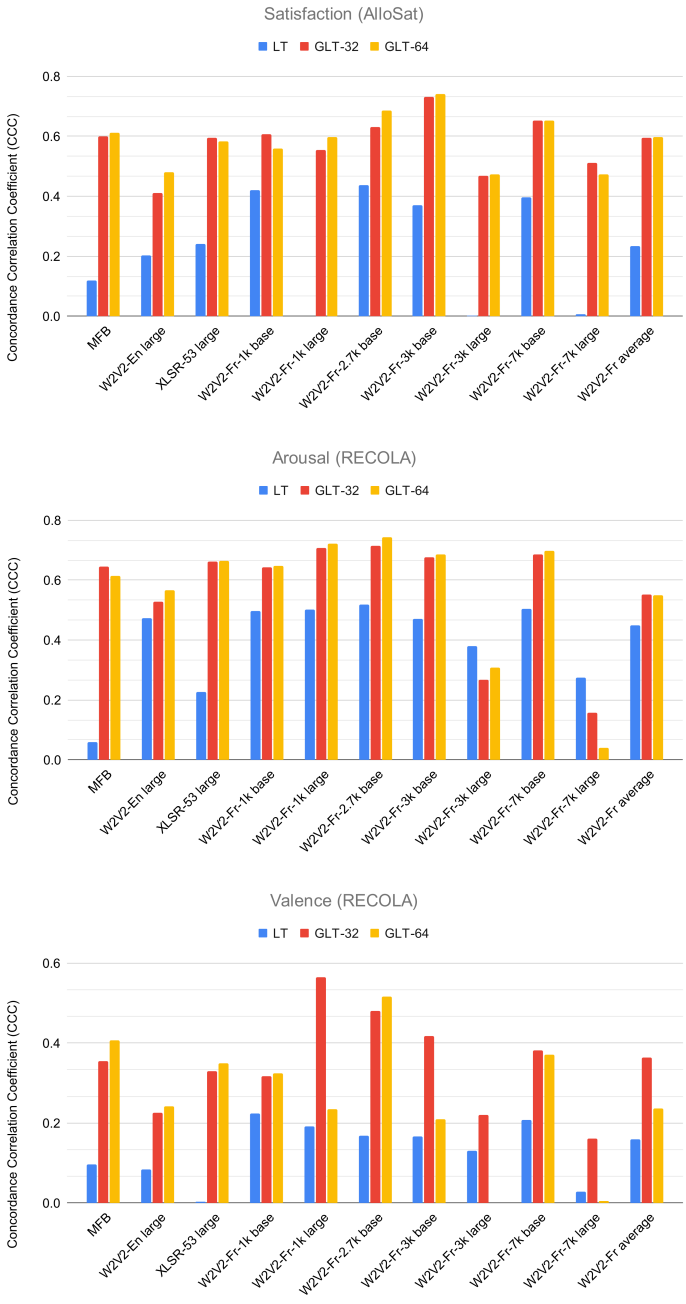


Figure 4.3: The results of arousal and valence prediction for the RECOLA dataset and frustration-satisfaction prediction for the AlloSat dataset. Here, the performance of different Wav2vec2 representations trained on different datasets is evaluated for the AER task. The AER models used here are LT: Linear-Tanh, GLT: GRU-Linear-Tanh, where the GRU is a layer with either 32 or 64 nodes.

further shows that Wav2vec2 representations are better at modelling acoustic context than MFB features. Nevertheless, for the GLT models, comparing the average of the W2V2-Fr representations with the MFB features shows that MFB features are still able to achieve comparable performance to deep representations for predicting emotion dimensions, when trained with contextual models that are sufficiently complex for a given task.

Furthermore, the XLSR and W2V2-Fr representations trained on multilingual or French speech can, on average, perform better than the W2V2-En representations trained only on English speech. This shows that **the language of the utterances used to train the Wav2vec2 representations plays a role in their ability to predict dimensional emotions**. This is not surprising as we know that the emotional expressions are language-dependent, because it is through language that the contextual meaning of the bodily sensations of the world is conceptualised (Lindquist et al., 2015).

Finally, the results of this section are compared with the state of the art. For the prediction of the frustration-satisfaction dimension of the AlloSat dataset, the best result is achieved by using the GLT-64 model and the W2V2-Fr-3k base representation (CCC=.740). This is similar to the state-of-the-art results. In Macary et al. (2021), bidirectional LSTM models are used with Wav2vec representations, achieving a CCC of .730. In a later paper (Tahon et al., 2021), the wav2vec representations were found to be more performant when computed on segments of 250 ms rather than the whole conversation, using recurrent models such as LSTM or GRU, achieving a CCC of .806.

For the arousal dimension of the RECOLA database, the best result was obtained using the W2V2-Fr-2.7k base representation and the GLT-64 model (CCC=.744). As far as we know, this result is among the best arousal results ever achieved on the RECOLA test set, being statistically[1] on-par with the performance obtained in the work of Zhang et al. (2016), who achieved a CCC of .732 using MFCCs with LSTMs. This suggests that effective prediction of arousal from acoustic signals does not require representations as complex as deep pre-trained representations, which is consistent with the results of this study. Also for the valence dimension, the best result here was obtained using the W2V2-Fr-1k large representation and the GLT-32 model (CCC=.564), which is also statistically[2] on-par with the performance obtained in the work of AlBadawy and Kim (2018), who achieved a CCC of .555 with MFBs and LSTMs.

The experiments and results of this section are summarised and further discussed in what follows.

[1] A two-tailed statistical test using the Fisher r-to-z transform, and a degree of freedom corresponding to the number of tested utterances (df=529), provides the following results: z=0.43, p=0.6672.

[2] The same statistical test as used for arousal provides the following results for valence: z=0.21, p=0.8337.

4.1.4 Discussion

In this section, the effect of training data on deep representations for the AER task was investigated. A series of experiments were then specifically designed to investigate the effect of both the amount of training data and the type of data (read versus spontaneous speech) on pre-training deep representations. The deep pre-trained representations were then exploited to predict time-continuous emotional dimensions on the RECOLA (arousal and valence dimensions) and AlloSat (frustration-satisfaction dimension) corpora. The performance of the deep representations was also compared with the performance of the traditional features (MFBs) for the same tasks.

The results of these experiments showed that pre-training deep representations with more data does not necessarily lead to better AER performance of such representations. However, the type of data could play a role, as for the task of AER from French speech, deep representations pre-trained on French speech performed better on average than deep representations pre-trained on English speech. Also, Compared to MFB features, deep representations can be used with less complex models to achieve good AER performance. Moreover, by using deep representations trained for French speech, this study could achieve the best reported performance for predicting the arousal and valence dimensions of the RECOLA dataset.

It should be noted, however, that the experiments carried out in this section are subject to certain limitations. For example, this study could not further determine the cause of the better AER performance of the deep representations pre-trained on read speech compared to deep representations pre-trained on spontaneous speech. Furthermore, the deep representations are trained on isolated utterances, whereas the experiments performed here aim at predicting dimensional emotions over time on long conversations lasting several minutes. This is because the state of the art for training deep representations tends to be training on isolated utterances, whereas the state of the art for predicting dimensional emotions tends to be training AER models and evaluating them in a time-continuous format. As this discrepancy could be a problem for a fair evaluation of the pre-trained deep representations for AER, all the following experiments in this thesis target utterance-level emotion classification tasks. Since recent state-of-the-art research also suggest that joint acoustic-textual representations can perform better than acoustic representations alone for AER, the next section describes the experiments conducted in this thesis on joint acoustic-textual representations for utterance-level emotion classification.

4.2 Joint representation of acoustic signals and text

State-of-the-art research in AER suggests that by jointly using representations of speech signals and their transcriptions, one can achieve better performance than by

using only speech or textual representations (see Section 2.2.5). The better performance of using both acoustic and textual modalities, rather than either alone, is often attributed to the incorporation of more information related to both verbal and non-verbal communication (see "Joint representations of acoustic signals and text" in Section 1.1.2). In order to investigate this phenomenon further, this section attempts to reproduce results comparable to the state of the art, before advancing it in the next Section 4.3 by exploring the use of ASR transcriptions, and in Section 4.4 by exploring the use of speaker representations.

This section first studies the effect of different hyper-parameters on pre-trained deep acoustic and textual representations separately for acted AER using the IEMOCAP dataset (see Section 4.2.1). Then, several AER model architectures are proposed and evaluated in order to obtain an effective joint acoustic-textual representations (see Section 4.2.2). IEMOCAP was chosen here because it is the most commonly used dataset for joint acoustic-textual representations (Li and Lee, 2019; Siriwardhana et al., 2020; Zhang and Xue, 2021), and thus the results in this paper could be compared with state-of-the-art results. Furthermore, in order to evaluate the performance of the proposed joint acoustic-textual representations for AER on in-the-wild emotional expressions, the CMU-MOSEI dataset is used in Section 4.2.3.

4.2.1 Deep representations of acoustic signals and text

This subsection presents a series of preliminary experiments aimed at analysing the effectiveness of deep acoustic and textual representations for AER across different model complexities and training conditions. This is the first step to see how different AER models can perform with such representations, before experimenting with joint acoustic-textual representations in Section 4.2.2. The experiments designed for this purpose are described further below in "Dataset and representations" and "Method", followed by the "Results".

Dataset and representations

To study and compare the performance of acoustic and textual representations for AER, the **IEMOCAP dataset** was chosen as it provides speech signals, their corresponding human transcriptions and emotion annotations (see Section 3.1.6). Moreover, the IEMOCAP dataset has also been the target of several state-of-the-art studies on joint acoustic-textual representations for AER. To be consistent with these studies, only the expressions labelled **anger, happiness (+excited), sadness and neutral** are included here, with sessions 1 to 3 as the training set, session 4 as the development set and session 5 as the test set. A brief summary of the number of utterances and their duration for this partitioning of the IEMOCAP dataset is provided

Table 4.2: Statistics related to the training, development and test partitions of the IEMOCAP dataset, which contains acted emotional expressions.

Dataset	Number of utterances			Duration in hours:minutes		
	Train	Dev	Test	Train	Dev	Test
IEMOCAP	3259	1031	1241	4:11	1:16	1:33

in Table 4.2. Since IEMOCAP contains English utterances, for textual representation, the RoBERTa model is used here, which is pre-trained on English text (see Section 3.2) and has shown state-of-the-art results for AER from text (Siriwardhana et al., 2020). For acoustic representations, both the traditional MFB features and the W2V2-XLSR-56 are used (see Table 3.2). The choice of the multilingual Wav2Vec2 ("W2V2-XLSR-56") was made to be consistent with the experiments carried out later in in the chapter 5, which use datasets containing speech with different languages. Also, in our preliminary experiments, the W2V2-XLSR-56 model obtained better AER results than the W2V2-XLSR-53 or the W2V2-En model. It should also be noted that **the pre-trained W2V2-XLSR-56 and RoBERTa models used here are frozen**, meaning that they are used as acoustic and textual representations, and are not further trained with the AER models.

Method

To explore the effect of different model complexities and different training hyperparameters, we experimented with the combination of the following ranges of possibilities:

- Model-[number of layers x number of nodes]: GRU-1x64[1], GRU-2x128, GRU-4x256
- Learning Rate (LR): 0.001, 0.0001
- Gradient Accumulation (GA): 1, 10, 100

The experiments here use GRU models of varying complexity because GRUs have been shown to achieve state-of-the-art (or comparable) performance for AER, whether using the MFB or Wav2vec2 representations (see Section 4.1 and Section 2.2.8). This was mainly because of the ability of GRUs to effectively model sequential data (see "Recurrent layers" in Section 2.2.2). The acoustic and textual representations here are fed into the GRU model similarly to the previous experiment (see Figure 4.2). However, since the goal in this experiment is not continuous

[1]We started by increasing the model complexity of GRU-1x64 rather than less complex models, as the results of the previous experiments in Section 4.1 showed that GRU-1x64 was mostly more performant for AER than GRU-1x32 or a linear (fully connected) classifier.

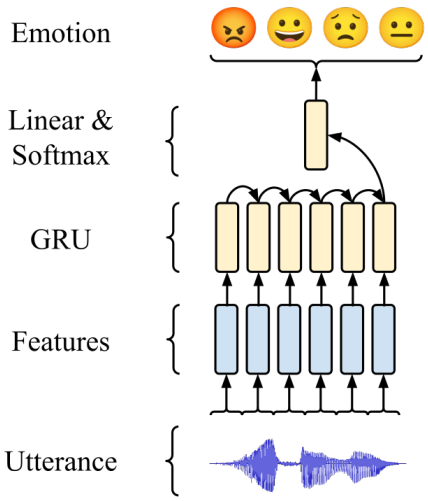


Figure 4.4: The pipeline of the method used to evaluate the deep representations for classifying emotion labels.

emotion prediction, but utterance-level emotion classification, the last output of the model is followed by a linear layer and then a Softmax layer to provide the probabilities of different emotion labels –happy, angry, sad and neutral for IEMOCAP– (see Figure 4.4). The use of attention layers instead of GRUs was also investigated, but they used more memory and did not perform as well as GRUs in our preliminary experiments (further experiments on GRUs with attention layers for AER are also done in Section 4.2.2). The GRU models were trained using the Adam optimiser, where the loss function was cross-entropy, which is a common choice for classification tasks (see Section 3.4.2 for more details). maximum number of epochs was chosen to be 250, and an early stopping of 15 epochs was applied if no improvement over the development set was observed. Also, we could not use more than one batch size due to memory constraints. However, instead of using different batch sizes, we used different GAs, which has a similar effect to varying batch sizes, but saves more memory (see Section 3.3). The results of these experiments are discussed below.

Results

The results of the experiments on the performance of acoustic and textual representations for AER are shown in Figure 4.5.

At first glance, the results show that the textual representations –RoBERTa– achieve consistently better results than the acoustic representations –XLSR-56 and MFB– across all the different setups. This suggests that **the verbal message of the IEMOCAP corpus is a more efficient channel for predicting emotion categories**, which may be due to the scripted nature of more than half of the dialogues in this dataset. In addition, upon further investigation, it was realised that there were

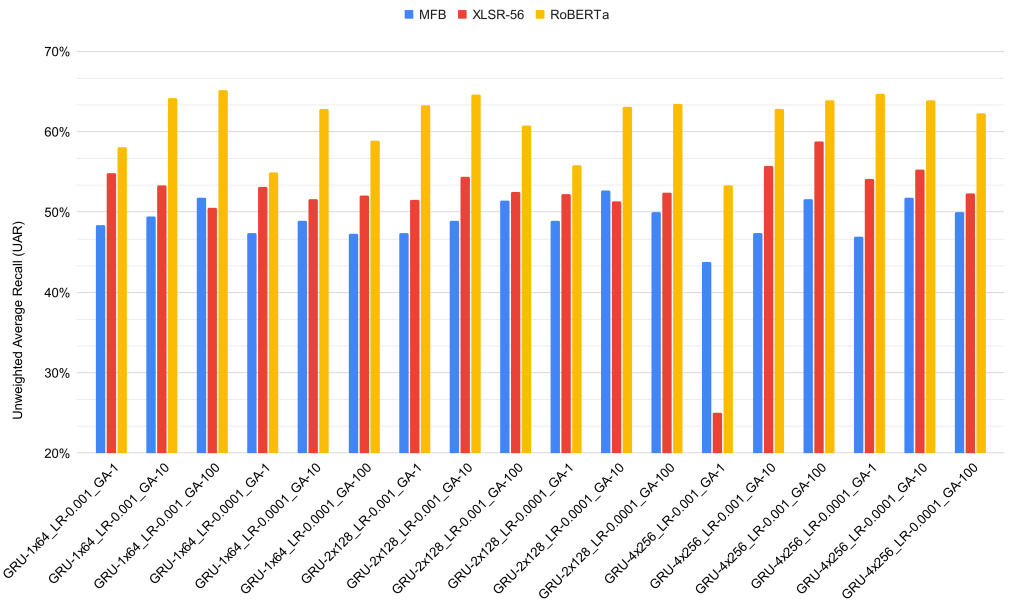


Figure 4.5: The results for emotion recognition on the IEMOCAP dataset, for different representations, models, and hyper-parameters. From left to right, the complexity of the model, and then hyper-parameters increases.

helpful tags in the transcriptions, such as a "[LAUGHTER]" tag, which may have played a role in aiding the AER from the transcriptions for the IEMOCAP dataset. Comparison of the two acoustic representations also shows that while XLSR-56 outperforms MFB in most cases, this is not always the case and MFB can perform as well or better, when having a high LR, a low GA, or a more complex model. This is in line with the results of the experiments done on the AlloSat and RECOLA datasets in Section 4.1.

Moreover, previous observations of the state of the art also suggest that, regardless of the modality used, a more complex architecture is not necessarily always more effective, and that how a model is trained may play a more important role than the model architecture (Evain et al., 2021b; Liu et al., 2019). Therefore, in order to study the effect of training hyper-parameters, LRs and GAs are the subject of this study. The averages of the results for LR-0.001 compared to LR-0.0001 show that a lower learning rate during training can be more effective for acoustic representations, but not for textual ones. This, together with the aforementioned results that textual representations can outperform acoustic ones, may suggest that textual representations are better correlates of emotion for the IEMOCAP dataset, and can therefore be exploited by training a model with a higher learning rate. Beside the learning rate, the study of the other hyper-parameter here, i.e. the GAs, shows

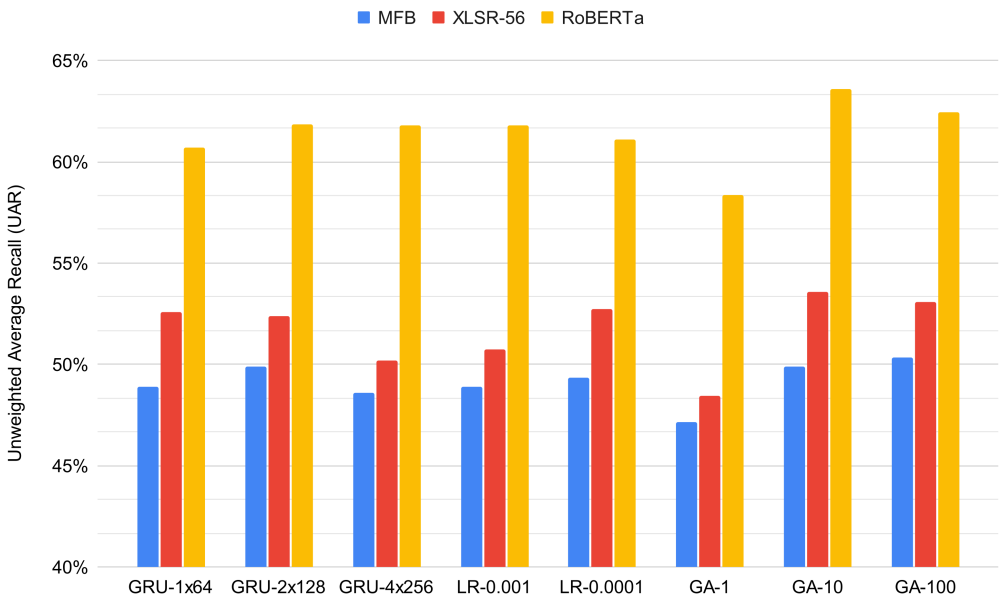


Figure 4.6: The emotion recognition results on the IEMOCAP dataset, averaged for different hyper-parameters.

that although GA-1 is not as effective for the IEMOCAP dataset as having higher GAs, there does not seem to be much difference between having 10 (GA-10) or 100 (GA-100) utterances influencing the weight update of the model for each training iteration.

The best results here for W2V2-XLSR and RoBERTa respectively are 59% and 65% in terms of UAR with GRU-4x246 and GRU-1x64 (on fixed train-dev-test partitions). These results are comparable or better than the state-of-the-art results on the IEMOCAP dataset. For example, in Li and Lee (2019), LLDs for audio and GloVe for text are used with LSTMs to achieve 57 % and 67 % UAR respectively (using 5-fold cross-validation). Also in Ho et al. (2020), MFCC and BERT representations are used with GRUs to achieve 57 % and 67 % UAR respectively (using 10-fold cross-validation). A more recent work (Cai et al., 2021) shows that by fine-tuning Wav2vec2 representations, and also training them for both AER and ASR at the same time, one can achieve 78 % accuracy (UAR not reported) by using 10-fold cross-validation.

The above results suggest that for IEMOCAP, the use of both acoustic signals and their transcriptions can be beneficial for AER. Therefore, the use of both acoustic and textual modalities can be expected to more efficiently capture both types of verbal and non-verbal communication. Therefore, by using acoustic and textual modalities, one can expect to more efficiently capture both types of verbal and non-

verbal communication. The following presents the experiments conducted in this thesis to obtain an efficient joint acoustic-textual representation.

4.2.2 Joint acoustic-textual representations

Acoustic and textual information can be fused at the input, decision or model level. Input-level fusion refers to techniques where the representations of different modalities can be concatenated at the signal level. Input-level fusion is technically difficult to implement for acoustic and textual representations because the representations of acoustic signals are based on audio frames, whereas the textual representations are based on tokens (see Section 2.2.5). On the other hand, decision-level fusion refers to training different models for different modalities in complete isolation from each other, but the final decision is made based on the decisions of each isolated model. In this way, the information from the acoustic and textual modalities may not really be used in a complementary way and therefore may not be in line with the aim of this section. Finally, model-level fusion refers to techniques where the representations of different modalities are concatenated in a latent space, which may be technically easier to implement than signal-level fusion, while also allowing the fusion of acoustic and textual information before producing the output. In addition, a comparison of the different fusion strategies (at input, model, or decision level), for acoustic and textual modalities in the state of the art, has shown no significant difference in performance for AER (Atmaja et al., 2022). Nevertheless, **the model level fusion strategy is chosen here, because it allows the integration of different sources of information in the latent space, where they may be inherently different at the signal-level.** In the following, the experiments carried out with the model-level fusion of acoustic and textual representations are explained in detail.

Experiments

Figure 4.7 depicts the AER model used here, which fuses acoustic and textual representations at the model-level. The acoustic and textual representations used here are **pre-trained (frozen) W2V2-XLSR-56 and RoBERTa representations**, similar to previous experiments in Section 4.2.1. In addition, the use of attention mechanism was also explored (see Figure 2.6), in order to further investigate whether there are emotionally salient parts, either in the sequence of data unrolled in time (*Sequential Attention* in Figure 4.7), or across different modalities for different utterances (*Modality Attention* in Figure 4.7).

The training strategy used here is similar to the previous hyper-parameters experiments on the IEMOCAP dataset, which was discussed in Section 4.2.1. Therefore, based on the hyper-parameters experiments, here the GRU-1x64 model is used because it could achieve comparable results to state of the art (see Figure 4.6), while being the simplest architecture with a faster training convergence compared to using

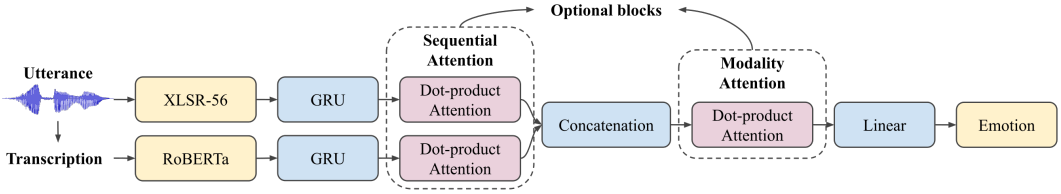


Figure 4.7: The joint acoustic-textual AER model. The *Sequential Attention* is applied to the sequence of outputs of the GRU model for acoustic and textual representations in order to find salient parts of the sequence. On the other hand, *Modality Attention* is applied to the latent representation of each modality in order to put more focus on an acoustic or textual modality, depending on a given input utterance.

more layers. In fact, GRU-1x64 was more than four times faster than GRU-4x256 for both training and inference. Furthermore, the LR and GA were chosen to be 0.0001 and 100 respectively, which gave the best results. The maximum number of epochs was also chosen as 50, because in the previous experiment with the same **model (GRU-1x64) and dataset (IEMOCAP)**, convergence was achieved in less than 50 epochs. The results of these experiments are discussed below.

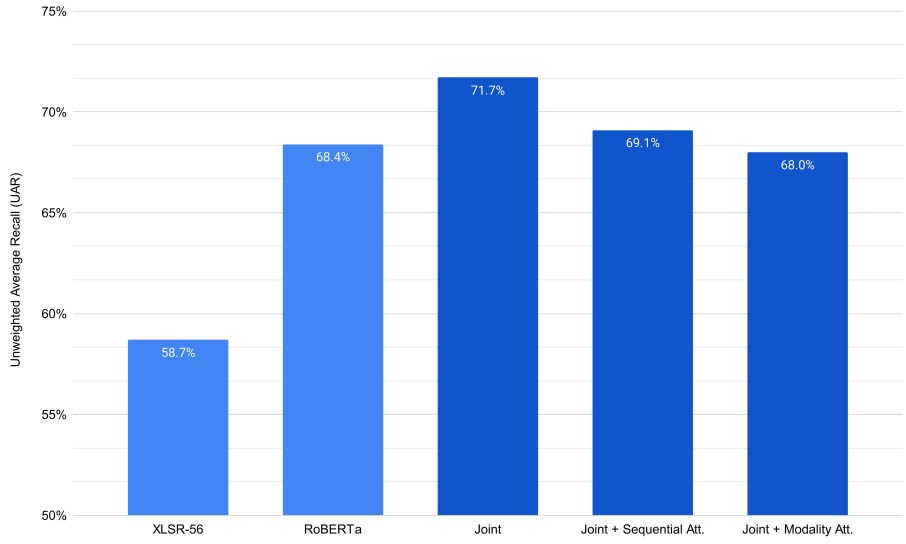


Figure 4.8: Emotion recognition performance comparison of different strategies for joint representation of acoustics and text, as well as, using the representation of each modality alone, on the IEMOCAP dataset.

Results

The results presented in Figure 4.8 show that **by using both acoustic and textual representations, an AER system can better distinguish acted emotional expressions from the IEMOCAP dataset than by using the representation of each modality alone**, which is in line with other studies (Li and Lee, 2019; Ho et al., 2020; Siriwardhana et al., 2020). Also, the use of dot-product attention, either for the textual and acoustic sequences, or on the latent joint representation space does not further improve the results. Although it is difficult to pinpoint the reason without further experiments, but this may be because the GRU is already sufficient to model the context of the data, and thus the use of attention vectors complicates the training process by adding extra trainable parameters (see "Training neural networks" in Section 2.2.2). Nevertheless, both the Wav2Vec2 and RoBERTa representations use attention-based models (via transformers) to achieve state-of-the-art deep representations of acoustics and text. However, **the use of attention layers at the back-end (for the emotion classification task), where there may not be enough training data, does not seem to be beneficial**. This is in line with some of the previous studies, such as Lieskovská et al. (2021), which show that the benefits of using the attention mechanism is not always clear and it is not an essential element for achieving state-of-the-art results for AER.

Furthermore, the results obtained here are comparable with the state-of-the-art results on joint acoustic-textual representations (see Table 4.3). For example, in Li and Lee (2019), LLDs are used as acoustic features and GloVe as textual representations with LSTM models for AER on IEMOCAP. However, for the joint representations, they also apply attention with "personal embeddings" obtained through a "Linguistic Inquiry Word Count (LIWC)", which consists of 64 semantic word categories correlated with what often concerns individuals, as well as spoken patterns. This further demonstrates the effectiveness of using personal embeddings, which are further explored in Section 4.4 of this thesis. In another similar work, (Ho et al., 2020), a multi-head attention mechanism is used with GRUs, and has been shown to achieve better performance than the method proposed here. This is despite the fact that the use of each modality alone in the work of Ho et al. (2020) does not yield better performance than the proposed method here. This may be because here the hyper-parameters were optimised for each modality rather than the joint acoustic-textual representations. In more recent work, Siriwardhana et al. (2020) shows that by jointly fine-tuning[1] the pre-trained deep representations of Wav2vec2 and RoBERTa, state-of-the-art performance can be achieved on a given dataset. This is because the fine-tuning process further adapts the deep representations to the specific dataset being evaluated. However, fine-tuning deep repre-

[1]Fine-tuning here refers to the continued training of deep representations, such as Wav2vec2 or RoBERTa, alongside the AER model.

Table 4.3: Emotion recognition performance on the IEMOCAP dataset, based on acoustic, textual or joint acoustic-textual representations using various methods based on this work and the state of the art.

Method	Audio	Text	Joint
LLD & GloVe + LSTM + personal attention (Li and Lee, 2019)	57.1%	67.1%	70.3%
MFCC & BERT + GRU + multi-head attention (Ho et al., 2020)	55.9%	67.2%	73.2%
VQ-Wav2Vec & RoBERTa + Linear (Siriwardhana et al., 2020)	-	-	**75.5%**
W2V2-XLSR-56 & RoBERTa + GRU (this work)	**58.7%**	**68.4%**	71.7%

sentations may come at the cost of reduced performance compared to using only pre-trained deep representations in a cross-corpus setting (see Section 5.1).

So far, the experiments in this section have only been based on the IEMOCAP dataset, which contains acted emotional expressions recorded in a laboratory environment. Since one of the main goals of this thesis is to study the effect of different deep pre-trained representations on emotional expressions in the wild, the following experiments are presented on the CMU-MOSEI dataset, which contains Youtube videos recorded with different speakers, microphones, and in different environments (see Section 3.1).

4.2.3 Joint representations for emotion recognition in the wild

The aim here is to see the effect of joint acoustic-textual representations for emotional expressions in the wild. The experiments and results are explained below.

Experiments

The experiments conducted here are similar to the previous experiments on the IEMOCAP dataset, with the base architecture and without the use of any attention layer. But here the experiments are on the **CMU-MOSEI dataset** (see Section 3.1.2). A statistical summary of this dataset is provided in Table 4.4. Similar to previous experiments, **pre-trained (frozen) W2V2-XLSR-56 and RoBERTa representations are used as acoustic and textual representations** for the experiments in this subsection. Three experiments are performed here, consisting in the use of joint acoustic-textual representations, acoustic representations alone, and textual representations alone. In order to compare our results with the state of the art, the sentiment dimension is used[1], which is considered as a continuous number ranging

[1]The experiments carried out here for the CMU-MOSEI dataset are only with the sentiment dimension and not with the emotion categories, because sentiment annotations seemed to be more reliable, as our preliminary experiments with the emotion categories showed rather low results (40 % for seven classes with RoBERTa and chance level with W2V2-XLSR-56) and were therefore not considered reliable for further studies.

Table 4.4: Statistics related to the training, development and test partitions of the CMU-MOSEI datasets, which contains in-the-wild emotional expressions.

Dataset	Number of utterances			Duration in hours		
	Train	Dev	Test	Train	Dev	Test
CMU-MOSEI	18,542	1377	3340	38	3	8

from -3 to +3. The sentiment values are then used in two forms: 1) to classify the sentiment value for less than zero or equal and greater than zero in a two-category classification paradigm, and 2) to predict continuously in a regression paradigm by simply removing the Softmax layer at the end and using MSE as the loss function, to follow other works, such as Siriwardhana et al. (2020); Sun et al. (2020). All the other parameters are the same as the previous experiment (see Section 4.2.2). The results of the experiments are discussed below.

Results

The results of this study, alongside the state of the art, is presented in Table 4.5. The results of the experiments here, together with the experiments from (Sun et al., 2020), suggest that the **acoustic representations (whether LLDs or W2V2-XLSR-56) can not work well for classification of the low and high sentiment values for expressions in the wild**. The ineffectiveness of acoustic representations for AER in the wild and not acted emotion recognition, may be because the acoustic representations used here (W2V2-XLSR-56) are trained for read speech in no-noise environments, and fail to work well on different noisy environments and spontaneous speech existing in the CMU-MOSEI dataset. Nevertheless, the W2V2-XLSR-56 representations seem to work much better compared to LLDs for continuous prediction of the sentiment dimension.

On the other hand, the **textual representations of BERT and RoBERTa seem to be able to provide effective representations for sentiment classification in the wild**. The better performance of textual representations such as RoBERTa compared to acoustic representations such as W2V2-XLSR-56 may be due to the fact that textual representations are not affected by the environmental noise that affects a recorded audio or the way a sentence is uttered. Moreover, as the aim here is sentiment, which can be related to valence, since both are usually annotated on a scale from negative to positive. Thus, it can be argued that the sentiment dimension is easier to predict from text than from speech, since arousal is usually attributed to vocal parameters, whereas valence is often influenced by linguistic information (Goudbeek and Scherer, 2010). For example, in Triantafyllopoulos et al. (2023), it is shown on two acted datasets of IEMOCAP and MSP-Podcast that the arousal dimension achieves significantly higher CCC using acoustic rather than textual features, and vice versa, the valence dimension achieves higher CCC using textual

Table 4.5: Emotion recognition performance on the CMU-MOSEI dataset, comparing the use of acoustic, textual and joint acoustic-textual representations. The best results for each column are shown in bold. The Mean Absolute Error (MAE) is the result of predicting the sentiment dimension continuously (from -3 to +3), and the UAR/Accuracy is the result of classifying low (below zero) and high (equal or above zero) sentiment values.

Method	MAE	UAR / Accuracy
LLD of audio + LSTM (Sun et al., 2020)	1.430	45.1% (Accuracy)
BERT + LSTM (Sun et al., 2020)	0.897	80.8% (Accuracy)
LLD of audio & BERT + LSTM (Sun et al., 2020)	0.909	80.6% (Accuracy)
W2V2-VQ & RoBERTa + Linear (Siriwardhana et al., 2020)	0.577	**88.3**% (Accuracy)
W2V2-XLSR-56 + GRU (this work)	0.680	57.3% (UAR)
RoBERTa + GRU (this work)	**0.531**	72.3% (UAR)
W2V2-XLSR-56 & RoBERTa + GRU (this work)	0.542	72.2% (UAR)

rather than acoustic features.

Furthermore, despite the previous results obtained in section 4.2.2, the use of **joint acoustic-textual representations does not perform better for AER in the wild than the use of textual representations**. This may be because acoustic representations here are not nearly as useful as textual representations for classifying low and high sentiments in the wild, and therefore using acoustic representations in addition to textual representations may increase the complexity of training a model rather than provide additional useful information (see "Training neural networks" in Section 2.2.2).

In this section, several experiments were conducted and analysed to show the effectiveness of joint acoustic-textual representations for acted emotion recognition on the IEMOCAP corpus and AER in the wild on the CMU-MOSEI corpus. The results showed that although joint acoustic-textual representations can perform significantly better than either modality alone for acted emotional expressions, this better performance of joint acoustic-textual representations does not necessarily extend to predicting emotional expressions in the wild. These findings are discussed further below.

4.2.4 Discussion

The experiments in this section started by analysing different model complexities and training hyper-parameters for training AER models for acoustic or textual modalities alone. The results showed that even a simple model (here GRU-1x64) can be effectively trained for AER on deep pre-trained acoustic (W2V2-XLSR) and textual (RoBERTa) representations. Next, a method was proposed to fuse the acoustic and textual representations at the model level using the dot-product attention

mechanism applied to either the sequence or latent representation of the two modalities. However, the results showed that simply concatenating the acoustic and textual representations at the model level is effective for AER on acted emotional expressions using the IEMOCAP dataset. It was also shown that the joint acoustic-textual representations obtained in this way can perform better than using either acoustic or textual modalities alone. The performance of the joint acoustic-textual representations was then further evaluated on the CMU-MOSEI dataset, which consists of emotional expressions in the wild. The results showed that the joint acoustic-textual representations did not perform better than just using the textual representations for continuous prediction of the sentiment dimension or for detecting low and high sentiment values. This was mainly because the acoustic representations could not effectively predict the sentiment dimension. This may be because the W2V2-XLSR acoustic representations used here were trained on read speech in noiseless environments, which is different from the recordings in the CMU-MOSEI dataset.

However, it should be noted that the study of AER in the wild is limited here to predicting the sentiment dimension and detecting low and high sentiment values from the CMU-MOSEI dataset. Although the CMU-MOSEI dataset also provides emotion categories (see Section 3.1.2), the results of our preliminary experiments on the emotion categories were too small to draw any conclusions and were therefore not considered in this thesis. The work of Sun et al. (2020) for the classification of the seven basic emotions of the CMU-MOSEI dataset also showed very low accuracies of 16.2%, 35.9%, 35.08% for acoustic (LLDs), textual (BERT) and acoustic+textual representations respectively. Nevertheless, these results on the emotion categories are also in line with the results on the sentiment dimension of the CMU-MOSEI dataset, which might suggest that the AER from acoustic representations is not yet ready to be used on emotional expressions observed in the wild. On the other hand, textual representations have been shown to be capable of reasonable performance for AER in the wild. Therefore, the state of the art in recent years has been to first try to obtain textual transcriptions of a speech signal by exploiting an ASR system, and then to try to use AER models trained on textual representations. This is further explored in the next section of this chapter.

4.3 Exploiting automatic speech recognition

Joint acoustic-textual representations have been shown to be more effective than using acoustic representations alone for AER of acoustic signals, both for acted emotional expressions and for recordings made in the wild. As shown in the previous section (see Section 4.2), the effectiveness of joint acoustic-textual representations is mainly due to the use of textual representations, especially for noisy acoustic signals in the wild, where the verbal message may be a more effective source of information for AER. However, as human transcriptions are not always available,

ASR models are often used to provide us with transcriptions. These transcriptions can later be used to obtain the joint acoustic-textual representations according to the model presented in Section 4.2.2 (see the model in Figure 4.7). The use of joint acoustic-textual representations, where the text is extracted from ASR models, has been shown to perform better than using only the acoustic representations for AER on acted data (Heusser et al., 2019; Yoon et al., 2019; Wu et al., 2021; Peng et al., 2021). However, the state of the art has not yet explored the effectiveness of this method for emotional expression in the wild. Therefore, in this section, several experiments are conducted to investigate the effect of using ASR transcriptions for joint acoustic-textual representations on AER for both acted and in-the-wild emotional expressions. These experiments are described in more detail below, followed by the results and a brief discussion.

4.3.1 Experiments

The experiments in this section use the two **IEMOCAP and CMU-MOSEI datasets** from the experiments in section 4.2.2 and section 4.2.3, with the same **GRU-1x64 model** and the same training strategy of LR=0.0001 and GA=100. The only difference from the previous experiments is that Google's ASR[1] is used to transcribe the utterances for the IEMOCAP and CMU-MOSEI datasets, because the goal here is to evaluate the effectiveness of ASR transcriptions, instead of using human transcriptions. Also, Google's ASR was chosen because it provides reliable transcriptions in many languages and is easy to use. Experiments will then be carried out on ASR-based textual representations and joint acoustic-textual representations based on ASR transcriptions.

4.3.2 Results

First, the Word Error Rate (WER) of Google's ASR with respect to human transcriptions was calculated by counting the number of substitutions, deletions and insertions over the total number of words per phrase (see Section 3.4.4). The results are shown in Figure 4.9. As can be seen from the figure, the WERs are quite high for the CMU-MOSEI dataset, but even higher for the IEMOCAP dataset, which may make the use of it for AER on the IEMOCAP dataset rather ineffective.

The results of the AER experiments are shown in Figure 4.10. For the IEMOCAP dataset, we can observe a large drop in performance by using the ASR transcriptions compared to using the human transcriptions. This was mainly because the ASR transcriptions did not pick up on all the words in many cases, which might

[1]Here the SpeechRecognition python wrapper was used (https://pypi.org/project/SpeechRecognition), which uses the Google Cloud speech-to-sext APIs (https://cloud.google.com/speech-to-text).

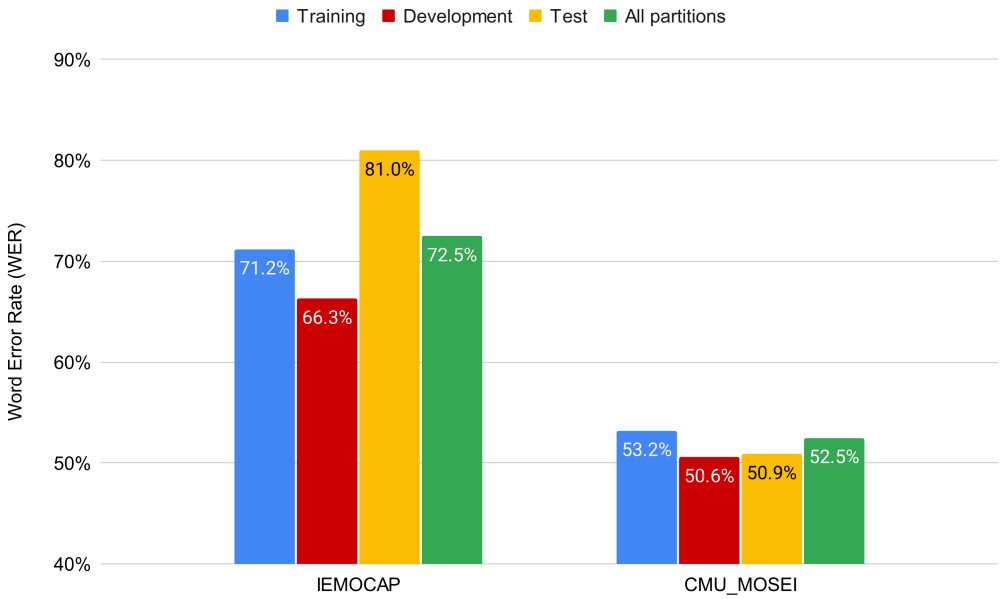


Figure 4.9: The WER of Google's ASR with respect to human transcriptions, calculated for the IEMOCAP and CMU-MOSEI datasets.

have been associated with certain emotions. For example, an utterance labelled as "anger" with the human transcription of "Adders don't snap, they sting.' was automatically transcribed as "snappy today", which loses it's meaning. Moreover, the IEMOCAP transcriptions contain "[LAUGHTER]" tags, which can further help the AER from text, especially for recognition of "happy" expressions. Despite the poor performance of the ASR transcriptions compared to the human transcriptions, using the ASR transcriptions to provide us with the joint acoustic-textual representations still performs better than using only the acoustic representations. This suggests that **even a hint of the verbal message, in this case around one word out of five, can help the AER from acoustic signals.**

On the other hand, the AER results for the CMU-MOSEI dataset show little difference between using the ASR transcriptions or the human transcriptions as input to RoBERTa. This is despite the fact that there is a 51 % WER for the ASR transcriptions, with respect to the human transcriptions[1]. This result suggests that **using ASR to obtain textual representations can be effective for AER in the wild, even if only half of the uttered words are detected**. Moreover, using only ASR-based RoBERTa representations shows a significant improvement (below 0.0001 using p-value test) over W2V2-XLSR-56 representations. Also, the use of joint W2V2-

[1]Methodologically, only Youtube videos with manual transcriptions provided by the uploader were collected for the CMU-MOSEI dataset (Zadeh et al., 2018).

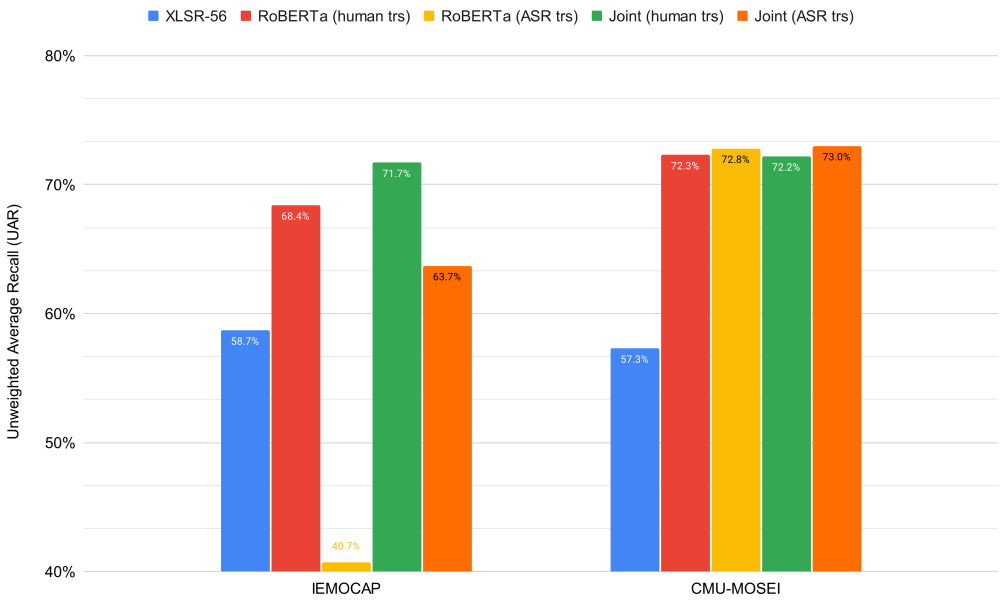


Figure 4.10: The comparison of using ASR transcriptions versus human transcriptions on the performance of AER from joint acoustic-textual representations on the IEMOCAP and CMU-MOSEI corpora.

RoBERTa representations does not significantly improve the AER results compared to the use of RoBERTa representations alone, which is consistent with previous results obtained in Section 4.2.3. The results of this section are summarised and discussed below.

4.3.3 Discussion

The aim of this section was to evaluate the effectiveness of using ASR models to provide us with transcriptions that can later be used in a joint acoustic-textual representation to improve AER from acoustic representations. The experiments conducted here, on the IEMOCAP and CMU-MOSEI datasets, have shown that this method can improve the performance of AER from acoustic representations for both acted and in-the-wild emotional expressions, demonstrating the importance of considering the verbal message alongside the acoustic changes of the signal. However, this improvement in performance for in-the-wild emotional expressions was shown to be mainly due to the use of the verbal message, as the W2V2-XLSR-56 acoustic representations and LLDs were experimentally evaluated as not being robust enough to be used reliably for in-the-wild emotional expressions.

The experiments so far in this chapter have shown that the joint acoustic-textual

representations can achieve better AER performance than the acoustic representations alone. Furthermore, since different people express the same emotion in different ways, recent studies have shown that incorporating speaker information into acoustic representations can also improve AER performance (Pappagari et al., 2020; Assunção et al., 2020; Ta et al., 2022). Therefore, in the next section, the effect of integrating speaker representations into joint acoustic-textual representations will be investigated.

4.4 Speaker-aware deep representations

One of the main aims of this thesis is to study the effect of speaker-aware joint acoustic-textual representations on AER, where the text can be either human transcription or generated by an ASR (see Section 1.2.1). Recent related state of the art has shown that the use of pre-trained speaker representations, which can be latent representations from a speaker recognition model, can be used in a model-level fusion strategy (see Section 4.2.2) to improve the performance of AER from acoustic signals compared to using only the acoustic representations (Ta et al., 2022). Since a model-level fusion strategy can also be effective for fusing both acoustic and textual representations (see Section 4.2), this section advances the state of the art by investigating the effectiveness of a model-level fusion strategy for acoustic, textual, and speaker representations (see Figure 4.11). To achieve this goal, this section first studies the effectiveness of model-level fusion of speaker representations based on self-supervised representations in Section 4.4.1, and then evaluates this method for joint acoustic-textual representations in Section 4.4.2.

4.4.1 Speaker-aware acoustic representations

In what follows, the dataset and representations are first explained, followed by the method and results.

Dataset and representation

Here, the IEMOCAP dataset is used for the experiments, since the related state of the art has also focused on this dataset, and also to be consistent with the previous experiments in Section 4.3. However, the partitioning of the IEMOCAP dataset used here differs from the partitioning used in the previous experiments. This is because in the previous experiments the task was speaker-independent emotion recognition, whereas **here the goal is speaker recognition**, which is not possible with speaker-independent partitioning. Therefore, here **all IEMOCAP utterances are randomly assigned, considering a 70 %-15 %-15 % distribution for training,**

Table 4.6: Statistics related to the training, development and test partitions of the IEMOCAP dataset used for the speaker-aware acoustic representations experiments. Sessions 1-5 of the IEMOCAP dataset are randomly assigned, taking into account a 70 %-15 %-15 % distribution for training, development and test partitions, and ensuring that the same speaker appears in all partitions.

Dataset	Target	Number of utterances			Duration in hours:minutes		
		Train	Dev	Test	Train	Dev	Test
IEMOCAP	**Speaker**	3871	830	830	4:56	0:57	1:06

development and test partitions, with all speakers involved in all partitions. A statistical summary of this partitioning is given in Table 4.6.

Regarding the representation employed in this subsection, **only acoustic W2V2-XLSR-56 is used** (following previous sections), since the focus of the study here is on the training strategy and not on different modalities. This study is further extended for both acoustic (W2V2-XLSR-56) and textual (RoBERTa) modalities in Section 4.4.2.

Method

The proposed method for speaker-aware emotion recognition is shown in Figure 4.11. **The speaker and emotion recognition models are the same model (GRU-1x64)** used in Section 4.2, which has already been shown to be able to effectively categorise different emotions. In this subsection, different strategies for training the speaker and emotion recognition models are investigated.

Furthermore, to train the model shown in Figure 4.11, three different training strategies are investigated:

1. **Separate training**: Here we first train the speaker recognition model, and then use the last vector in the output sequence of the trained GRUs to provide us with speaker representation vectors. The speaker representations are then concatenated with the acoustic and textual latent representations to train the AER model.

2. **Separate training with fine-tuning**: Here, the same separate training strategy as above is applied, but with the difference that we allow the W2V2-XLSR-56 weights for the speaker recognition model to be updated (i.e. "fine-tuned") in order to further adapt the W2V2-XLSR-56 model to the speakers of the IEMOCAP dataset and thus obtain a more pertinent learnt speaker representation.

3. **Joint training**: Here the speaker and emotion recognition models are trained jointly in a stepwise fashion, i.e. for each training iteration there are two

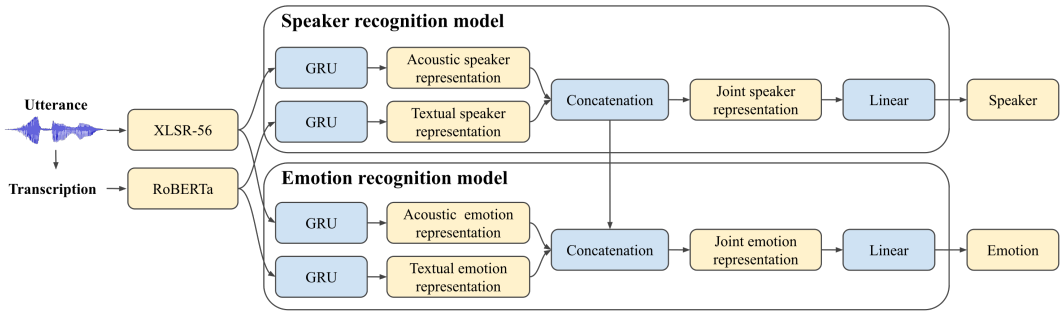


Figure 4.11: The proposed model for speaker-aware emotion recognition based on joint acoustic-textual representations. The speaker recognition model can be trained separately or together with the emotion recognition model.

steps: 1) a forward pass to compute the loss and gradients for the speaker model based on the input and update its weights when it has reached the GA cap (here 100 iterations), and 2) a forward pass to compute the loss and gradients for the emotion model based on the input and update its weights when it has reached the GA cap. The joint training strategy when having multiple tasks is a common practice in deep learning, and it has shown to obtain state-of-the-art results in similar works such as Moine et al. (2021); Ta et al. (2022).

All the experiments in this subsection are performed on the IEMOCAP corpus with W2V2-XLSR-56 representations, the GRU-1x64 model, LR of 0.0001 and GA of 100, similar to the experiments in Section 4.3. In the following, the results of the different training strategies for speaker-aware acoustic representations are presented.

Results

The results of the experiments described above are shown in Figure 4.12. The results show that **the use of speaker representations can improve the performance of AER from acoustic representations, for both separate and joint training strategies**. This demonstrates the effectiveness of integrating speaker information for AER from acoustic signals. The joint training strategy also performs better than training the speaker and emotion recognition models separately, because in joint training the speaker recognition model also learns to produce more relevant latent representations for emotion recognition. It should be noted that although joint training achieves the best performance, it cannot be used in speaker-independent emotion recognition paradigms, where the speakers are assumed to be different for different partitions. Furthermore, the results show that the separate training with fine-tuning of the W2V2-XLSR-56 weights for the speaker recognition model does

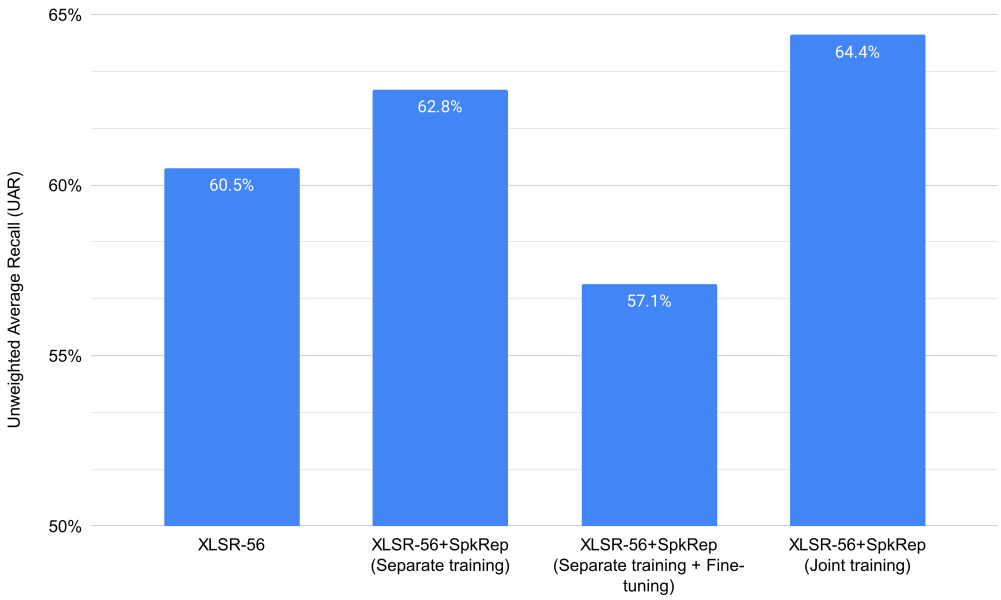


Figure 4.12: Performance comparison of different strategies for the proposed speaker-aware emotion recognition model. This model is trained using three different strategies: 1) separate training, where a speaker recognition model is trained separately to provide speaker representations (SpkRep), 2) separate training with fine-tuning, where the W2V2-XLSR-56 weights for the speaker recognition model are fine-tuned during training, and 3) joint training of the speaker and emotion recognition models.

not seem to be effective. This may be due to the fact that the number of instances for the IEMOCAP dataset is too small to further train the large number of parameters used in the Wav2vec2 architecture, which would result in a lower generalisability of the speaker representations when fine-tuning the pre-trained W2V2-XLSR-56 model for the speaker recognition task on the IEMOCAP dataset (This is further studied in Section 5.1).

Here, different training strategies have been evaluated for speaker-aware AER from acoustic representations; in the following, these experiments are further extended to include textual representations, either from human or ASR-based transcriptions.

4.4.2 Speaker-aware joint acoustic-textual representations

The aim of this section is to evaluate the effectiveness of using speaker information to improve the performance of AER from joint acoustic-textual representations. In

section 4.4.1 it was discussed that speaker information can be obtained from the latent representations of a speaker recognition model, which can be further fused with acoustic representations to better recognise different emotion categories compared to using acoustic representations alone (see Figure 4.11). Following the method proposed in 4.4.1, and according to the aforementioned aim of this section, several experiments are carried out to first compare the effectiveness of acoustic, textual and joint acoustic-textual representations in a speaker recognition model. Then, the latent speaker representations of the speaker recognition models are fused with the representations of different modalities, which are then called speaker-aware representations. Further experiments are then conducted to evaluate the performance of the speaker-aware acoustic, textual and joint acoustic-textual representations for the AER task. The dataset, representations, the method, and the results are presented below.

Dataset and representations

Despite the previous experiments in Section 4.4.1, in this subsection the **IEMOCAP dataset** is partitioned differently for the speaker and emotion recognition models. This is because training a speaker recognition model requires the same speakers to be present in different partitions. On the other hand, the goal here is to train speaker-independent emotion recognition models, which requires the speakers to be unique for each partition. Given these criteria, the partitioning of the IEMOCAP dataset for speaker and emotion recognition models is defined as follows (see Table 4.7 for a statistical summary):

- **Regarding the partitioning of the speaker recognition models**, sessions one to four of the IEMOCAP dataset are randomly assigned to training, development and test partitions with 70 %-15 %-15 % distributions for each partition respectively, leaving the speakers of session five of the IEMOCAP dataset completely unseen by the speaker recognition model. Session five of the IEMOCAP dataset was intentionally omitted so that the latent speaker representations would not be influenced by the speakers of this session, as this session is considered as the test set for the AER models.

- **Regarding the partitioning of the emotion recognition models**, sessions one to three are considered as the training set, session four as the development set, and session five as the test. This is consistent with the partitioning used for experiments in Section 4.2 and Section 4.3, which enables further comparisons of the effectiveness of fusing speaker representations for AER from joint acoustic-textual representations.

Also, the main acoustic and textual representations used here are **W2V2-XLSR-56 and RoBERTa**, following the previous experiments. In addition to the W2V2-

Table 4.7: Statistics related to the training, development and test partitions of the IEMOCAP dataset used for the speaker-aware joint acoustic-textual representations experiments.

Dataset	Target	Number of utterances			Duration in hours:minutes		
		Train	Dev	Test	Train	Dev	Test
IEMOCAP	**Speaker**	3003	643	644	3:48	0:51	0:48
IEMOCAP	**Emotion**	3259	1031	1241	4:11	1:16	1:33

XLSR-56 representations, the **encoder representations of the Whisper model** (see Section 3.2.3) are also used as acoustic representations. As the Whisper model is partly trained on ASR, where the speaker tags were often found in the transcriptions, it was therefore indirectly trained to recognise different speakers (Radford et al., 2022). Since Whisper has been trained to predict both spoken words and speaker identities, it is hypothesised here that the corresponding representations should be aware of both the verbal content and the speaker peculiarities, and therefore its performance should be on a par with the speaker-aware joint W2V2-RoBERTa representations.

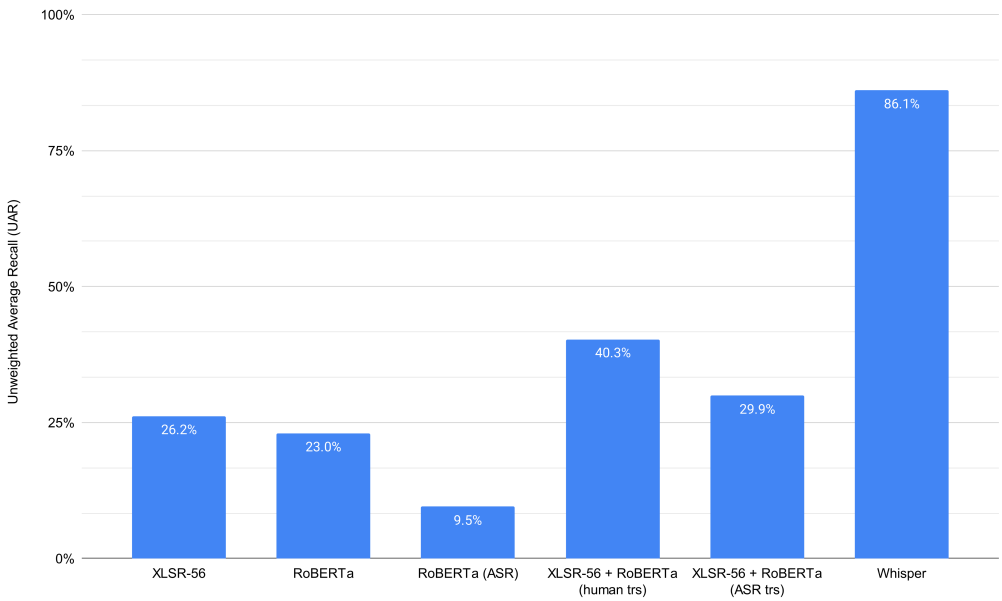


Figure 4.13: The results for the speaker recognition model based on the eight speakers from sessions one to four of the IEMOCAP dataset. The model uses GRU-1x64 with W2V2-XLSR-56, RoBERTa, and Whisper representations. Also, the RoBERTa representations are computed based on either human or ASR transcriptions.

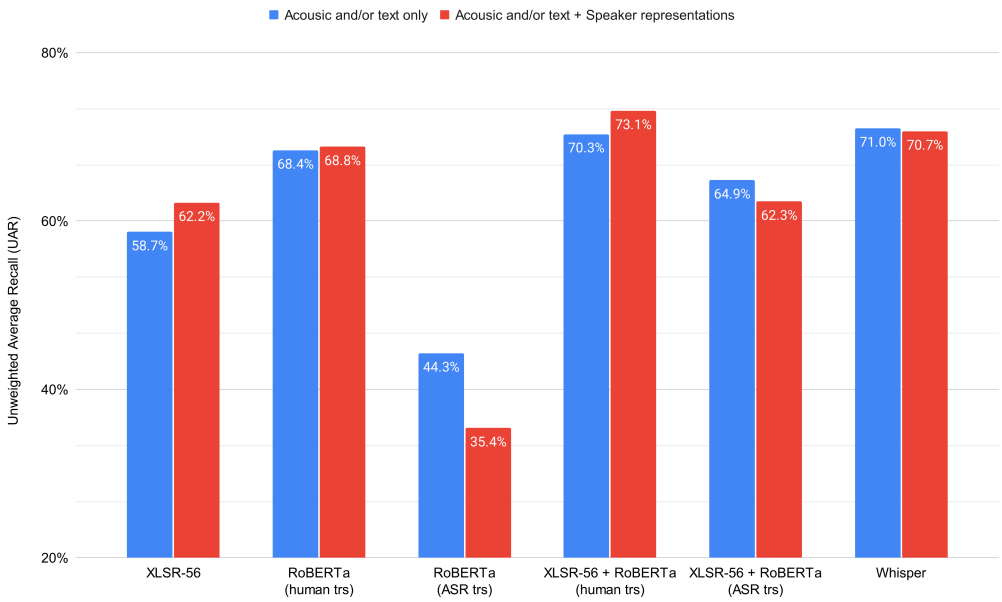


Figure 4.14: The results for the emotion recognition model based on W2V2-XLSR-56, RoBERTa and Whisper representations as baseline and their fusion with their respective speaker representations.

Method

The experiments conducted here are based on the separate training strategy of the method proposed in the Section 4.4.1 (see Figure 4.11). This method is used as described in the previous experiments, with a **GRU-1x64 model**, which is trained with LR of 0.0001, and GA of 100. The results of the experiments in this subsection are presented below.

Results

The results of the speaker recognition model for the W2V2-XLSR-56, RoBERTa and Whisper representations are depicted in Figure 4.13. The results show that the **Whisper representations can distinguish different speakers with a significant superior performance compared to W2V2-XLSR-56, RoBERTa and joint W2V2-RoBERTa representations**. This may be because the Whisper encoders have been indirectly trained for the task of speaker identification and are therefore able to provide representations that can better distinguish between different speakers (see the "Experiments" section above). Moreover, a comparison between the W2V2-XLSR-56 and RoBERTa representations shows that deep acoustic representations are marginally (but not significantly) more effective for speaker recogni-

tion. This may be due to the fact that acoustic representations contain correlates of speaker pitch, whereas the use of RoBERTa representations in a speaker recognition model can only correlate the use of specific words and phrases to different speakers. This explanation is further supported by comparing the results of the RoBERTa representations for human and ASR transcriptions, where the RoBERTa representations of ASR transcriptions achieve chance-level accuracy for the speaker recognition task on the IEMOCAP dataset. This is not surprising as the WER of the automatic transcriptions, with respect to human transcriptions, was very high for the IEMOCAP dataset (see Figure 4.9). Furthermore, the ineffectiveness of using transcriptions may be due to the fact that more than 50% of the utterances in the IEMOCAP dataset are scripted, where the choice of words might not correlate with the speakers (actors) who utter them. Nevertheless, the results further show that **using joint acoustic-textual representations achieves better performance for speaker recognition than using each modality alone**, which is concordant with the previous experiments performed for the AER task in section 4.2.2 and section 4.3.

The results of the AER model for the W2V2-XLSR-56, RoBERTa and Whisper representations are shown in Figure 4.14. The results show that the W2V2-XLSR-56 representations fused with the speaker representations can achieve a marginal improvement (p value of .075 in a z score test) in the performance of the AER model compared to using only the W2V2-XLSR-56 representations. This improvement might demonstrate the effectiveness of fusing speaker information with acoustic representations for AER from acoustic signals. Furthermore, the training set of the speaker recognition model contained only eight speakers, which can mean that even a small number of speakers can help to generalise different speaker styles, which can be further exploited to "personalise" AER models by the proposed model. On the other hand, fusing speaker representations with the RoBERTa representations for the human transcriptions does not significantly improve the AER performance. Moreover, using ASR transcriptions in this paradigm significantly reduces the AER performance when using speaker representations that are trained on the RoBERTa representations of the ASR transcriptions. This may be because there is little correlation between different speakers and textual representations in the IEMOCAP dataset (see the paragraph above).

The results also show that joint W2V2-RoBERTa representations achieve better performance than W2V2-XLSR-56 representation alone (when using human transcriptions), with and without fusion with speaker representations. This shows that the acoustic representations can be made "verbal-aware" and "speaker-aware" by simply fusing textual and speaker representations at the model level, resulting in better performance for AER. However, as can be seen from the results of the joint W2V2-RoBERTa representations from ASR transcriptions, this method relies heavily on the performance of the ASR model. Furthermore, by comparing the results of the speaker and emotion recognition models found in Figure 4.13 and Figure

4.14, one can see a correlation between the performance of the speaker recognition and the emotion recognition models. This suggests that **the effectiveness of the speaker representations used in the AER model depends on the accuracy of the speaker recognition model that provides the speaker representations.**

Moreover, the results for AER on the IEMOCAP dataset show that **the Whisper representations are as effective as using speaker-aware joint W2V2-RoBERTa representations**. This may be because the Whisper representations are trained directly for ASR and indirectly for speaker recognition (see "Experiments" above), which means that such representations **already contain speaker and verbal information and do not need to be fused with such information** in an extra step. This hypothesis is further supported by the results of the Whisper representations fused with speaker representations, which show no improvement over using only the Whisper representations for the AER task.

The experiments and results presented in this section are briefly summarised and further discussed below.

4.4.3 Discussion

The aim of this section was to investigate the effectiveness of speaker awareness for deep pre-trained representations of acoustic signals and text. To achieve this goal, a method was proposed in which speaker representations trained from a speaker recognition model are fused at the model level with acoustic and textual representations. Several experiments were then carried out on the IEMOCAP dataset to evaluate this method. The results showed that although speaker recognition models perform better from acoustic signals than from text, the use of both modalities still outperforms the use of either modality alone for the speaker recognition task. The latent speaker representations from the pre-trained speaker recognition model can then be fused with acoustic and textual representations to provide speaker information. The results of this method for the AER task show that the fusion of speaker representations with acoustic-textual representations can perform marginally (but not significantly) better than the joint acoustic-textual representations without speaker representations, but only when the text is human transcriptions. The use of ASR transcriptions instead of human transcriptions was shown to be less effective in this method due to the lack of accuracy of the automatic transcriptions. These results suggest that for an effective speaker-aware AER model, in the absence of human transcriptions and an accurate ASR model, it is better to use only speaker representations computed from acoustic signals, rather than automatic transcriptions. Moreover, the results suggest that the use of deep acoustic representations already trained on speaker and verbal information, such as the Whisper representations, would not need to be further fused with speaker or text representations to achieve comparable results to speaker-aware W2V2-RoBERTa representations for the AER task.

It should be noted that the study presented in this section has some limitations. One such limitation is that the method used here was only evaluated on the acted IEMOCAP dataset and not on emotional expressions in the wild. This is mainly because the CMU-MOSEI dataset, which was used in previous experiments to represent data in the wild, could not be used here as this dataset does not provide speaker IDs. Furthermore, the IEMOCAP dataset uses only 10 speakers, eight of which were used to train the speaker recognition model and the other two speakers were used to test the speaker representations for the AER model. The evaluation of the "speaker awareness" method with only two speakers can also be seen as a limitation of this study.

4.5 Summary

Recent advances in DNNs have shown that deep pre-trained representations, and in particular Wav2vec2 and RoBERTa representations, can achieve state-of-the-art performance in a wide range of speech and text tasks, including emotion recognition (see Chapter 2). As such deep representations are formed incrementally on large amounts of data, the relationship between their training data and their performance for a downstream task such as emotion recognition is difficult to understand and control. To this end, in Section 4.1, several Wav2vec2 models were pre-trained on different amounts and types of speech, and their performance for dimensional emotion recognition was analysed on the RECOLA and AlloSat datasets. The results showed that the type of data plays a more important role than the amount of data. In particular, Wav2vec2 representations pre-trained on French data perform better on French emotion prediction tasks than Wav2vec2 representations pre-trained on English data. However, pre-training the deep representations on more data does not necessarily lead to better performance for AER. Moreover, compared to traditional MFB features, such deep representations rely on less complex models to achieve good AER performance.

Furthermore, as emotions are conveyed by both verbal and non-verbal communication, the joint acoustic-textual representation of acoustic and textual representations has been shown to be more effective than using each modality alone (see section 2.2.5). Therefore, in section 4.2, several methods for fusing acoustic (W2V2-XLSR-56) and textual (RoBERTa) representations were investigated on the IEMOCAP (for acted expressions) and CMU-MOSEI (for in-the-wild expressions) datasets. The results showed that simply concatenating latent representations of pre-trained acoustic and textual representations achieved better AER performance for acted and in-the-wild emotional expressions than using the representation of each modality alone. These results are consistent with those observed in similar studies such as Siriwardhana et al. (2020).

However, joint acoustic-textual representations often rely on human transcriptions of acoustic signals, which are not always available in a realistic application of AER models. Therefore, several studies have investigated the use of ASR transcriptions to provide joint acoustic-textual representations for AER on acted emotional expressions (Heusser et al., 2019; Yoon et al., 2019; Wu et al., 2021; Peng et al., 2021). These studies have shown that although the joint acoustic-textual representations based on ASR transcriptions are not as effective as human transcriptions for AER, they still perform better than relying on acoustic representations alone. To advance the aforementioned state-of-the-art studies, in Section 4.3, this method is further evaluated here for emotional expressions in the wild, by exploiting the CMU-MOSEI dataset. The results show that this method is also effective for emotional expressions in the wild, but mainly due to the use of textual representations,

as the acoustic representations, whether W2V2-XLSR-56 or traditional LLDs, were not shown to be robust enough to be used reliably for emotional expressions in the wild.

The use of verbal messages from textual representations is not the only source of information that can improve the performance of AER from acoustic representations. The use of speaker representations has also been shown to be effective for AER (Pappagari et al., 2020), as it provides information about the style of different speakers. Therefore, in section 4.4, several experiments were performed on the acted IEMOCAP dataset to investigate the effectiveness of fusing speaker representations into joint acoustic-textual representations on acted AER, by using W2V2-XLSR-56 and RoBERTa representations. The results showed that speaker-aware acoustic-textual representations can achieve marginally better AER performance than acoustic-textual representations without speaker representation. However, it was also shown that using ASR transcriptions to compute speaker-aware acoustic-textual representations does not lead to significantly better results than using speaker-aware acoustic representations, and therefore the accuracy of the ASR model is a limiting factor in this paradigm. Furthermore, it was shown that deep acoustic representations such as Whisper, which have been pre-trained directly for ASR and indirectly for speaker recognition, can contain both verbal and speaker information and do not need to be fused with textual and speaker representations to achieve comparable performance to speaker-aware joint W2V2-RoBERTa representations for acted AER.

Chapter 5

Generalisation beyond emotion schemes

From pixabay.com

AER aims to predict the numerical representation of an affective state of mind, from a numerical representation of a given acoustic or textual emotional expression (see Section 1.1). In particular, the use of deep pre-trained representations of acoustic signals and text, such as Wav2Vec2 or RoBERTa, have shown great performances in AER in Chapter 4 for acted and in-the-wild emotional expressions. The better performance of deep representations over traditional hand-crafted features is usually attributed to the fact that deep representations are trained on large amounts of data, in order to form a data-driven feature extraction that can effectively model different variations in the acoustic signals or text, rather than designing feature extraction methods based on our limited acoustic or textual knowledge (see Section 2.2.3).

Furthermore, the state-of-the-art Automatic Emotion Recognition (AER) models usually use supervised data-driven machine learning methods to map the deep representations of acoustic signals or text to specific emotion annotations defined for each specific dataset. This is because numerical representations of emotion are defined and annotated in different subjective ways from one dataset to another, and thus it is challenging to consider multiple datasets to train AER models (see Section 1.1.1). For example, the RECOLA dataset is annotated based on arousal and valence dimensions of emotion according to six specific annotators, while the EmoDB's labels are related to an actor's perception of a set of emotion labels – anger, anxiety, boredom, disgust, happiness, neutral, and sadness– (see Section 3.1). Also, the emotional expressions in each dataset represent a limited range of the vast possibilities of all the emotional expressions that can be observed in the wild. Therefore, it is important to exploit multiple datasets in order to generalise across a wide range of emotional expressions. Although self-supervised representations have been shown to generalise well across datasets, there do not appear to be sufficiently large datasets of emotional expressions to train a self-supervised representation to generalise across emotional expressions without using diverse emotion annotations (see Section 4.1). But how can a supervised machine learning model be trained to generalise across different emotion annotations from different datasets?

To address the challenge of generalisation across different emotion annotation schemes, MTL is often used in the state of the art to consider different classifiers for different annotation schemes of each dataset, while sharing a main model across all used datasets (see Section 2.2.7). For example, in Xia and Liu (2015); Kim et al. (2017), MTL paradigms have been successfully used by assigning different classifiers for different emotion dimensions and showing superior performance of AER from acoustic signals than using each emotion dimension as the target alone. Also, in Zhang et al. (2017b) different classifiers are assigned to the set of emotion labels of multiple emotion datasets in order to better represent emotional expressions from acoustic features across the used datasets. However, they considered that all the tasks are defined for arousal and valence, which requires all the used datasets to follow the same annotation scheme. Later in Zhu and Sato (2020), a MTL paradigm is used to first compute a latent emotion representation (called "emotion embedding") from acoustic signals via a common model, which can then be mapped to different sets of emotion labels by separate classifiers. However, their work is limited as they only used two small acted datasets to evaluate their method. The aforementioned studies on MTL for AER show that this area of research has not yet been explored for acted and in-the-wild AER from deep acoustic and textual representations.

Therefore, to advance the state of the art, this chapter proposes and evaluates a method based on MTL and deep representations of acoustic and textual data that can go beyond different emotion schemes, by relating similar emotions across different corpora. In Section 5.1, the proposed method, which uses MTL for AER

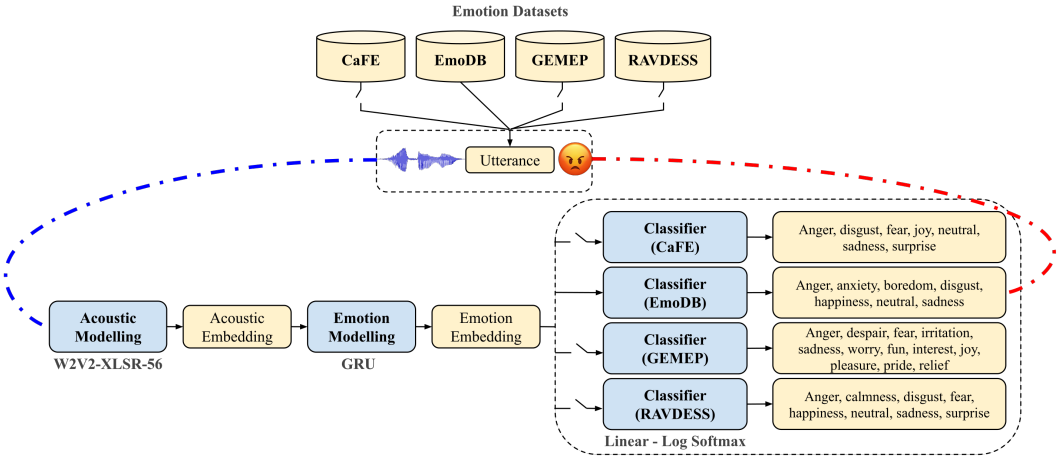


Figure 5.1: The proposed stepwise multi-corpus emotion recognition model for CaFE, EmoDB, GEMEP, and RAVDESS datasets. The acoustic and the emotion modelling steps are shared for all the utterances of all datasets. However, each set of emotion categories of each corpus, use a specific classifier to predict the probabilities of the emotion classes, based on the emotion embedding.

from deep acoustic representations, is further explained, and several experiments are performed to evaluate the proposed method for multiple acted datasets. Then, in Section 5.2, this method is further evaluated for AER in the wild, and how acted emotional expressions can further improve the AER performance from both acoustic and textual representations. Finally, the chapter is summarised in Section 5.3.

5.1 Multi-corpus acted emotion recognition from acoustic signals

In this section, a method integrating MTL with state-of-the-art deep acoustic representations (Wav2vec2 and Whisper) is proposed and evaluated[1]. This method is described in Section 5.1.1. The datasets used to evaluate this method are presented in Section 5.1.2, followed by the experimental setup explained in Section 5.1.3. The method is then evaluated for both within-corpus and cross-corpus settings in Section 5.1.4 and Section 5.1.5, respectively.

5.1.1 Method

Figure 5.1 shows the proposed method. The following describes the different parts of this model and how they work together:

[1]The work presented in this section is published in (Alisamir et al., 2022c)

1. **Emotion datasets**: The proposed method is based on the use of multiple emotion datasets with different sets of emotion labels, which can be predicted by classifiers dedicated to the labels of each dataset. For example, the experiments in this section use four acted datasets from CaFE, EmoDB, GEMEP and RAVDESS, and as can be seen from the Figure 5.1, each dataset uses a different set of emotion labels (see Section 5.1.2).

2. **Acoustic modelling**: For an utterance from a given dataset, the model first computes the *acoustic embedding* by using different acoustic representations such as MFB, Wav2vec2, or Whisper (see Section 3.2). For the Wav2vec2 representation, the W2V2-XLSR-56 is used here as the acoustic modelling step, because it can achieve state-of-the-art results (see the results of different experiments in Chapter 4), and it is trained on 56 different languages, which is important since the datasets mentioned above also use various languages of English, French, and German.

3. **Emotion modelling**: Given an acoustic embedding, the emotion modelling step calculates the emotion embedding. Here GRU is used for emotion modelling, because it has been shown to be able to effectively model acoustic data for AER (see Section 2.2.8). Since GRUs learn to model sequential data in a recurrent manner, each item in a sequence produces a corresponding output, which is also influenced by the outputs of the previous items in the sequence. Therefore, the last output of the GRU is used here to represent the *emotion embedding*, because it contains information about the entire input utterance. It should be noted that the GRU used for emotion modelling is shared for all the datasets, regardless of their emotion annotation schemes. The hypothesis is that, by sharing the emotion model during the training, it can learn to represent an "understanding" of the underlying perceivable emotion across the different corpora.

4. **Classifiers**: Each corpus uses a specific classifier to map the emotion embedding to the specific way it defines emotion categories. Here, the classifiers consist of a linear (fully connected) layer followed by a log-softmax function to estimate the class probabilities associated with each dataset's set of emotions.

The training strategy used to train the different models involved in the proposed method is described below.

Training strategy

The GRU and linear models mentioned above are trained using the Adam optimiser (see Section 3.3), where the loss function is cross-entropy (see Section 3.4.2). The

training is performed for each utterance of each dataset independently, in a stepwise manner. This means that during training, for each utterance, the acoustic and then the emotion embedding is computed, then only the appropriate classifier assigned to the dataset of the input utterance is used to compute the set of emotion probabilities. Then the cross-entropy loss, which is commonly used for classification tasks, is calculated and its resulting gradients are sent backwards through the associated classifier as well as the shared emotion model (here the GRU). In this way, the emotion model would continue to be trained regardless of which corpus the input utterance belongs to, but the classifiers would only be optimised according to which corpus the input utterance belongs to. It should be noted that the stepwise training strategy used here is slightly different from common MTL paradigms, where the loss function consists of multiple terms, where each term refers to the loss function computed for the sample of each task. Because this common MTL training strategy uses multiple samples for each training iteration, it consumes more memory than the stepwise training strategy, which is why the stepwise training strategy is used here.

Unless otherwise stated, the Wav2Vec2 representations used here are considered "frozen" during the training of the AER models, which means that their weights do not change and are used only as pre-trained models. However, in some fine tuning experiments with the Wav2Vec2 representations, the loss is also sent back through the Wav2Vec2 model. Thus, based on the gradients calculated for each training iteration, the weights of the Wav2Vec2 model can be updated alongside the weights of the emotion model and the classifiers. In this way, the Wav2Vec2 representations are influenced by the utterances used during training. The process of fine-tuning has recently become popular as it allows the pre-trained representations to be adapted to a specific task with a specific data distribution. In what follows, the datasets used to evaluate this method are presented.

5.1.2 Datasets

The experiments in this section use four acted datasets from CaFE, EmoDB, GEMEP and RAVDESS, which are described in detail in Section 3.1. In this section, we focus only on acoustic expressions of emotion in order to study the interplay of the proposed method more clearly, before moving on to use data recorded in the wild and using textual transcriptions in Section 5.2, which would make the method more difficult to analyse in detail, as expressions in the wild contain more subtle emotions as well as environmental noise.

Table 5.1 presents the generic statistics related to the training, development and test partitions of the datasets used here. As can be seen from this table, different datasets contain different amounts of data. Preliminary experiments have shown that the amount of data for each corpus has a direct correlation with the influence that

Table 5.1: Statistics related to the training, development and test partitions of the acted datasets of CaFE, EmoDB, GEMEP, and RAVDESS, which are used for multi-task learning experiments from deep acoustic representations.

Dataset	Number of utterances			Duration in minutes		
	Train	Dev	Test	Train	Dev	Test
CaFE	624	156	156	44	12	13
EmoDB	292	116	127	13	5	6
GEMEP	648	216	216	26	9	8
RAVDESS	1080	180	180	66	12	11

Table 5.2: Details of the mappings of the original emotion targets of each corpus to negative, neutral, and positive classes, as were used here in the cross-corpus evaluations.

Corpus	Negative emotions	Neutral emotions	Positive emotions
CaFE	anger, disgust, fear, sadness	neutral, surprise	joy
EmoDB	anger, anxiety, boredom, disgust, sadness	neutral	happiness
GEMEP	anger, despair, fear, irritation, sadness, worry	-	fun, interest, joy, pleasure, pride, relief
RAVDESS	anger, disgust, fear, sadness	neutral, surprise	calmness, happiness

corpus has on the weights of the shared GRU model after training. For instance, it was found that the trained model performed better on the RAVDESS dataset, which contains more data than other datasets used during training. To solve this problem, the utterances of the underrepresented corpora are randomly selected and duplicated for the training and development partitions in order to equalise the number of training utterances for all datasets.

Furthermore, to evaluate the proposed method in cross-corpus settings (see Section 5.1.5) in a coherent way, the output predictions of each utterance is mapped to either a negative, neutral or positive class, instead of using the original predictions of each dataset, which depend on the specific set of emotion labels for each dataset. The mapping between the original emotion targets of each dataset, and the negative, neutral or positive classes are presented in Table 5.2. Further details of the training and experimental setup is presented below.

5.1.3 Experimental setup

In order to define the setup of the proposed method (see Figure 5.1), different possible models and hyper-parameters were evaluated. Specifically, a grid search was performed over the different possible setups with different ranges of complexity, as follows:

- Emotion model: GRU, Transformer (8 heads)

- Classifiers: GRU, Linear
- Hyper-parameter of the GRU and Transformer models: 1 layer with 64 nodes, 2 layers with 128 nodes, 4 layers with 256 nodes
- Learning rate: 0.01, 0.001, 0.0001

After training each possible setup for 50 epochs on two datasets of GEMEP and RAVDESS independently (without the MTL), and by using the W2V2-XLSR-56 acoustic representations, the best performing setup was found to be the GRU model using 1 layer with 64 nodes, and learning rate of 0.0001. The linear layers for the emotion model were not investigated as previous experiments in Section 4.1 clearly show superior performance of GRUs compared with linear layers. On the contrary, for the classifiers, no improvements were observed when using GRUs instead of linear layers, which may be because a simple GRU with 1 layer and 64 nodes is sufficient for modelling emotion for the GEMEP and RAVDESS datasets. Therefore, **the proposed method described in Section 5.1.1 uses the GRU model using 1 layer with 64 nodes, and linear (fully-connected) classifiers** (see Figure 5.1).

The experiments defined in this section will then train the proposed method both separately for each dataset (single-corpus) and by using all four datasets – CaFE, EmoDB, GEMEP, and RADVESS– presented above (multi-corpus). In addition, the fine-tuning of the W2V2-XLSR-56 is also investigated here, in order to further investigate the effect of fine-tuning the acoustic representations within the MTL context. These experiments are designed to show whether the use of multiple datasets, even small ones, can help to achieve a more generalised emotion embedding than training a model for each dataset alone. To evaluate this hypothesis, the trained models for single-corpus and multi-corpus (using MTL) methods are evaluated for within-corpus and cross-corpus settings. The within-corpus results involve training and testing each dataset with its own classifier, while the cross-corpus results involve training the classifier of each dataset for the same dataset, but testing it with the classifier of another dataset. The within-corpus and cross-corpus results are presented and analysed in Section 5.1.4 and Section 5.1.5 respectively.

5.1.4 Within-corpus results

Figure 5.2 presents the results of the within-corpus evaluation of the proposed multi-corpus (using MTL) versus single-corpus training strategies.

Regarding the results of W2V2-XLSR-56 with frozen weights (no fine-tuning), we can see that a weakly significant[1] UAR is observed for the EmoDB dataset, but for the other datasets the results are either the same or comparable. As the EmoDB

[1]P-value of .052 using the Z-score statistical measure

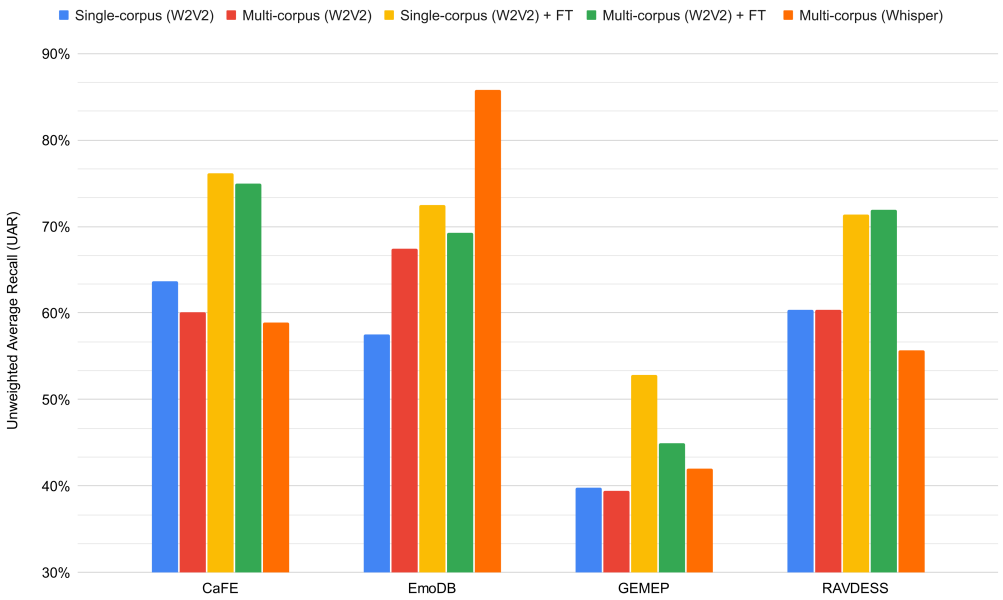


Figure 5.2: The results of the within-corpus evaluation for multi-corpus versus single-corpus training methods, by using Whisper, and W2V2-XLSR-56 with and without Fine-Tuning (FT).

dataset contains the least number of training examples, this result suggests that the proposed MTL method with multi-corpus training is especially helpful for very small datasets, even if the additional datasets used do not share the same emotion annotation scheme. Also, this finding is in line with previous studies such as Zhang et al. (2017b, 2022); Zhu and Sato (2020), which have shown that multi-corpus training can improve performance on a specific corpus, but not necessarily for all datasets involved.

Contrary to the results of the experiments for W2V2-XLSR-56 without fine-tuning, the results of the experiments with fine-tuning show that training and testing on a single corpus produces better or comparable results to the multi-corpus training strategy. This could be because with fine-tuning, the W2V2-XLSR-56 learn to predict more corpus-specific acoustic representations, rather than learning more generic representations that could be beneficial in cross-corpus settings. This is further explored in Section 5.1.5, where the cross-corpus results are analysed.

Moreover, the results by using the Whisper representations only show a big improvement for the EmoDB dataset, which is the smallest dataset used here, compared to the W2V2-XLSR-56 representations, but this improvement does not extend to other datasets. This may be due to the fact that the Whisper model was trained for more than ten times data compared to the W2V2-XLSR-56 model (see Table

Table 5.3: Results of the within-corpus experiments for CaFE, EmoDB, GEMEP and RAVDESS datasets. In order to fairly compare the results of this study with the state of the art, the reported result of this work (last row) is based on the Leave One Speaker Out (LOSO) cross validation.

Method	Evaluation Metric	CaFE	EmoDB	GEMEP	RAVDESS
Subspace learning and extreme learning Xu et al. (2018)	UAR (Random Partitioning)	-	-	43.3 %	
Prosodic and spectral features + SVM El Seknedy and Fawzi (2021)	Accuracy (10 fold Cross-Validation)	70.6 %	86.0 %	-	70.6 %
MFCC/GTCC features with echo state network Ibrahim et al. (2021)	UAR (LOSO)	-	86.8 %	-	73.1 %
W2V2-XLSR-56+FT (single corpus training)	UAR (LOSO)	**77.2 %**	**90.5 %**	**55.8 %**	**82.2 %**

3.2), and thus it might be able to generalise better, when exploited in a supervised model that is trained on smaller amounts of data. Also, the Whisper model is mainly trained for ASR-related tasks, such as speech transcription and translation, whereas the datasets used here have no correlation between the uttered phrases and the emotion, because in the CaFE, EmoDB, GEMEP and RAVDESS datasets different actors utter a fixed number of phrases in different ways to express different emotions. Furthermore, because the best results here, on average, were achieved with fine-tuned W2V2-XLSR-56 representations with single-corpus training, in what follows, this paradigm is compared to the state of the art.

In the experiments performed here, the CaFE, EmoDB, GEMEP and RADVESS datasets were divided into training, development and test sets. However, as these datasets are rather small, state-of-the-art works usually use Leave One Speaker Out (LOSO) cross-validation to compare different methods. In LOSO cross-validation, the model is trained on all but one of the speakers involved, on which it is tested. This process is repeated until all the speakers have been tested, and then the average results for all the speakers is reported. As the state of the art uses LOSO cross-validation, the best within-corpus method here, which is the W2V2-XLSR-56 representation with single-corpus training, is also evaluated by using LOSO cross-validation. The results are presented in Table 5.3. The results show that the proposed method can achieve better performance than the state of the art for all the datasets, which seems to be mainly due to the use of highly contextualised W2V2-XLSR-56 representations compared to other techniques. It should be noted that the

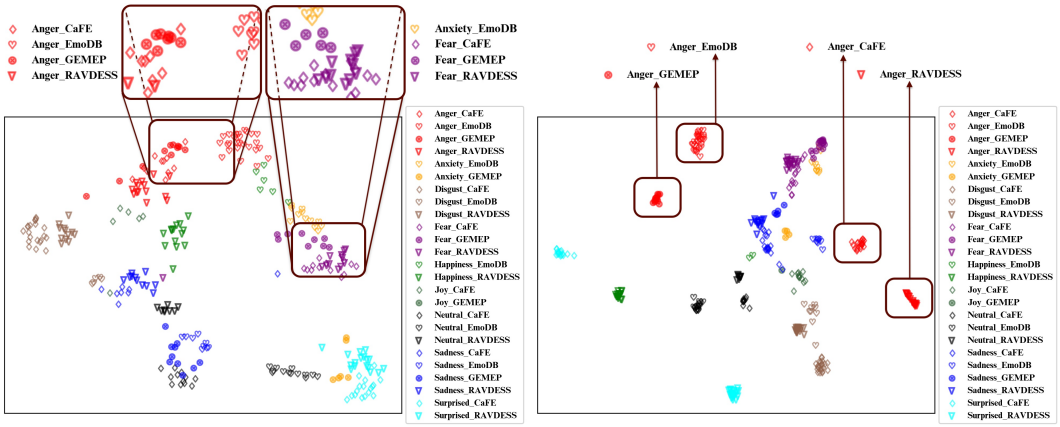


Figure 5.3: Emotion embeddings of different correctly classified utterances of the test partitions of the studied corpora. Only emotions that were used at least in two different corpora are shown. On the left: emotion embeddings of the baseline model with W2V2-XLSR-56 weights frozen during the training; clusters of similar emotions across different datasets can be identified in this space. On the right: emotion embeddings of the baseline model with fine-tuning of the W2V2-XLSR-56 weights during the training; clusters of emotion are now specific to each dataset and similar expressions are located in different parts of the emotion space.

results in the rest of this section are based on the partitioning of the Table 5.1 in order to keep the results of the different experiments in this section comparable.

To further investigate the generalisability of the emotion embeddings and the effect of fine-tuning the deep acoustic representations on them, the emotion embeddings are visualised in Figure 5.3. Here, UMAP (McInnes et al., 2018) is used to reduce the dimensions of the emotion embeddings and display them in a two-dimensional space. From the figure, we can see that for the multi-corpus training paradigm (without fine-tuning), the emotion embeddings of utterances from different corpora are mostly placed closer to each other in the embedding space when they represent the same or similar emotions. For example, different utterances representing anger are placed closer together in the upper part of the figure. We can also observe that utterances labelled as Happiness and Joy are close to utterances labelled as Anger. This may be because both happiness and anger are associated with high arousal according to Russell's theory of core affect (see Figure 2.1). Furthermore, we can observe that utterances labelled as anxiety in EmoDB are placed close to utterances labelled as fear from the other corpora in the emotion embedding space. The relationship between anxiety and fear has been studied in psychology, where anxiety is seen as a more general case of cue-specific fears (Lang et al., 2000). **These observations suggest that the proposed multi-corpus method, without fine-tuning acoustic representations, can in most cases provide a sense of the**

underlying emotion across different corpora.

However, as the acoustic representations are fine-tuned, this generalisation in the emotion embedding space seems to disappear, and instead different utterances from the same corpus tend to be closer together in this space. This may suggest that as acoustic representations are fine-tuned, acoustic embeddings become more corpus-specific, which may lead to emotion embeddings becoming more corpus-dependent. **This may mean that fine-tuning the W2V2-XLSR-56 representations makes the multi-corpus training method less focused on capturing the underlying emotion across corpora**. To investigate this further, the models trained with the single-corpus and multi-corpus strategies are also evaluated in cross-corpus settings below.

5.1.5 Cross-corpus results

One way to study the generalisation of emotion embeddings across corpora is to take an utterance from one corpus as input to the proposed MTL-based method, and observe the output of the classifier dedicated to another corpus. However, since different corpora have different sets of emotion labels, the cross-corpus results cannot be quantified in the same way as in the within-corpus evaluation above. Therefore, here the output predictions of each utterance are mapped to either a negative, neutral or positive class (see Table 5.2), after the training is complete by using the original labels for each dataset. In this way, **all the outputs of all the classifiers are mapped to the three classes –negative, neutral and positive– in order to quantify the cross-corpus results** and make them useful for different comparisons. For example, if an utterance labelled irritation in the GEMEP dataset is predicted as disgust by the CaFE classifier, then this prediction is counted as correct, as both are considered negative emotions. The results of this evaluation are shown in Figure 5.4.

The results in Figure 5.4 for multi-corpus training, show that although the W2V2-XLSR-56 representation with fine-tuning achieves the best performance for within-corpus evaluation over three negative, neutral and positive classes in Figure 5.4, it has one of the worst performances for cross-corpus evaluation. On the other hand, the W2V2-XLSR-56 representation without fine-tuning achieves the best results for cross-corpus evaluation. **This result suggests that the multi-corpus training with the W2V2-XLSR-56 representation without fine-tuning can be effectively used to compute generalisable emotion embeddings**, which is consistent with the observation of emotion embeddings for the W2V2-XLSR-56 representation without fine-tuning in Figure 5.3.

From Figure 5.4, it can also be seen that in the case of single-corpus training, fine-tuning the Wav2vec2 representations can slightly improve the performance of the model in the cross-corpus evaluation, while in the case of multi-corpus train-

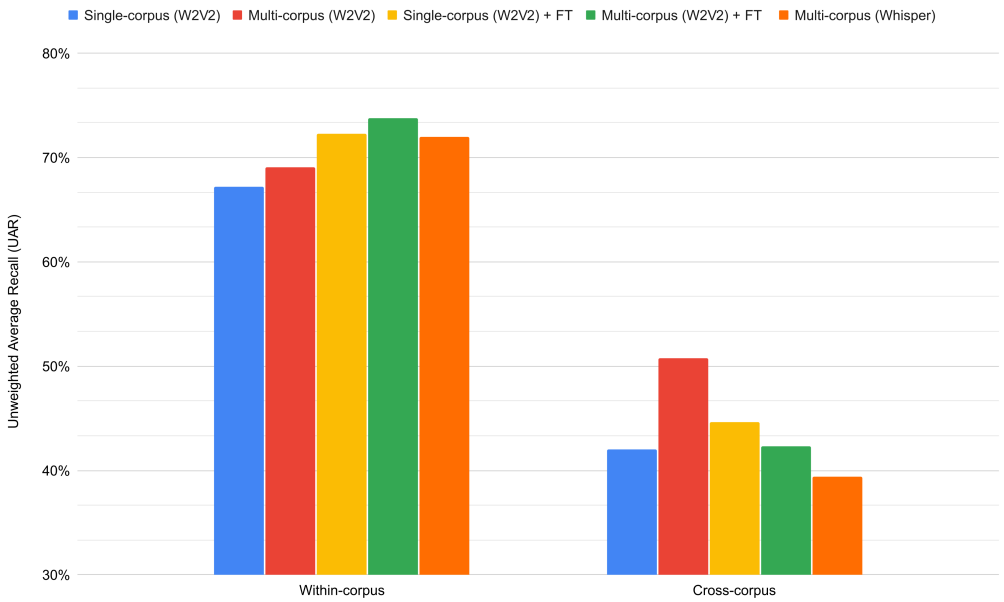


Figure 5.4: The results of the multi-corpus training method for cross-corpus evaluations, where emotion label predictions (after the training is complete) are mapped to three classes of negative, neutral and positive. The results are presented for W2V2-XLSR-56 representations with and without Fine-Tuning (FT), plus the Whisper representations.

ing, fine-tuning seems to worsen the results. This may be due to the fact that the datasets used here are similar in terms of being acted clean data. Thus, fine-tuning the W2V2-XLSR-56 with single-corpus training may help to generalise to other similar corpora. However, in multi-corpus training, since all the datasets are trained together, the model seems to learn to distinguish the data from different datasets in order to improve performance on a given utterance, which may only come from one of the four datasets used in this section.

To further investigate why fine-tuning deep acoustic representations for multi-corpus training does not generalise emotion embedding across different corpora, we can also analyse a specific confusion matrix where the input utterance and its classifier differ. For example, Figure 5.5 shows confusion matrices where the utterances of the CaFE dataset are predicted by the GEMEP classifier, for frozen, and fine-tuned W2V2-XLSR-56 representations (See all the confusion matrices in Appendix A). For the fine-tuned representations, the predictions do not look accurate, as "sadness" utterances in CaFE are classified as "joy" in GEMEP. On the other hand, the cross-corpus predictions for W2V2-XLSR-56 representations without fine-tuning seem to be mostly correct. For example, the four common emotions between the

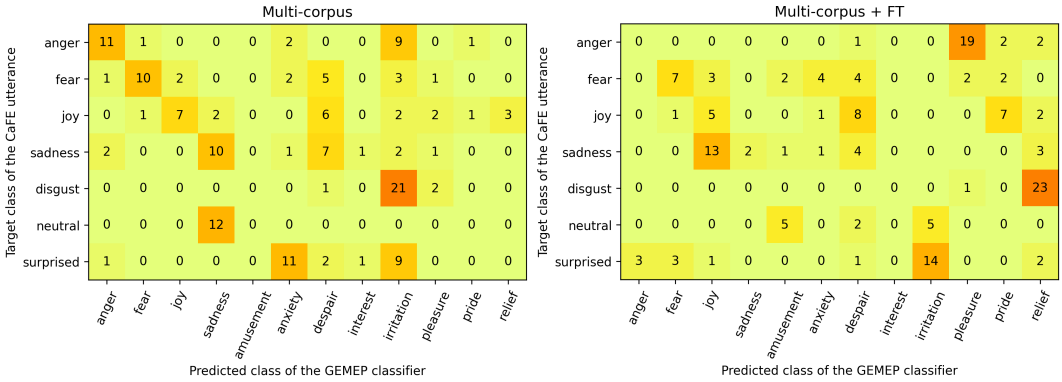


Figure 5.5: Confusion matrices of the predictions of the GEMEP classifier from the CaFE utterances. Here, the original target label for each utterance is based on the set of emotion labels of the CaFE dataset, while the prediction of the model is based on the set of emotion labels of the GEMEP dataset. The confusion matrix of the proposed multi-corpus method without fine-tuning of the W2V2-XLSR-56 representations is shown on the left, while the confusion matrix of the same method with fine-tuning of the W2V2-XLSR-56 representations is shown on the right.

two datasets –anger, fear, happiness, and sadness– are mostly predicted correctly. Interesting results are also observed when analysing emotion labels that differ between the two datasets. For example, disgust utterances from the CaFE dataset are mainly labelled as irritation, suggesting that **the proposed method is able to generalise to some extent across different emotion labels in different corpora, even if they use different labelling schemes, and without being explicitly trained to do so**. These results are also consistent with what we observed by visualising the emotion embeddings in Figure 5.3.

The proposed methodology and related findings of this section are discussed further in the discussion section below.

5.1.6 Discussion

The aim of this section was to compute a generic emotion embedding that could demonstrate the generalisation of emotion concepts across different corpora with different sets of emotion labels. To achieve this goal, a method has been proposed that uses deep acoustic representations, GRUs, and linear classifiers dedicated to each corpus. The proposed method was trained following a MTL paradigm to learn a generalised emotion embedding across four small acted datasets –CaFE, EmoDB, GEMEP, and RAVDESS–, where each dataset uses different sets of emotion labels. The results were then evaluated quantitatively in both within-corpus and cross-corpus settings, as well as by visualising the emotion embeddings and examining cross-corpus emotion recognition confusion matrices. The results suggest that this

method, by using W2V2-XLSR-56 as acoustic representations (without any fine-tuning), is able to effectively produce an emotion embedding that is generalised across different datasets, and thus can improve cross-corpus emotion recognition where the emotion labels may be different.

However, the study in this section has certain limitations. All the datasets used to evaluate the method were rather small acted datasets, where actors were hired to convey different emotions by saying the same sentences in different ways. This is not realistic, as the words chosen in a natural emotional expression are often correlated with specific emotions (Lindquist et al., 2015). Therefore, in the next section, the effect of both acoustic changes in the speech signal and the verbal message on emotional expressions in the wild is investigated.

5.2 Exploiting acted data for emotion recognition in the wild

In the previous section, it was shown that the proposed method, by using MTL, can achieve emotion embeddings that can generalise the representation of the same or similar emotional expressions across different acted datasets, which may use different sets of emotion labels. However, the experiments in the previous section only focused on analysing MTL in order to generalise the prediction of acted emotional expressions from acoustic representations. In this section, the proposed MTL-based method is further analysed to see if the use of both acted and in-the-wild emotional expressions, can help the generalisation ability of the trained model for AER in the wild.

In addition, the experiments in Section 4.2 have shown the benefits of using transcriptions of a speech signal to improve the performance of AER for both acted and in-the-wild emotional expressions. Furthermore, it has been shown that this improvement is also effective when the transcriptions are automatically extracted using an effective ASR system (see Section 4.3). Therefore, the experiments in this section also focus on the use of deep acoustic and textual representations, where the text is either human or ASR transcriptions, for both acted and in-the-wild AER. These experiments are described in more detail below.

5.2.1 Experiments

The experiments in this section, combine the proposed MTL method of Section 5.1 with the joint acoustic-textual representations explored in Section 4.2 and Section 4.3, in order to focus on using MTL to predict emotional labels across acted and in-the-wild emotion datasets, using deep acoustic and textual representations. Figure 5.6 depicts the method used for experimentation in this section. To stay consistent

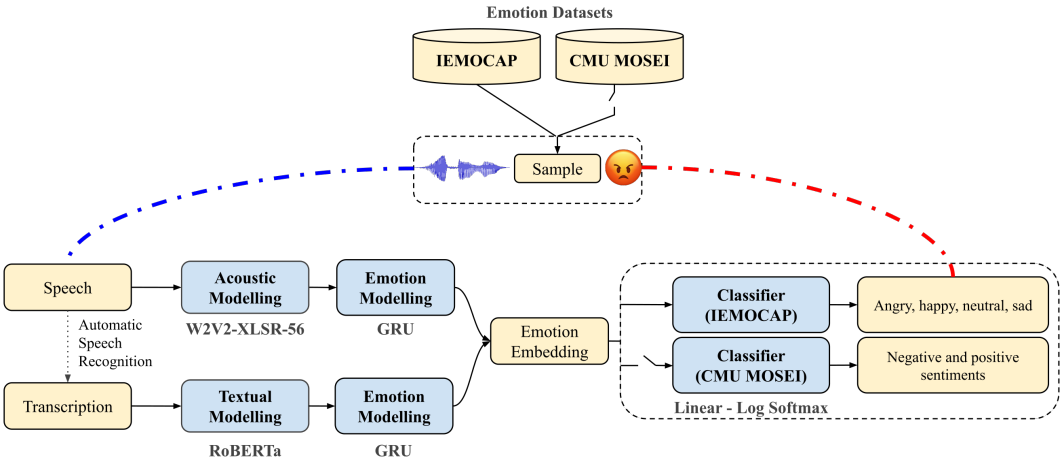


Figure 5.6: The proposed multi-corpus model used to recognise different emotion categories for IEMOCAP and CMU-MOSEI datasets from acoustic and textual representations. The acoustic, textual, and the emotion modelling steps are shared for all the utterances of all datasets. However, each set of emotion categories of each corpus, use a specific classifier to predict the probabilities of the emotion classes, based on the emotion embedding.

with the previous experiments in Section 4.3, the same datasets of CMU-MOSEI (in-the-wild) and IEMOCAP (acted, controlled environment) are investigated here (see Table 5.4 for a statistical summary of the datasets). The four acted datasets –CaFE, EmoDB, GEMEP, and RADVESS– used for the experiments in Section 5.1 are not used here because the verbal message for these datasets is not correlated with emotion by design, whereas the study in this Section focuses partly on the use of textual representations for AER. Nonetheless, to define the setup of the MTL-based method used here, this section follows the setup and training in Section 5.1. This means that a GRU-1x64 model is used for emotion modelling and linear classifiers are assigned to each dataset and trained following the proposed stepwise MTL strategy with LR=0.0001 and GA=100. In addition, the deep acoustic and textual representations used in this section are W2V2-XLSR-56 and RoBERTa (see Section 3.1), so the results of this section are comparable to the results of Section 4.3, where the GRU-1x64 model is used for the same acoustic and textual representations in a single-corpus setting. The results of this section are presented below.

5.2.2 Results

The results of the experiments in this section are shown in Figure 5.7. It should be noted that for single-corpus training, the IEMOCAP and CMU-MOSEI results in the figure are from Section 4.3, where it was shown that the ASR transcriptions

Table 5.4: Statistics related to the training, development and test partitions of the acted datasets of IEMOCAP, and CMU-MOSEI, which are used for multi-corpus experiments from deep acoustic and textual representations.

Dataset	Number of utterances			Duration in hours		
	Train	Dev	Test	Train	Dev	Test
CMU-MOSEI	18,542	1377	3340	38	3	8
IEMOCAP	3259	1031	1241	4	1	2

can be used to significantly improve the performance of AER from acoustic signals. The multi-corpus training experiments in this section also show that **existing ASR methods can be used to improve current AER from speech waves, even if the ASR transcriptions are not accurate** (see Figure 4.9 for WERs related to the ASR).

The results further show that the transcriptions of the acted IEMOCAP dataset appear to be particularly helpful in improving the AER performance for the human or ASR transcriptions of in-the-wild spoken emotional expressions of the CMU-MOSEI dataset. The improvement in AER results for the CMU-MOSEI dataset, when trained with IEMOCAP data using the MTL paradigm, seems to be mainly due to the use of transcriptions. **This finding suggests that transcriptions of acted emotional expressions can be helpful in improving AER of in-the-wild emotional expressions**.

On the other hand, considering only the results for the acoustic representations –W2V2-XLSR-56–, the IEMOCAP results show a small improvement from single-corpus to multi-corpus training, while this is not the case for the CMU-MOSEI dataset. This finding is consistent with the results of Section 5.1.4, where the proposed MTL paradigm improved the performance of the smallest dataset –EmoDB– in multi-corpus training. However, as discussed in the paragraph above, multi-corpus training for transcriptions can improve the results for the CMU-MOSEI dataset compared to single-corpus training, but not for the IEMOCAP dataset. This result means that while multi-corpus training helps the smallest dataset for acoustic representations, this trend does not extend to textual representations. This can be explained by the fact that text only contains the verbal message, which is the same in all datasets, and thus by using a smaller (and mostly scripted) dataset, one may be able to improve the recognition of different emotions from textual representations, even when extracted with an ASR system applied to acoustic signals captured in the wild. The same argument cannot be made for acoustic representations, as the acoustic signals of each word vary depending on the speaker, the ambient noise, and even the microphone used. Moreover, the deep acoustic representations used here –W2V2-XLSR-56– do not prove to be robust enough for in-the-wild AER, either within single-corpus or multi-corpus training paradigms. The experiments and results of this section are discussed further below.

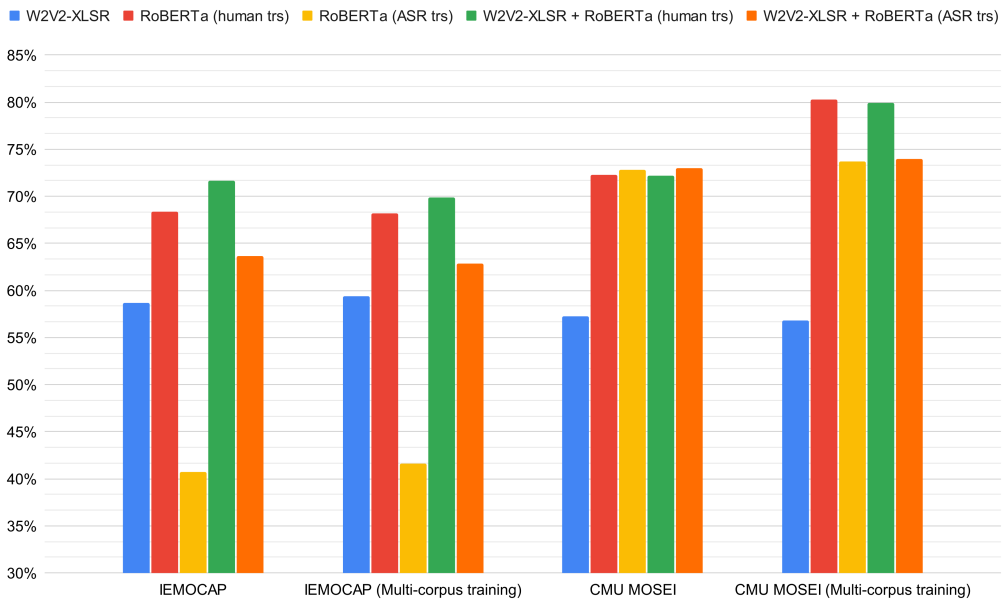


Figure 5.7: The results for using joint deep representations of acoustics and text, in a multi-corpus training paradigm, where both IEMOCAP and CMU-MOSEI corpora were used to train a shared emotion recognition model, with different classifiers exclusive to each corpus.

5.2.3 Discussion

This section has investigated the use of both acoustic and textual representations for multi-corpus training on acted and in-the-wild datasets with different sets of emotion labels. The MTL method introduced in Section 5.1.1 was then used to experiment with CMU-MOSEI (in the wild) and IEMOCAP (acted) datasets for AER from W2V2-XLSR-56 and RoBERTa representations. The results showed that by using the MTL method, the transcriptions of the IEMOCAP dataset can help to improve the AER performance of the CMU-MOSEI dataset from textual representations.

However, the study of in-the-wild emotional expressions in this section is limited to the prediction of positive and negative sentiments of the CMU-MOSEI dataset. Although positive and negative sentiments correspond somehow to the valence dimension of emotion, they do not adequately represent the full range of emotional expression in the wild. For example, the arousal dimension can further be studied for in-the-wild emotional expressions. However, annotations for the arousal dimension did not exist for the CMU-MOSEI dataset, which brings us to another limitation of this work, which is using only one dataset with emotional expressions in the wild. Other in-the-wild datasets such as Aff-wild (Kollias and Zafeiriou,

2018) or SEWA (Kossaifi et al., 2021), which contain both arousal and valence annotations, can be further added in the MTL paradigm to obtain more conclusive results. Moreover, the study here only focused on within-corpus experiments; to evaluate the effectiveness of the proposed method in a more realistic scenario, it is important to consider cross-corpus settings, where unseen datasets are used to evaluate a trained model.

5.3 Summary

Current state-of-the-art AER methods use data-driven machine learning models to map the acoustic or textual representations of data to numerical representations of emotion. However, because emotion is a subjective concept for which there is no agreed definition, the representation of emotion varies between different datasets. On the other hand, it is important to train AER models on multiple datasets, since each dataset represents a limited range of emotional expressions that can be observed in the wild. In order to use multiple datasets with different emotion annotation schemes, this section proposes an MTL-based method that predicts a generalised emotion embedding, which can then be mapped to different sets of emotions based on each dataset (see Section 5.1.1). To compute the emotion embeddings, the proposed method uses a shared GRU model, which is trained alongside linear (fully connected) classifiers that are each dedicated to the utterances of each dataset involved.

This method was first evaluated on four small acted datasets –CaFE, EmoDB, GEMEP, and RAVDESS– in within-corpus, and cross-corpus settings for AER from deep acoustic representations –W2V2-XLSR-56– (see Section 5.1). The within-corpus results showed that this method is especially effective for improving the AER performance of small datasets. On the other hand, the cross-corpus results suggest that this method can effectively compute emotion embeddings that can generalise beyond different emotion labels in different corpora.

The proposed MTL-based method was then further evaluated for in-the-wild emotional expressions and also with deep textual representations –RoBERTa–, in addition to deep acoustic representations –W2V2-XLSR-56– (see Section 5.2). Instead of the four acted datasets, the CMU-MOSEI (in the wild) and IEMOCAP (acted) datasets were then used to investigate whether acted emotional expressions, when used in a MTL paradigm, were useful for AER in the wild. The results showed that by using deep acoustic representations, no improvement is observed for the CMU-MOSEI dataset when trained in a multi-corpus setting with the IEMOCAP data compared to when trained alone (single-corpus training). This was attributed to the fact that W2V2-XLSR-56 representations were not robust enough to generalise across different speakers, environments or microphones. On the other hand, the use of deep textual representations of the acted IEMOCAP dataset helped to significantly improve the performance of AER for the in-the-wild CMU-MOSEI dataset. This finding suggests that although deep acoustic representations may not yet be robust enough to be used for in-the-wild emotional expressions, whether trained with single-corpus or multi-corpus training strategies, the transcriptions of speech signals may significantly benefit from the proposed multi-corpus training method to provide a better in-the-wild AER than using the speech signals alone.

Chapter 6

Conclusion

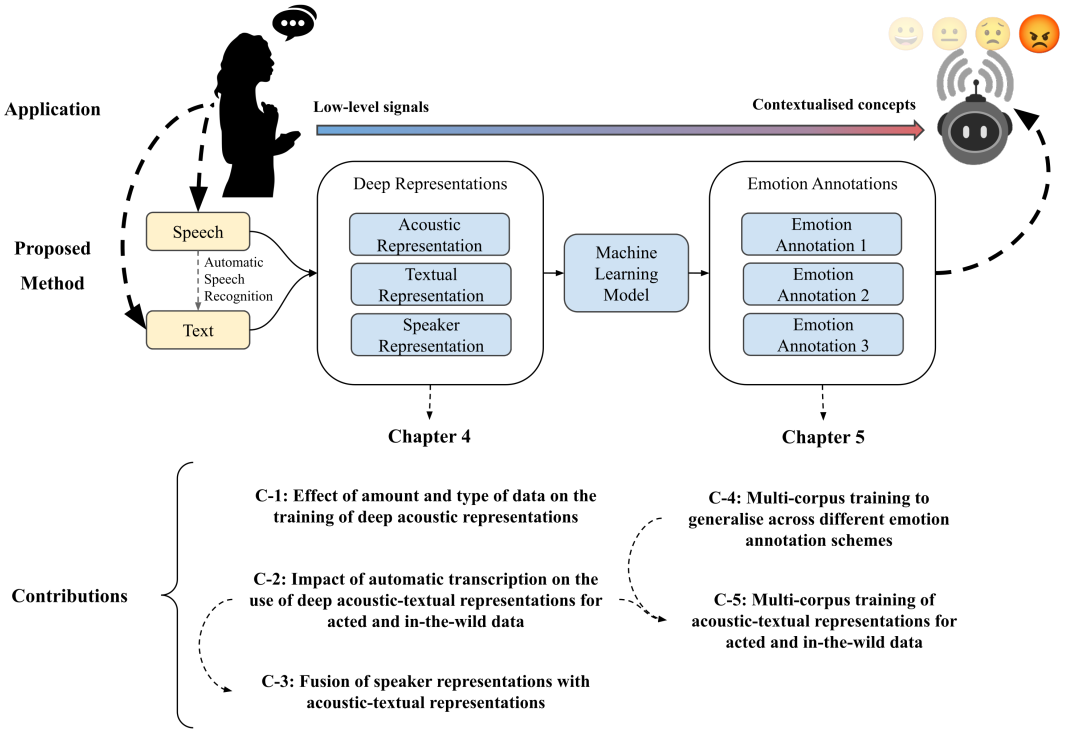


Figure 6.1: Overview of the contributions of this thesis. The intended application of the thesis is shown at the top, and the used method is shown in the middle. At the bottom, the contributions (C-1 to C-5) of the experimental chapters (4 and 5) to the proposed method are shown, with dashed lines indicating the relationship between different contributions.

In this thesis, several studies have been conducted to advance current research on machine learning of emotional expressions in the wild from acoustic signals

and text. Since AER aims to predict emotion annotations from acoustic or textual representations, this thesis has aimed to advance the state of the art in two areas:

1. The use of deep acoustic and textual representations (see Chapter 4)
2. Generalisation beyond emotion schemes (see Chapter 5)

An overview of the main contributions of this thesis to these two areas is presented in Figure 6.1. The remainder of this chapter summarises the research questions (denoted as Q-1 to Q-5 in the left margin), contributions (C-1 to C-5) and limitations (L-1 to L-5) related to the aims of this thesis in more details (see Section 1.2). Furthermore, some future research directions that can be pursued from some of the limitations of this study are also presented (denoted as F-1 to F-3 in the left margin). At the end, final concluding remarks, and envisioned applications and studies beyond this thesis are discussed in Section 6.2

6.1 Contributions, limitations, and future studies

The aim of AER can be defined as the automatic recognition of human emotions from different modalities, such as an acoustic signal or text (see Section 1.1). Current state-of-the-art AER methods use data-driven machine learning techniques to map the numerical acoustic or textual representations to numerical emotion representations. Recent advances in DNNs have put the spotlight on the deep pre-trained acoustic and textual representations, such as Wav2vec2 and RoBERTa, which have dominated state-of-the-art AER performances in recent years (see Section 2.2). This is because such representations can be trained on large amounts of data to brute-force an approximation of an effective feature extraction method, which in most cases can work better than traditional "hand-crafted" features. However, because deep representations are data-driven, rather than engineered from human knowledge of acoustic signals or text, they are difficult to interpret, study and control. This has raised our first research question, which was as follows (see Section 1.2.1):

Q-1 What is the effect of different amounts and types of acoustic signals used to train deep representations, on the performance of such representations for the AER task?

C-1 To answer the question above, in Section 4.1, several Wav2vec2 models were first trained using different types –professional, read, spontaneous, telephonic, and acted– and amounts –1k, 3k, and 7k hours– of French speech. The performance of the pre-trained French Wav2vec2 models, along with two other Wav2vec2 models trained for English and multilingual speech, was then evaluated on two French datasets, AlloSat (telephonic, in the wild) and RECOLA (spontaneous, in-the-lab), to continuously predict the frustration-satisfaction and arousal/valence dimensions of emotion, respectively. The findings showed that using more amounts of data (after 3k hours) to train deep representations does not necessarily lead to improved

performance of such representations on the downstream task of AER. However, the language of the data used to train the models, regardless of the type of speech, seemed to play an important role, as deep representations pre-trained on French speech were better at predicting French emotional expressions than deep representations pre-trained on English speech.

L-1 This study, however, had certain limitations. For example, the use of purely emotional and spontaneous speech to train deep representations and their performance for AER was not fully investigated. Also, only one self-supervised architecture, Wav2vec2, was used for deep representation learning in this thesis, whereas several others do exist. Moreover, the Wav2vec2 representations were trained on isolated utterances, but then they were used to predict continuous dimensional annotations over long conversations. This discrepency between the pre-training and evaluation of the Wav2vec2 representations might have effected the results and thus findings of this study. This, in turn, has motivated the rest of this study to focus only on the recognition of emotion labels per utterance.

The state of the art on AER from speech signals has further shown that using human transcriptions of a speech signal can further improve the performance of AER for both acted and in-the-wild emotional expressions (see Section 2.2.5). However, the human transcriptions are not always available, while the existing ASR methods have recently become robust enough to be reliably applied to in-the-wild data. This has inspired several studies, such as Heusser et al. (2019); Yoon et al. (2019); Wu et al. (2021); Peng et al. (2021), which show that ASR transcriptions can be used to improve AER performance on acted data. However, this paradigm has not been further investigated for emotional expressions in the wild. Thus, the second research question of this thesis is as follows:

Q-2 How can automatic transcriptions from an existing ASR model be exploited to improve the performance of AER models from speech signals, for both acted and in-the-wild emotional expressions?

C-2 To answer the above question, in Section 4.3, Google's ASR was first used to extract automatic transcriptions of speech signals from two datasets, IEMOCAP (acted) and CMU-MOSEI (in the wild). Then, deep acoustic (W2V2-XLSR-56) and textual (RoBERTa) representations of both human and automatic transcriptions were used in both isolated and joint ways for the AER task. The results showed that joint acoustic-textual representations yielded better performance than either acoustic or textual representations alone for both acted and in-the-wild emotional expressions, regardless of whether the text was human or automatically transcribed. This better performance for in-the-wild emotional expressions was mainly due to the use of textual representations, as acoustic W2V2-XLSR-56 representations were shown to be ineffective for in-the-wild AER.

L-2 The limitation of this study may be the use of only one in-the-wild dataset – CMU-MOSEI– to study the effect of automatic transcriptions for in-the-wild AER.

Moreover, only positive and negative sentiment values were used for this study. The negative-positive scale has shown to be more influenced by the linguistic information, rather than acoustic changes of a speech signal (Goudbeek and Scherer, 2010). This might have been the reason why acoustic representations used in this study could not perform well to recognise sentiment polarities.

F-1 Furthermore, several studies have shown that the valence dimension is detected more accurately from text, whereas the detection of the arousal dimension is more accurate using acoustic signals (Schuller et al., 2020; Triantafyllopoulos et al., 2023). For example, in Triantafyllopoulos et al. (2023), for the scripted utterances of the IEMOCAP dataset, textual features could predict arousal/valence with a CCC of .444/.757, whereas acoustic features predicted arousal/valence with a CCC of .661/.333. Furthermore, the concept of predicting the valence dimension is similar to predicting sentiment via text, for which many datasets already exist. Therefore, a future study could involve using textual datasets to predict sentiment, while training another model on a separate set of audio datasets to predict the arousal dimension.

Nevertheless, state-of-the-art AER models and the studies in Chapter 4 have shown that both verbal and non-verbal information can be useful for predicting both arousal and valence dimensions, or different emotion categories. In other words, emotion can be conveyed by both what is said and how it is said. Recent studies further suggest that how a sentence is uttered, and the words that make up that sentence, can be greatly influenced by who the speaker is (see Section 2.2.6). In particular, latent speaker representations computed from speaker recognition models have been shown to perform better for AER than traditional acoustic representations. However, the fusion of speaker representations with deep acoustic and textual representations, which has recently been achieved in state-of-the-art performances (see C-2 above), has not yet been explored. This has raised another question in this thesis, which is as follows:

Q-3 Given that speaker recognition models can provide us with latent speaker representations, how can we improve the performance of deep acoustic and textual representations for AER, by fusing them with such speaker representations?

C-3 The question above was investigated in Section 4.4 by first training speaker recognition models from acoustic –W2V2-XLSR-56–, textual –RoBERTa–, and joint acoustic-textual representations (see C2 above). Then, the latent speaker embeddings was fused with latent acoustic, textual, or joint acoustic-textual latent representations to form speaker-aware representations for AER on the IEMOCAP dataset. The results showed that the fusion of speaker representations with joint acoustic-textual –W2V2-RoBERTa– representations achieves the best results among the aforementioned scenarios, although it was marginally better than joint acoustic-textual representations without the speaker representations. In addition, whisper representations pre-trained directly for speech recognition and indirectly for speaker recognition were also investigated and showed similar results to the

speaker-aware joint W2V2-RoBERTa representations. These results indicated that by integrating both linguistic and speaker information into acoustic representations, one could expect better results for AER than by relying on acoustic information alone. It was further shown that by using deep representations such as Whisper, which have been pre-trained directly for ASR and indirectly for speaker recognition, one might not need to further fuse textual and speaker representations to achieve comparable performance to speaker-aware joint W2V2-RoBERTa representations for acted AER.

L-3 The limitations of this study were that the evaluation was performed on only one acted dataset (IEMOCAP), the speaker recognition models were trained on only eight speakers, and the speaker-aware representations for the AER task were evaluated on only two speakers.

F-2 Therefore, a future direction may be to consider in-the-wild emotional expressions with more speakers for speaker-aware acoustic and textual representations.

Furthermore, the research questions above (Q1 to Q3) focus only on how to achieve more effective representations of acoustic signals or text in order to predict specific emotion annotations, based on each dataset used separately. However, as each dataset covers a limited range of emotional expressions, it is important to use multiple datasets to train AER models in order to obtain generalised models. Training machine learning models for different emotion datasets can be challenging, because numerical representations of emotion are defined in various subjective manners across different datasets (see Section 1.1.1). To overcome this challenge, current state of the art uses MTL to train multi-corpus AER models by using traditional acoustic features. In particular, MTL can be used to first provide us with a latent emotion representation that is shared across corpora. On the other hand, deep pre-trained representations such as Wav2vec2 have recently been shown to be more performant for AER than traditional acoustic features. However, MTL-based multi-corpus training has not been investigated for acted or in-the-wild AER from acoustic representations or text. This motivated the next research question:

Q-4 How effective is the latent emotion representation, computed by using the MTL-based method using deep representations, in recognising the same or similar emotions across different corpora that might use different emotion annotation schemes?

C-4 To answer the above question, in Section 5.1, an MTL-based method was proposed that uses deep acoustic representations, a shared GRU model for all datasets, and linear classifiers dedicated to each dataset. The proposed method was then evaluated using four acted datasets from CaFE, EmoDB, GEMEP and RAVDESS, all of which use different sets of emotion labels. The results suggested that the proposed method could produce a latent emotion representation that could improve the performance of AER across datasets with different sets of emotion labels. Furthermore, fine-tuning the deep representations while training the shared GRU model further showed that each fine-tuned model learns high specificity of each dataset, at

the cost of lower generalisation across the datasets.

L-4 However, in the experiments mentioned above, all the datasets used to evaluate the proposed method were rather small acted datasets. Also, only the effectiveness of deep acoustic representations was studied. In order to further investigate this method for in-the-wild emotional expressions, and both acoustic and textual modalities, the next research question was as follows:

Q-5 By using the proposed multi-corpus training method, can acted emotional expressions be useful in improving in-the-wild AER from acoustic signals and text?

C-5 The above question was investigated in Section 5.2 by evaluating the method mentioned in C-4 (proposed in Section 5.1) on the IEMOCAP (acted) and CMU-MOSEI (in the wild) datasets, and for deep acoustic –W2V2-XLSR-56– and textual –RoBERTa– representations. The results suggested that by using the proposed MTL-based method, the AER performance of the CMU-MOSEI dataset could be improved when using textual or joint acoustic-textual representations, but not with acoustic representations alone. This was due to the fact that the W2V2-XLSR-56 acoustic representations used in this study did not perform well for AER on the CMU-MOSEI dataset, either in multi-corpus or single-corpus training strategies.

L-5 Similar to what was said above for L2, this study also suffers from using only one in-the-wild dataset –CMU-MOSEI– with negative and positive sentiments as emotion targets. Moreover, in this study, the proposed MTL-based method was only evaluated for within-corpus settings.

F-3 Therefore, cross-corpus analysis of multi-corpus training for different in-the-wild datasets can be considered for a future research direction. Also, as the above study only exploited datasets with per-utterance emotion labels, multi-corpus training for datasets with continuous dimensional annotations may also be a future research direction.

In summary, this thesis has made several contributions to solving some of the existing challenges in AER from acoustic signals and text for emotional expressions in the wild. The results of the experiments carried out during this thesis have suggested that current ASR technologies, combined with deep textual representations, can be particularly effective. Furthermore, even if annotated in a different and subjective way, acted data can be used with the MTL-based method proposed in this thesis, in order to achieve good performance for AER in the wild. In addition, this section suggested several other works as future research directions. Beyond this thesis and the future works discussed above, what is underway after the writing of this thesis is discussed below.

6.2 Beyond this thesis

This thesis has focused on developing AER in the wild from acoustic signals and text. Although there are still some limitations to this technology, it can be applied

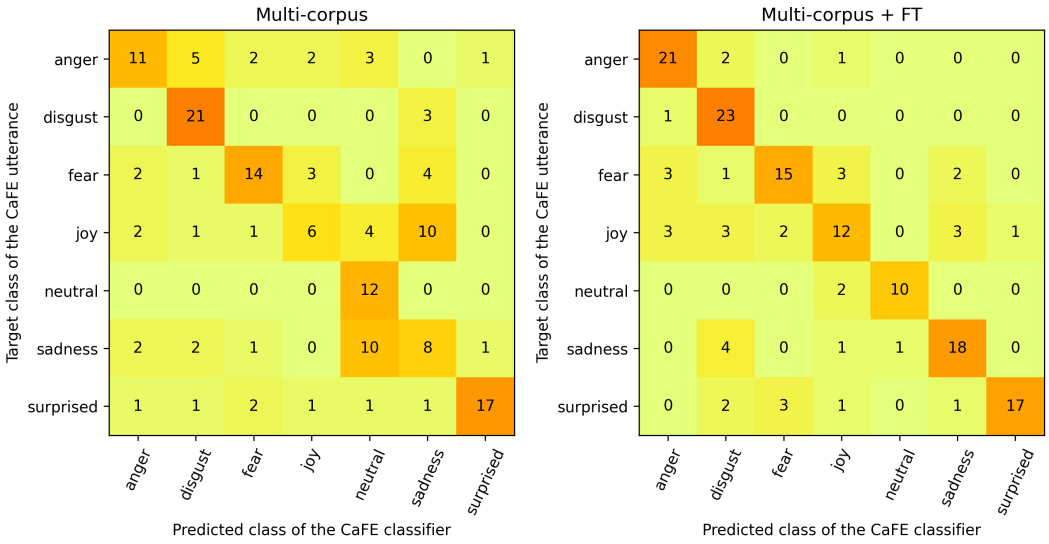
to many domains, including education, entertainment, marketing, and health (see Chapter 1). One of the applications of this thesis is in AI-enhanced digital therapy, within the THERADIA project (Tarpin-Bernard et al., 2021). THERADIA consists of the development of an empathic virtual therapeutic assistant that acts as an interface between the patient and the patient's therapist or carer. The virtual interactive assistant is specifically targeted at people suffering from cognitive disorders such as Alzheimer's disease, dyslexia and stroke. The effectiveness of interactive therapies can be improved by allowing patients to go beyond their limited face-to-face therapy sessions by pursuing them at home while being accompanied by an empathic virtual assistant. The empathic assistant will not only monitor the patient's emotions, but will also be able to respond to them.

The technological building blocks for implementing an AER system for an application such as THERADIA have also been developed during this thesis. And an interactive demo has also been put online, which is publicly available for everyone and is completely open source (see Appendix C). Furthermore, the use of the video modality and the wide range of emotional annotations, collected for the THERADIA project according to the appraisal theory (see section 2.1.1), are among the next steps beyond this thesis.

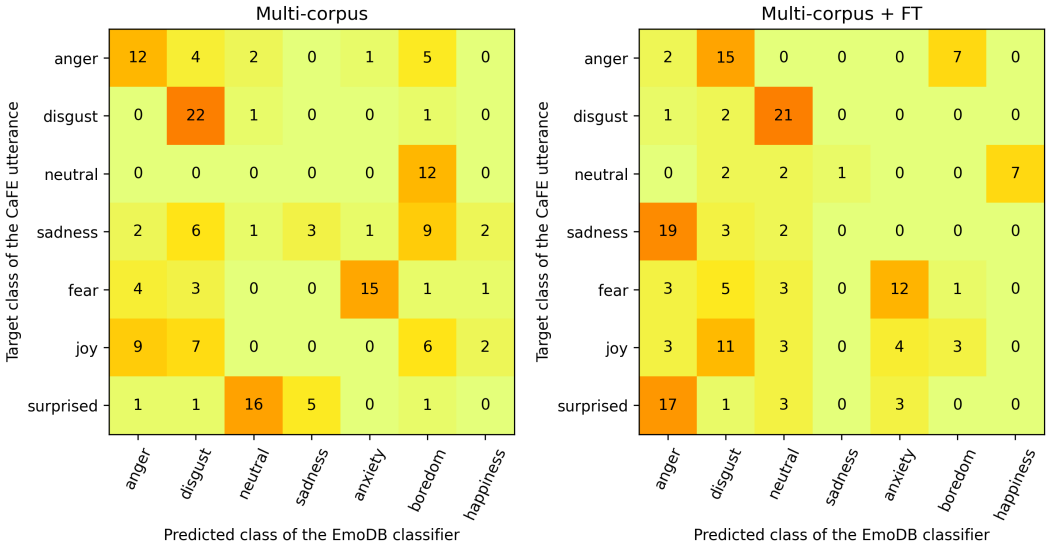
Appendix A

Supplementary results

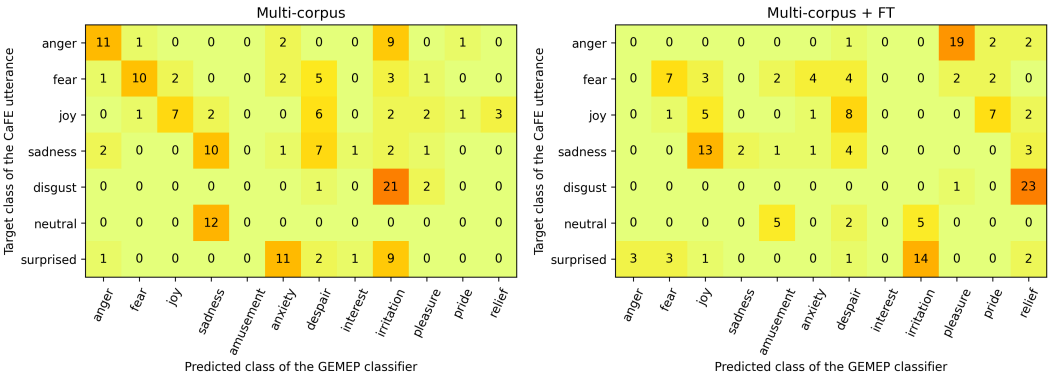
This appendix presents all the confusion matrices related to the cross-corpus evaluation of the multi-corpus training method of section 5.1, without (frozen) and with fine-tuning (FT) of the W2V2-XLSR-56 representations.



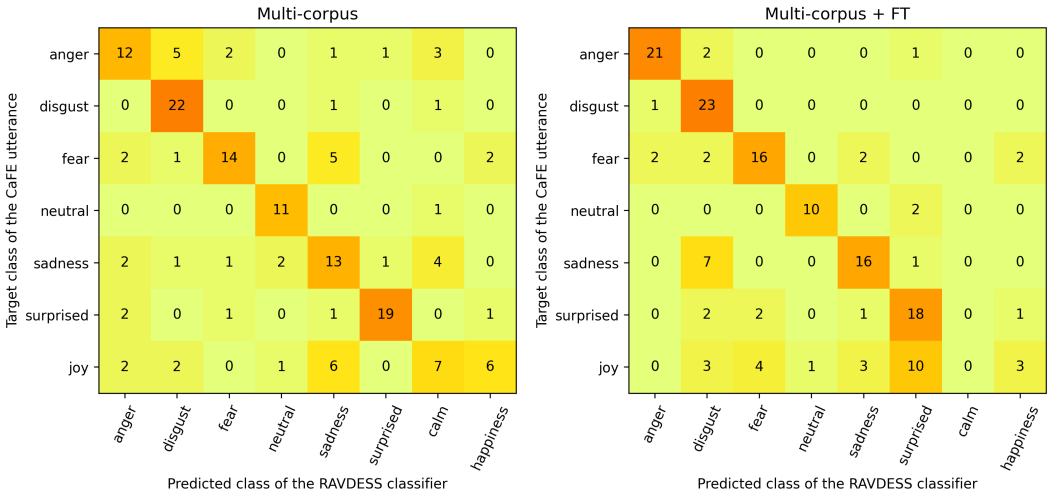
Confusion matrices of the predictions of the CaFE classifier from the CaFE utterances.



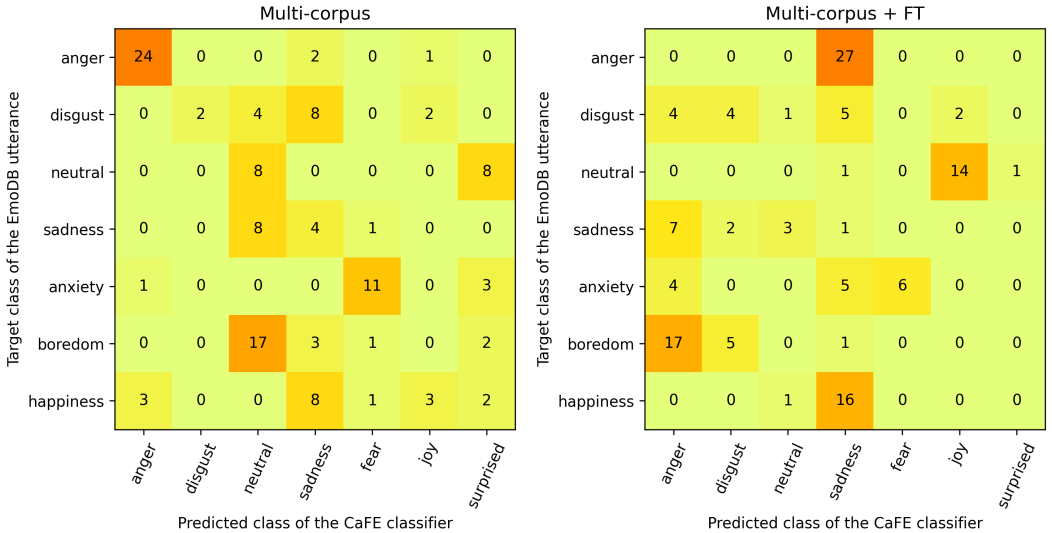
Confusion matrices of the predictions of the CaFE classifier from the EmoDB utterances.



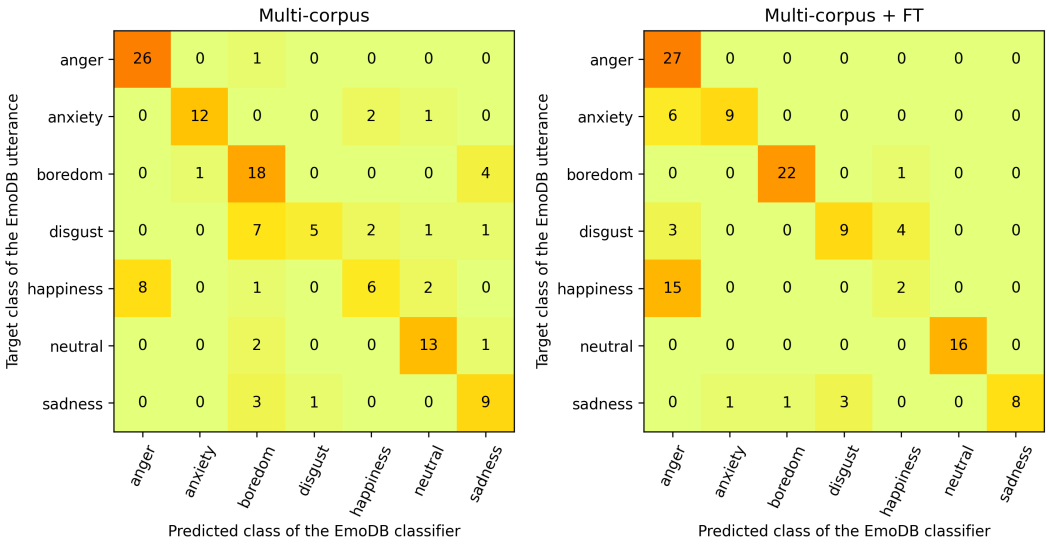
Confusion matrices of the predictions of the CaFE classifier from the GEMEP utterances.



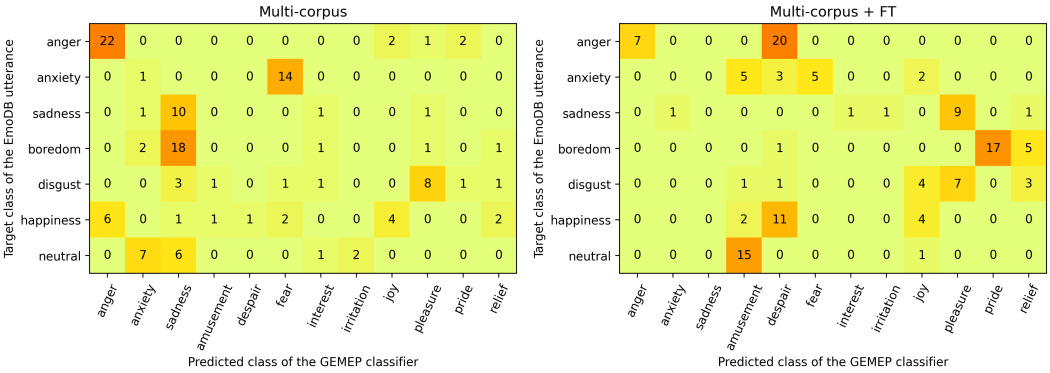
Confusion matrices of the predictions of the CaFE classifier from the RAVDESS utterances.



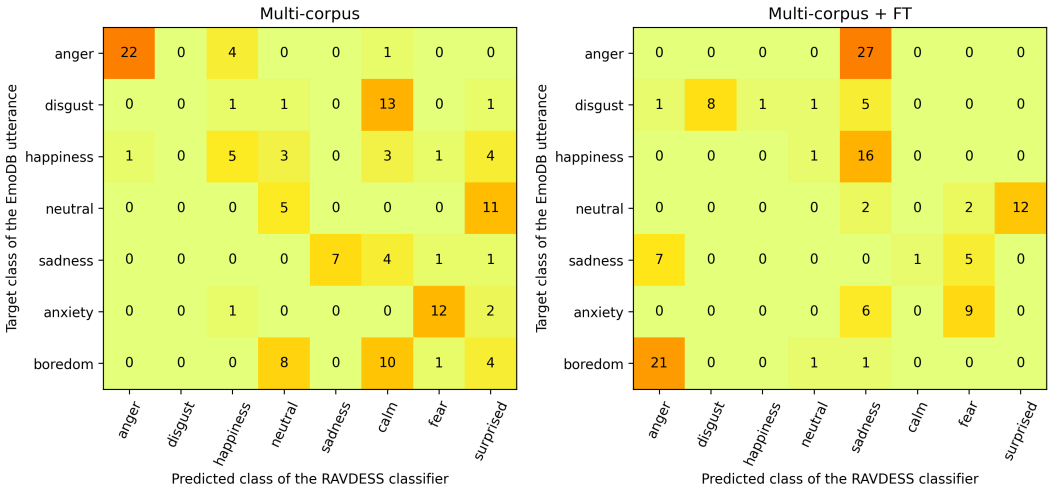
Confusion matrices of the predictions of the EmoDB classifier from the CaFE utterances.



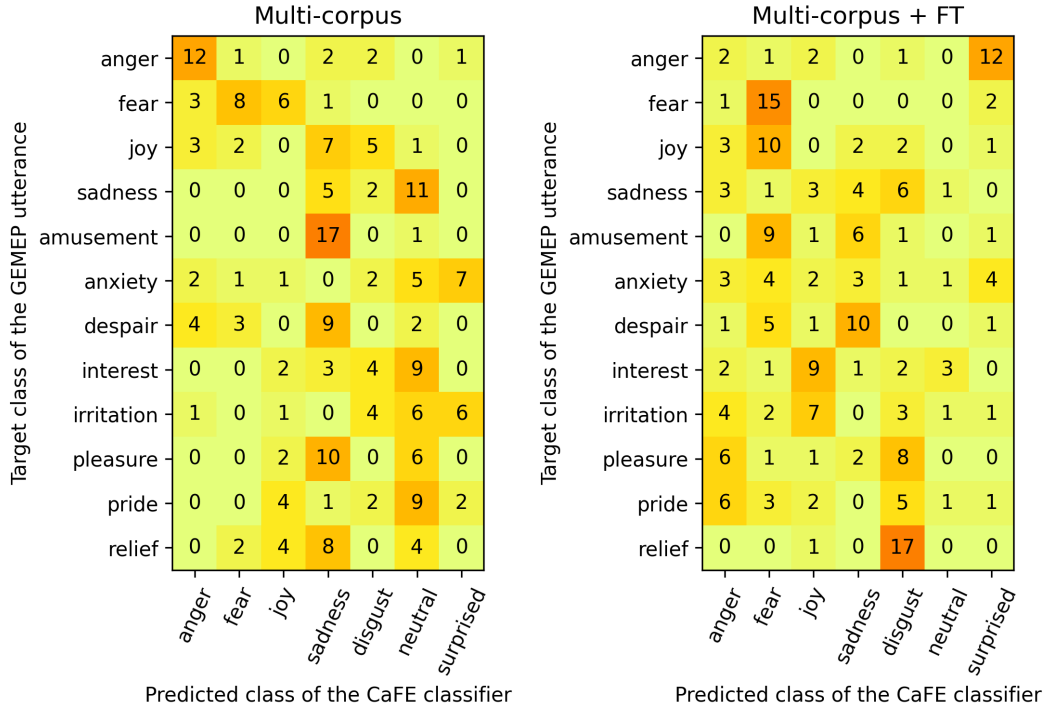
Confusion matrices of the predictions of the EmoDB classifier from the EmoDB utterances.



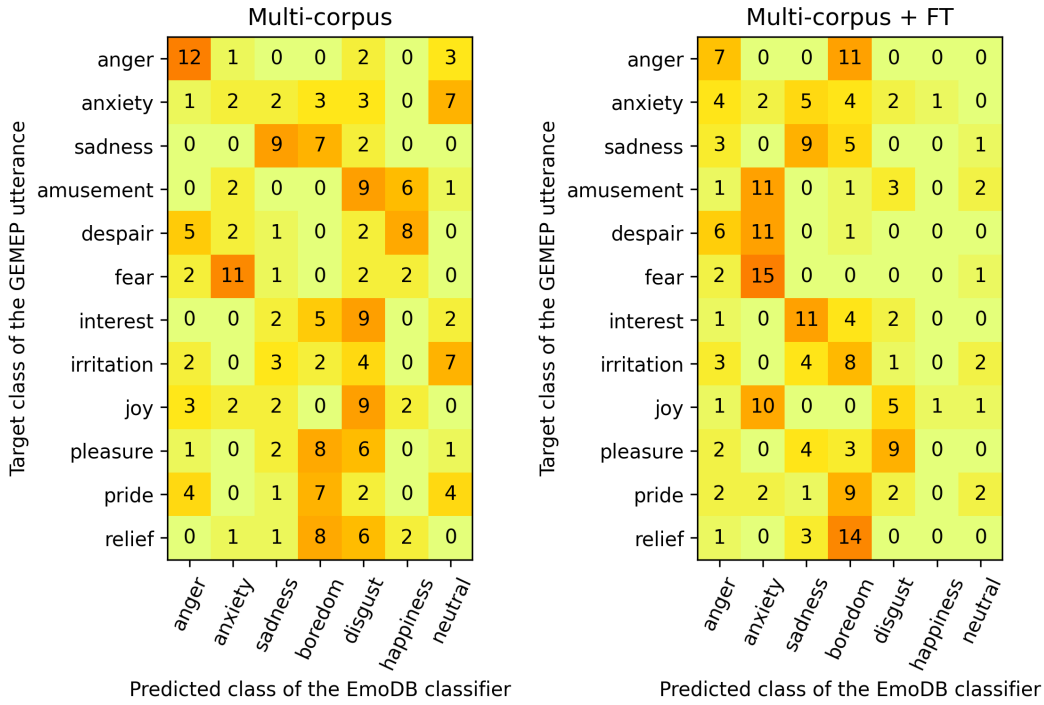
Confusion matrices of the predictions of the EmoDB classifier from the GEMEP utterances.



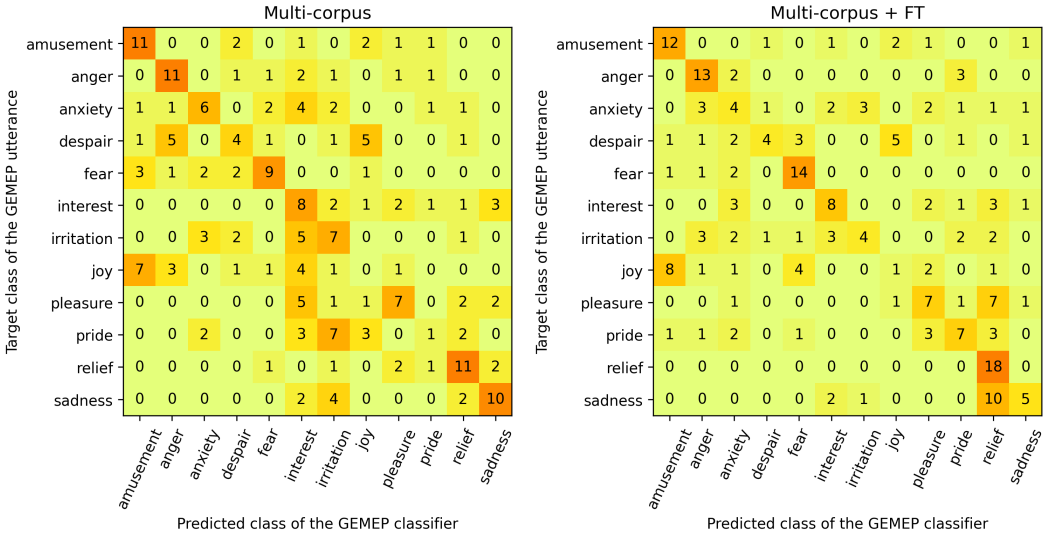
Confusion matrices of the predictions of the EmoDB classifier from the RAVDESS utterances.



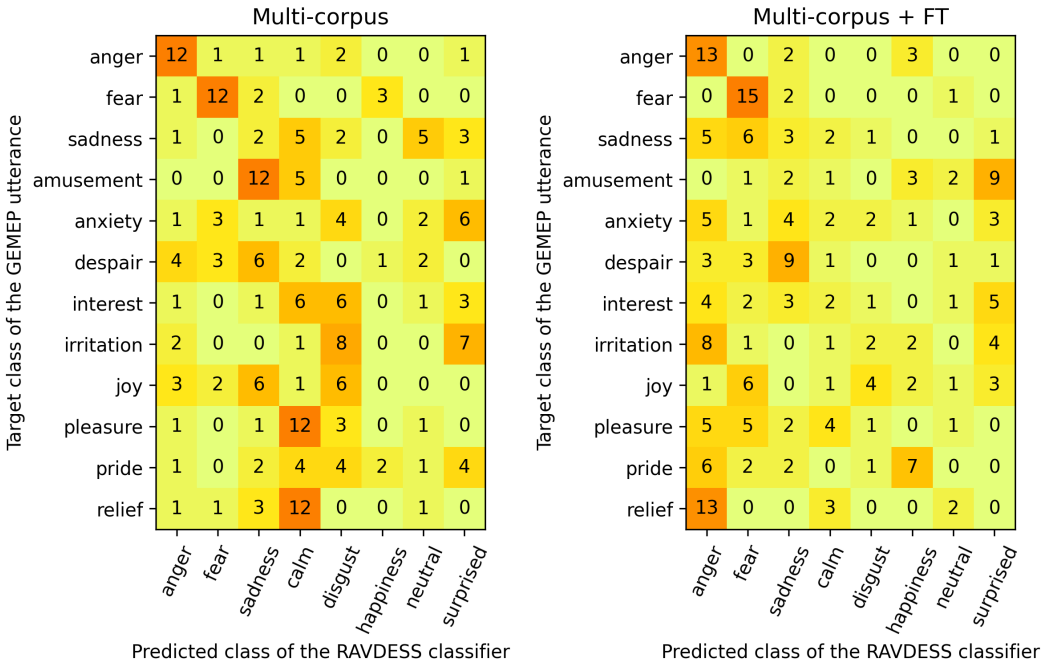
Confusion matrices of the predictions of the GEMEP classifier from the CaFE utterances.



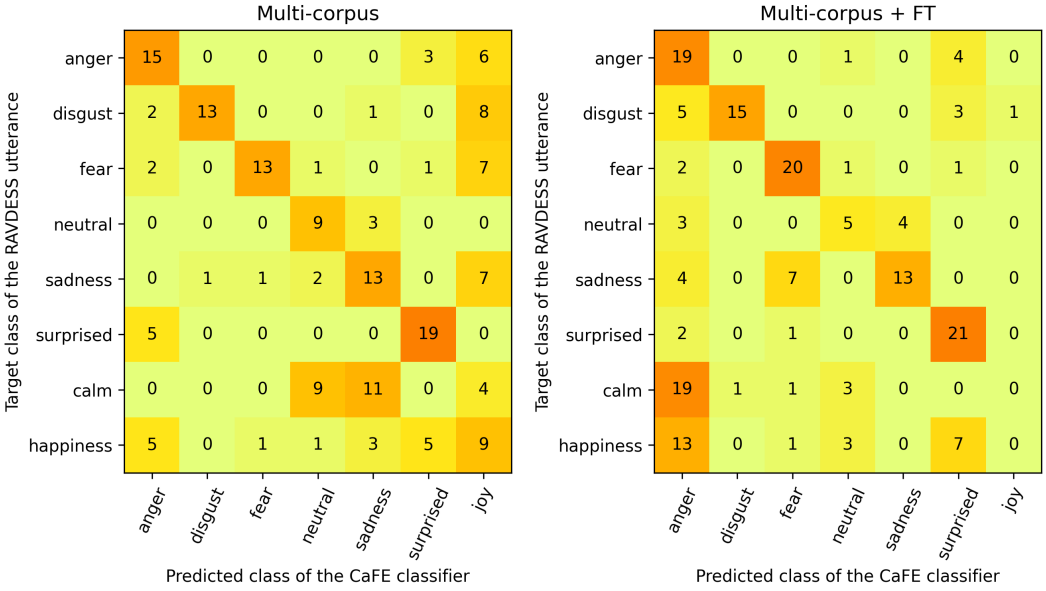
Confusion matrices of the predictions of the GEMEP classifier from the EmoDB utterances.



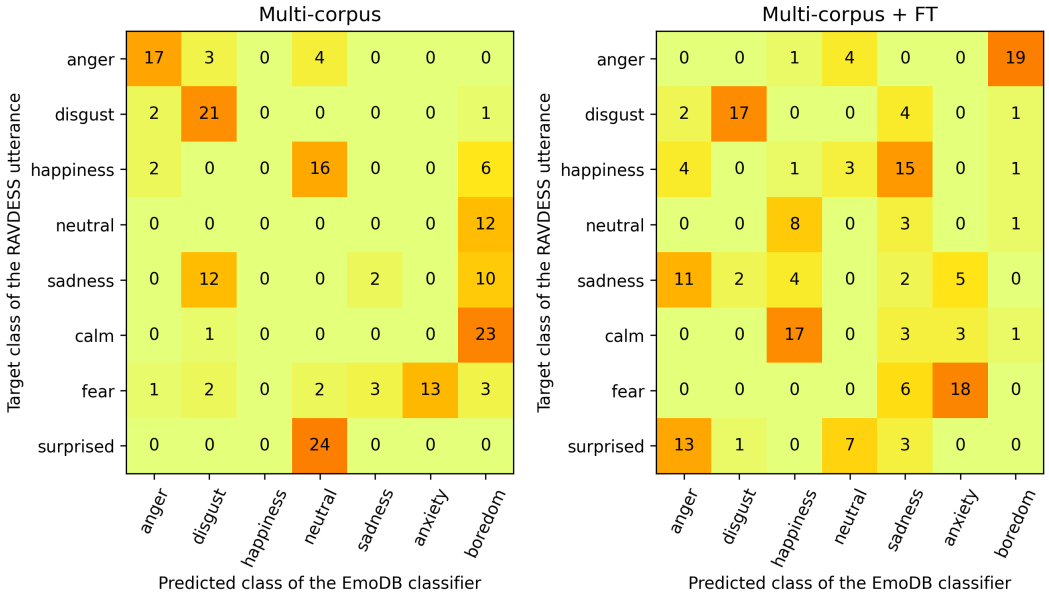
Confusion matrices of the predictions of the GEMEP classifier from the GEMEP utterances.



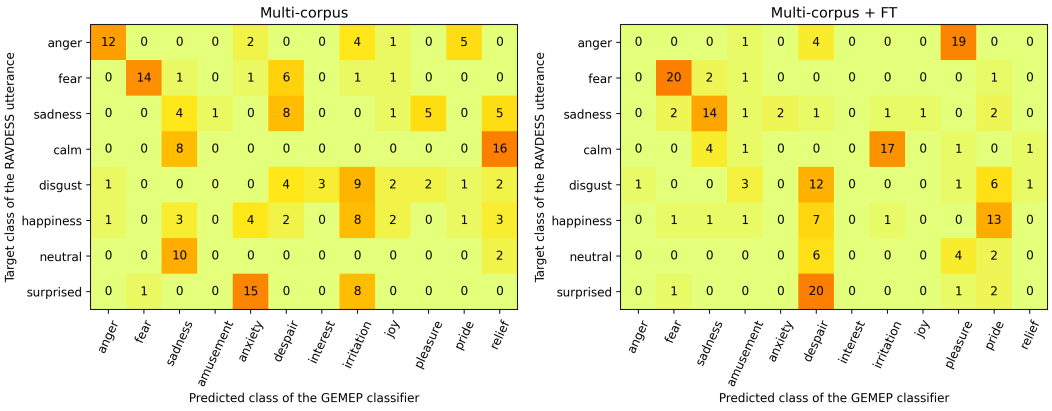
Confusion matrices of the predictions of the GEMEP classifier from the RAVDESS utterances.



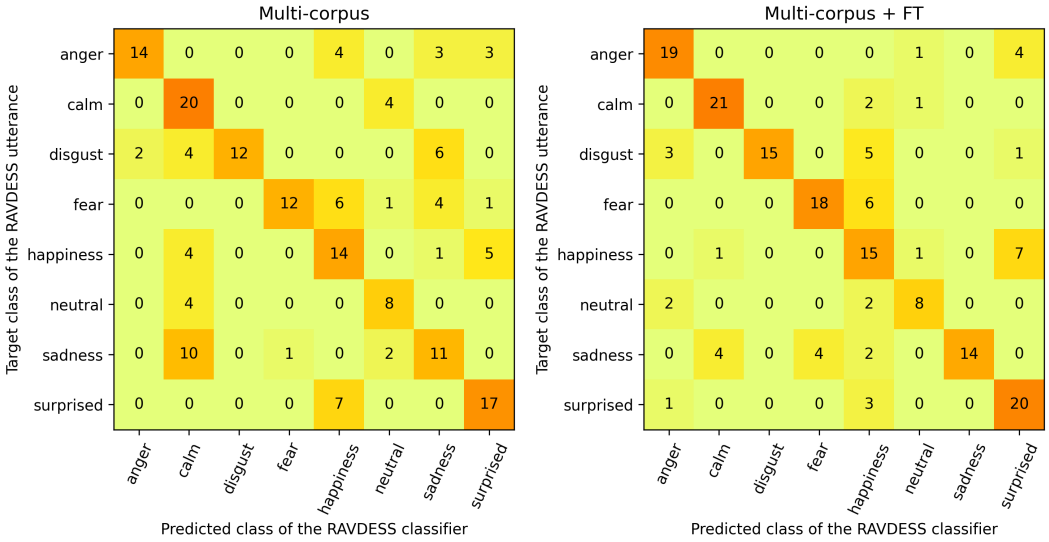
Confusion matrices of the predictions of the RAVDESS classifier from the CaFE utterances.



Confusion matrices of the predictions of the RAVDESS classifier from the EmoDB utterances.



Confusion matrices of the predictions of the RAVDESS classifier from the GEMEP utterances.



Confusion matrices of the predictions of the RAVDESS classifier from the RAVDESS utterances.

Appendix B

Publications

As part of this thesis, the following papers have been published:

- Evain, S., Nguyen, H., Le, H., Boito, M. Z., Mdhaffar, S., **Alisamir, S.**, Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021a). Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech. In *Proc. Interspeech 2021*, pages 1439–1443. Brno, Czech Republic.

- Evain, S., Nguyen, M. H., Le, H., Zanon Boito, M., Mdhaffar, S., **Alisamir, S.**, Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021b). Task agnostic and task specific self-supervised learning from speech with lebenchmark. In Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021), NeurIPS 2021 Datasets and Benchmarks Track, Online, USA.

- **Alisamir, S.** and Ringeval, F. (2021). On the evolution of speech representations for affective computing: A brief history and critical overview. *IEEE Signal Processing Magazine*, 38(6):12–21.

- **Alisamir, S.**, Ringeval, F., and Portet, F. (2022). Multi-corpus affect recognition with emotion embeddings and self-supervised representations of speech. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)*. Nara, Japan. IEEE.

- Evain, S., Nguyen, M. H., Le, H., Zanon Boito, M., Mdhaffar, S., **Alisamir, S.**, Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2022). Modèles neuronaux pré-appris par auto-supervision sur

des enregistrements de parole en français. In *JEP 2022*. Île de Noirmoutier, France.

Several other related studies were also carried out during this thesis, which are not included in the main text, including the following two *arXiv*[1] papers:

- **Alisamir, S.**, Ringeval, F., and Portet, F. (2022). Cross-domain Voice Activity Detection with Self-Supervised Representations. *arXiv preprint arXiv:2209.11061*.
- **Alisamir, S.**, Ringeval, F., and Portet, F. (2022). Dynamic Time-Alignment of Dimensional Annotations of Emotion using Recurrent Neural Networks. *arXiv preprint arXiv:2209.10223*.

[1] https://arxiv.org/

Appendix C

Demo

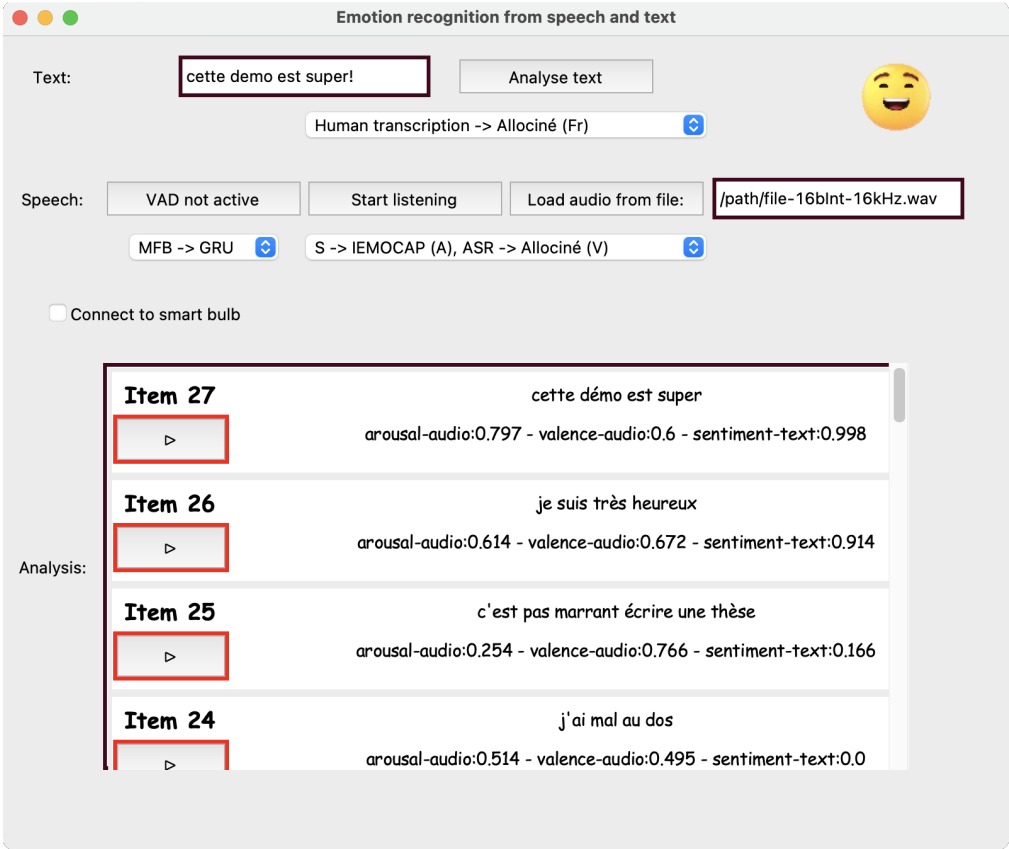


Figure C.1: The interface of the real-time emotion recognition demo built during this thesis. Based on a given text input or a spoken utterance, the demo can analyse and display the detected emotion in real time.

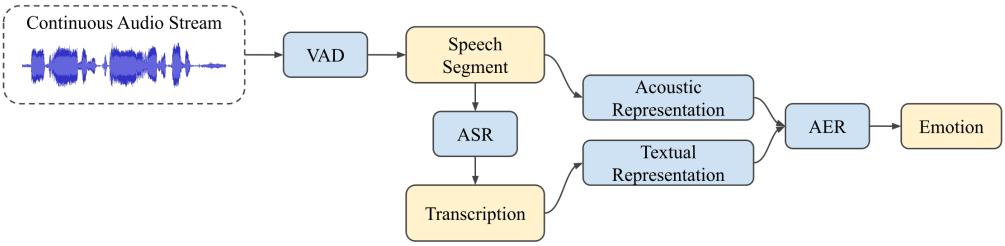


Figure C.2: The pipeline of the real-time emotion recognition demo built during this thesis.

During this thesis, a real-time[1] emotion recognition demo was also built to both present different trained models and to understand all the technical aspects and problems that can arise from theory to application. This demo is available on the Github[2] for anyone to try out. Figure C.1 shows the interface of this demo. The top part of the interface is where the user can interact and the bottom part is where the results of the analysis are displayed. At the very top, the user can enter a text as input and click on the button "Analyse text" to run the sentiment prediction model, which will display the results on a range from zero to one.

In addition, the user can also activate a Voice Activity Detection (VAD) trained during this thesis Alisamir et al. (2022a) by clicking on the button "VAD not active" to activate the VAD. Then, the microphone should be activated by clicking on the button "Start listening". The VAD will then automatically detect if each time a phrase is spoken, and the detected segment is passed to the emotion recognition model to make predictions based on the acoustic signal and the transcription[3]. For example, if the model selected is "S -> IEMOCAP, ASR -> Allociné", which means that the model trained on the IEMOCAP dataset based on speech is used and the text predictions are based on a model trained on the Allociné dataset. The model trained on the IEMOCAP dataset can predict the arousal and valence dimensions of emotion (on a scale from zero to one), and therefore the results presented in the Analysis section will be in the same format.

To better understand how the demo works, Figure C.2 shows its pipeline. The VAD first detects a speech segment from a constant active audio stream input. Then the Google's ASR transcribes the speech segment. The speech segment and the transcription are then used to extract acoustic and textual representations respectively, which are then used in the AER model to predict emotion. It should be noted that depending on the representation, the dataset and the model trained on it, which can be selected in the interface, the predicted emotion may have a different value and format.

[1] By "real-time", we mean roughly around 100ms up to ten seconds, following Eyben (2015).
[2] https://github.com/SinaAlisamir/Real-time-Emotion-Recognition-from-Speech-and-Text
[3] The transcriptions are obtained automatically by exploiting Google's ASR model

Appendix D

Résumé de thèse

D.1 Introduction

L'émotion et la raison ont longtemps été considérées comme deux parties indépendantes et rivales du cerveau humain. Ce n'est que récemment que la corrélation entre les deux a été clairement établie grâce à une lésion cérébrale (Damasio, 1994). Des découvertes psychologiques récentes suggèrent que ce n'est pas seulement le raisonnement qui est nécessaire pour obtenir une réponse émotionnelle, mais que l'émotion est également nécessaire pour raisonner et filtrer toutes les informations sensorielles constantes que notre cerveau reçoit à chaque seconde de notre vie. Permettre aux machines de percevoir les émotions humaines peut être une technologie révolutionnaire, ayant un impact sur de nombreux domaines différents, allant de l'éducation et de la santé au divertissement.

Pour que les différents cas d'utilisation technologique de la reconnaissance automatique des émotions (RAE) puissent voir le jour, il faut qu'elle soit performante sur les données naturelles[1], qui présentent plusieurs caractéristiques, à savoir que différents locuteurs ont des interactions naturelles, utilisent différents microphones et se trouvent dans différents environnements (Kossaifi et al., 2021). Cependant, à ce jour, la recherche sur la RAE s'est principalement concentrée sur les expressions actées des émotions, enregistrées dans des environnements de laboratoire. Néanmoins, ce paradigme évolue progressivement vers l'exploitation de données recueillies dans la nature. Ce changement de paradigme est principalement dû à de nouvelles techniques d'apprentissage profond, telles que les représentations profondes, qui peuvent nous fournir des représentations acoustiques et textuelles

[1] Ici, le terme "naturel" dans le contexte de la RAE fait référence aux expressions émotionnelles qui sont le résultat d'interactions naturelles entre des humains, ou des humains et une machine, enregistrées dans une variété d'environnements, tels que dans une salle de classe, en public, ou à la maison, et en utilisant différents microphones. Les données naturelles s'opposent aux expressions émotionnelles actées (ou induites), qui sont généralement recueillies dans des environnements de laboratoire contrôlés.

mieux adaptées aux expressions émotionnelles naturelles. Les nouvelles méthodes d'apprentissage profond, ainsi que les vastes possibilités technologiques de la RAE, ont suscité l'intérêt de l'industrie ces dernières années. Par exemple, Atos, une entreprise de services numériques basée en France et le partenaire industriel de cette thèse CIFRE[1], est intéressée par le développement de la technologie RAE pour les applications d'assistants virtuels qui fonctionnent avec des entrées acoustiques et textuelles. En se basant sur une telle application cible, cette thèse se concentre principalement sur les modalités acoustiques et textuelles, provenant de différents locuteurs, avec des données naturelles. Les objectifs de cette thèse sont donc les suivants :

1. L'état de l'art a montré des résultats prometteurs avec les transcriptions RAP[2] pour les représentations acoustiques-textuelles conjointes et les représentations acoustiques profondes tenant compte du locuteur. Ainsi, dans cette thèse (voir la section D.4), la recherche connexe est avancée en étudiant l'utilisation de représentations profondes pour les expressions émotionnelles actées et naturelles, lorsque le texte est transcrit soit par des humains, soit par un système RAP.

2. L'état de l'art a montré l'efficacité de l'utilisation de la MTL[3] pour former des modèles d'apprentissage automatique sur plusieurs corpus avec différents ensembles de catégories d'émotions. D'autre part, il a été démontré que les représentations profondes pré-entraînées se généralisent bien à travers différentes expressions émotionnelles. Par conséquent, dans la section D.5, cette thèse propose une méthode qui utilise des représentations acoustiques et textuelles profondes avec MTL pour prédire une représentation latente de l'émotion qui peut reconnaître les mêmes catégories d'émotion ou des catégories similaires dans différents ensembles de données.

Dans ce qui suit, la section D.2 présente l'état de l'art de la RAE à partir de signaux acoustiques et de textes. La section commence par discuter de la façon dont l'émotion est perçue en psychologie et comment elle est utilisée pour définir les annotations d'émotion pour l'apprentissage automatique. Il passe ensuite en revue plusieurs méthodes RAE de pointe, suivies d'une étude de cas pour les comparer quantitativement.

[1]CIFRE (Conventions Industrielles de Formation par la REcherche) est un mécanisme qui permet aux entreprises de droit français de recruter un doctorant dont le projet de recherche est mené en collaboration avec un laboratoire public (dans ce cas, l'Université de Grenoble Alpes), le ministère français de la recherche versant une subvention annuelle à l'entreprise (dans ce cas, Atos).

[2]RAP: Reconnaissance Automatique des Émotions

[3]Multi-Task Learning (MTL): l'Apprentissage multitâche

Basé sur l'étude de l'état de l'art, la section D.3 décrit ensuite la méthodologie des expériences dans cette thèse, à savoir les représentations, les ensembles de données, les méthodes d'apprentissage, les fonctions de perte et les métriques utilisées pour atteindre les objectifs de la thèse.

Puis dans la section D.4, des représentations profondes pré-entraînées de signaux acoustiques et de texte, qui ont montré les meilleures performances dans la littérature, sont utilisées pour RAE avec des expressions émotionnelles actées et naturelles. En outre, cette section expérimente l'utilisation de représentations acoustiques-textuelles conjointes pré-entraînées personnalisées, lorsque le texte est soit une transcription humaine, soit généré par un RAP.

Ensuite, dans la section D.5, la méthode proposée utilisant les représentations profondes pré-entraînées avec MTL est explorée. L'efficacité de l'utilisation d'expressions émotionnelles actées pour RAE avec des données naturelles est également expérimentée.

Enfin, la section D.6 conclut la thèse.

D.2 l'État de l'art

La RAE à partir de signaux acoustiques et de textes a été un domaine d'étude au cours des dernières décennies. Bien que l'émotion n'ait pas de définition standard en psychologie, la RAE cible généralement une vision catégorielle de l'émotion telle que la peur, la colère, le bonheur, la tristesse, le dégoût et la surprise, basée sur les travaux d'Ekman, ou les dimensions d'éveil et de valence (plaisir intrinsèque), basées sur les travaux de Russell. Afin de prédire les annotations catégorielles ou dimensionnelles des émotions, les méthodes traditionnelles de la RAE impliquent plusieurs étapes de transformation des données pour modéliser le signal acoustique ou le texte à différents niveaux. Ces étapes sont 1) l'extraction de caractéristiques de bas niveau, qui implique des techniques de modélisation au niveau du signal comme les MFB 2) des approches statistiques pour parvenir à une modélisation contextuelle, qui implique des méthodes comme BoAW pour les signaux acoustiques et TF-IDF pour le texte, et 3) la mise en correspondance des caractéristiques statistiques avec des représentations numériques d'une annotation d'émotion, en utilisant différentes méthodes statistiques, et plus tard des méthodes d'apprentissage automatique comme les SVM. Cependant, l'avènement des DNN a remis en question ce paradigme et a devenu populaires au cours des dernières années. Cela est dû en grande partie à la capacité des réseaux neuronaux profonds à approximer des fonctions complexes en utilisant uniquement des données (axées sur les données), par opposition aux techniques traditionnelles qui sont principalement "fondées sur la connaissance humaine". En outre, comme les couches neuronales profondes peuvent être montées en cascade, elles peuvent être formées pour remplacer efficacement toutes les étapes de transformation des données susmentionnées, en brouil-

lant les frontières entre chaque étape. En outre, l'utilisation de DNNs préformés sur de grandes quantités de données non étiquetées, tels que le modèle Wav2vec2 pour les signaux acoustiques et BERT pour le texte, a dominé les meilleures performances dans de nombreux domaines, y compris la RAE. Ces modèles préformés sont particulièrement efficaces pour la RAE car ils sont formés de manière non supervisée et ne sont pas formés par des annotations d'émotions subjectives, qui sont souvent bruyantes. En outre, l'utilisation de représentations acoustiques et textuelles conjointes, où le texte peut être soit des transcriptions humaines, soit produit par un RAP, s'est avérée plus efficace que l'utilisation d'une modalité acoustique ou textuelle seule. Des études récentes suggèrent également que cette amélioration peut être renforcée par l'exploitation des informations relatives au locuteur. Cependant, aucune étude n'a pu être trouvée sur l'utilisation de représentations acoustiques-textuelles profondes pré-entraînées augmentées d'informations sur le locuteur. L'utilisation de représentations acoustiques-textuelles pour les expressions émotionnelles naturelles semble également constituer une lacune dans l'état actuel de la RAE.

D.3 Méthodologie expérimentale et ressources

La revue de l'état de l'art a montré que les représentations profondes pré-entraînées des signaux acoustiques et du texte sont significativement plus performantes que les techniques traditionnelles pour la RAE. L'objectif de cette thèse est d'étendre l'état de l'art relatif à l'utilisation de représentations acoustiques et textuelles profondes pré-entraînées pour une large gamme d'expressions émotionnelles, qu'elles soient actées ou naturelles. Par conséquent, plusieurs représentations profondes pré-entraînées sont utilisées dans différentes expériences de cette thèse, ainsi que dans un large éventail d'ensembles de données. Les représentations profondes pré-entraînées utilisées ici sont Wav2vec2, et Whisper pour les signaux acoustiques et RoBERTa pour le texte, choisies en particulier pour leurs performances supérieures à l'état de l'art. Les ensembles de données utilisés dans cette thèse sont AlloSat, CMU-MOSEI, CaFE, EmoDB, GEMEP, IEMOCAP, RAVDESS et RECOLA, qui varient en termes d'environnement d'enregistrement, de locuteurs, d'annotations d'émotions et de contexte dans lequel les expressions émotionnelles sont collectées (actées, induites et naturelles). En outre, dans les différentes expériences menées tout au long de cette thèse, l'optimiseur Adam est choisi pour entraîner les modèles neuronaux pour différentes fonctions de perte, car il converge plus rapidement et peut obtenir des résultats comparables ou supérieurs à ceux de l'algorithme SGD[1] de base. Les fonctions de perte et les mesures utilisées pour former et évaluer les modèles varient en fonction de la tâche. Pour la prédiction continue dans le temps

[1] Stochastic Gradient Descent (SGD): Descente stochastique de gradient

de dimensions émotionnelles telles que l'excitation et la valence, $1-CCC$ est utilisé comme fonction de perte et CCC[1] comme métrique, parce qu'il peut mesurer à la fois la covariance des prédictions et des cibles et la distance entre leurs moyennes. Et pour catégoriser les étiquettes d'émotion telles que la joie, la colère, la tristesse et la neutralité, l'entropie croisée est utilisée comme fonction de perte et l'UAR[2] comme mesure d'évaluation. Le choix de l'UAR par rapport à la mesure de précision plus courante s'explique par le fait que la précision ne donne pas une bonne indication des performances d'un modèle si les données de test ne sont pas équilibrées. Enfin, toutes les expériences ont été réalisées à l'aide de Pytorch et de la boîte à outils SpeechBrain.

D.4 Représentations profondes pour la prédiction des émotions

Les avancées récentes dans le domaine des DNN ont montré que les représentations profondes pré-entraînées, et en particulier les représentations Wav2vec2 et RoBERTa, peuvent atteindre des performances de pointe dans un large éventail de tâches vocales et textuelles, y compris la reconnaissance des émotions. Comme ces représentations profondes sont formées de manière incrémentale sur de grandes quantités de données, la relation entre leurs données d'entraînement et leur performance pour une tâche telle que la reconnaissance des émotions est difficile à comprendre et à contrôler. À cette fin, dans cette thèse, plusieurs modèles Wav2vec2 ont été pré-entraînés sur différentes quantités et différents types de discours, et leurs performances dans reconnaissance dimensionnelle des émotions ont été analysées sur les ensembles de données RECOLA et AlloSat. Les résultats ont montré que le type de données joue un rôle plus important que la quantité de données. En particulier, les représentations Wav2vec2 pré-entraînées sur des données françaises sont plus performantes dans les tâches de prédiction des émotions en français que les représentations Wav2vec2 pré-entraînées sur des données anglaises. Cependant, le pré-entraînement des représentations profondes sur un plus grand nombre de données ne conduit pas nécessairement à une meilleure performance pour la RAE. De plus, par rapport aux caractéristiques MFB traditionnelles, ces représentations profondes s'appuient sur des modèles moins complexes pour obtenir de bonnes performances dans la RAE.

En outre, les émotions étant transmises à la fois par la communication verbale et non verbale, il a été démontré que la représentation conjointe acoustique-textuelle des représentations acoustiques et textuelles était plus efficace que l'utilisation de chaque modalité seule. Par conséquent, dans cette thèse, plusieurs méthodes de

[1] CCC: Concordance Coefficient de Corrélation

[2] Unweighted Average Recall (UAR): Moyenne non pondérée du rappel

fusion des représentations acoustiques (W2V2-XLSR-56) et textuelles (RoBERTa) ont été étudiées sur les ensembles de données IEMOCAP (pour les expressions agies) et CMU-MOSEI (pour les expressions naturelles). Les résultats ont montré que la simple concaténation des vecteurs latents des représentations acoustiques et textuelles pré-entraînées, permettait d'obtenir de meilleures performances pour les expressions émotionnelles actées et naturelles, que l'utilisation de la représentation de chaque modalité seule. Ces résultats sont cohérents avec ceux observés dans des études similaires telles que Siriwardhana et al. (2020).

Cependant, les représentations acoustiques-textuelles conjointes reposent souvent sur des transcriptions humaines des signaux acoustiques, qui ne sont pas toujours disponibles dans une application réaliste des modèles de la RAE. Par conséquent, plusieurs études ont examiné l'utilisation des transcriptions ASR pour fournir des représentations acoustiques-textuelles conjointes pour la RAE sur les expressions émotionnelles actées (Heusser et al., 2019; Yoon et al., 2019; Wu et al., 2021; Peng et al., 2021). Ces études ont montré que même si les représentations acoustiques-textuelles conjointes basées sur les transcriptions RAP ne sont pas aussi efficaces que les transcriptions humaines pour la RAE, elles sont toujours plus performantes que les représentations acoustiques seules. Afin de faire progresser les études de pointe susmentionnées, cette méthode est évaluée dans cette thèse pour les expressions émotionnelles naturelles, en exploitant l'ensemble de données CMU-MOSEI. Les résultats montrent que cette méthode est également efficace pour les expressions émotionnelles naturelles, mais principalement grâce à l'utilisation de représentations textuelles, car les représentations acoustiques, qu'il s'agisse de W2V2-XLSR-56 ou de LLD traditionnels, ne se sont pas révélées suffisamment robustes pour être utilisées de manière fiable pour les expressions émotionnelles naturelles.

L'utilisation de messages verbaux provenant de représentations textuelles n'est pas la seule source d'information susceptible d'améliorer les performances de la RAE à partir de représentations acoustiques. L'utilisation des représentations du locuteur s'est également avérée efficace pour la RAE, car elle fournit des informations sur le style des différents locuteurs. Par conséquent, dans cette thèse, plusieurs expériences ont été réalisées sur l'ensemble de données IEMOCAP afin d'étudier l'efficacité de la fusion des représentations du locuteur dans les représentations acoustiques-textuelles conjointes sur la RAE, en utilisant les représentations W2V2-XLSR-56 et RoBERTa. Les résultats ont montré que les représentations acoustiques-textuelles tenant compte du locuteur peuvent atteindre de meilleures performances dans la RAE que les représentations acoustiques-textuelles sans représentation du locuteur. Cependant, il a également été démontré que l'utilisation de transcriptions RAP pour calculer des représentations acoustiques-textuelles tenant compte du locuteur ne conduit pas à des résultats significativement meilleurs que l'utilisation de représentations acoustiques tenant compte du locuteur, et donc

la précision du modèle RAP est un facteur limitant dans ce paradigme. En plus, il a été démontré que les représentations acoustiques profondes telles que Whisper, qui ont été pré-entraînées directement pour la RAP et indirectement pour la reconnaissance du locuteur, peuvent contenir à la fois des informations verbales et des informations sur le locuteur et n'ont pas besoin d'être fusionnées avec des représentations textuelles et des représentations du locuteur pour atteindre des performances comparables aux représentations conjointes W2V2-RoBERTa tenant compte du locuteur pour la RAE actée.

D.5 Généralisation au-delà des schémas d'annotation des émotions

Les méthodes actuelles de la RAE utilisent des modèles d'apprentissage automatique pilotés par les données pour mettre en correspondance les représentations acoustiques ou textuelles des données avec des représentations numériques de l'émotion. Cependant, l'émotion étant un concept subjectif pour lequel il n'existe pas de définition commune, la représentation de l'émotion varie entre les différents ensembles de données. D'autre part, il est important d'entraîner les modèles de la RAE sur plusieurs ensembles de données, car chaque ensemble de données représente une gamme limitée d'expressions émotionnelles qui peuvent être observées naturellement. Afin d'utiliser plusieurs ensembles de données avec différents schémas d'annotation des émotions, cette thèse propose une méthode basée sur MTL qui prédit une intégration d'émotions généralisée, qui peut ensuite être mise en correspondance avec différents ensembles d'émotions basés sur chaque ensemble de données. Pour calculer l'intégration des émotions, la méthode proposée utilise un modèle GRU partagé, qui est entraîné avec des classificateurs linéaires, qui sont chacun dédiés aux énoncés de chaque ensemble de données concerné.

Cette méthode a d'abord été évaluée sur quatre petits ensembles de données actées –CaFE, EmoDB, GEMEP, et RAVDESS– dans des paramètres intra-corpus, et inter-corpus pour la RAE à partir de représentations acoustiques profondes –W2V2-XLSR-56–. Les résultats à l'intérieur du corpus ont montré que cette méthode est particulièrement efficace pour améliorer les performances de la RAE pour les petits ensembles de données. D'autre part, les résultats inter-corpus suggèrent que cette méthode peut calculer efficacement les représentations d'émotions qui peuvent se généraliser au-delà des différentes étiquettes d'émotions dans différents corpus.

La méthode proposée, basée sur le MTL, a ensuite été évaluée pour les expressions émotionnelles in-the-wild et également avec des représentations textuelles profondes –RoBERTa–, en plus des représentations acoustiques profondes –W2V2-XLSR-56–. A la place des quatre ensembles de données actées, les ensembles de données CMU-MOSEI (naturelles) et IEMOCAP (actées) ont ensuite été utilisés

pour étudier si les expressions émotionnelles actées, lorsqu'elles sont utilisées dans un paradigme MTL, sont utiles pour la RAE naturelle. Les résultats ont montré qu'en utilisant des représentations acoustiques profondes, aucune amélioration n'est observée pour l'ensemble de données CMU-MOSEI lorsqu'il est entraîné dans un contexte multicorpus avec les données IEMOCAP par rapport à l'entraînement avec un seul corpus. Cela a été attribué au fait que les représentations W2V2-XLSR-56 n'étaient pas suffisamment robustes pour se généraliser à travers différents locuteurs, environnements ou microphones. En revanche, l'utilisation de représentations textuelles profondes de l'ensemble de données IEMOCAP a permis d'améliorer de manière significative les performances de la RAE pour l'ensemble de données CMU-MOSEI. Ce résultat suggère que, bien que les représentations acoustiques profondes ne soient pas encore suffisamment robustes pour être utilisées pour les expressions émotionnelles naturelles, les transcriptions des signaux vocaux peuvent bénéficier de manière significative de la méthode de formation à plusieurs corpus proposée pour fournir une meilleure RAE naturelle qu'en utilisant les signaux vocaux seuls.

D.6 Conclusion

Dans cette thèse, plusieurs études ont été menées pour faire avancer la recherche actuelle sur l'apprentissage automatique des expressions émotionnelles naturelles à partir de signaux acoustiques et de textes. Plus spécifiquement, deux objectifs principaux ont été étudiés : 1) l'utilisation de représentations profondes pour la RAE et 2) la généralisation des modèles de la RAE au-delà des schémas d'annotations des émotions. Ces études ont donné lieu à cinq contributions principales qui sont énumérées ci-dessous :

1. Plusieurs modèles Wav2vec2 ont d'abord été entraînés en utilisant différents types et quantités de parole française. Les modèles pré-entraînés ont ensuite été utilisés comme représentations acoustiques pour la RAE et évalués sur deux jeux de données français, AlloSat (téléphonique, naturel) et RECOLA (spontané, en laboratoire). Les résultats ont montré que l'utilisation d'une plus grande quantité de données (après 3k heures) pour entraîner les représentations profondes ne conduit pas nécessairement à une amélioration de la performance de ces représentations pour la RAE. Cependant, la langue des données utilisées pour entraîner les modèles, indépendamment du type de discours, semble jouer un rôle important, puisque les représentations profondes pré-entraînées sur le discours français étaient plus performantes pour prédire les expressions émotionnelles françaises que les représentations profondes pré-entraînées sur le discours anglais.

2. Les représentations profondes acoustiques (W2V2-XLSR-56) et textuelles (RoBERTa) des transcriptions humaines et automatiques ont été utilisées de manière isolée et conjointe pour la RAE. Les résultats ont montré que les représentations acoustiques et textuelles conjointes offraient de meilleures performances que les représentations acoustiques ou textuelles seules pour les expressions émotionnelles jouées et naturelles, que le texte soit humain ou transcrit automatiquement. Cette meilleure performance pour les expressions émotionnelles naturelles est principalement due à l'utilisation de représentations textuelles, les représentations acoustiques s'étant révélées inefficaces pour la RAE naturel.

3. Les modèles de reconnaissance du locuteur ont été entraînés pour obtenir des vecteurs latents du locuteur. Ensuite, les vecteurs de locuteurs ont été concaténés avec des représentations latentes acoustiques-textuelles pour former des représentations conscientes du locuteur pour la RAE sur l'ensemble de données IEMOCAP. Les résultats ont montré que la fusion des vecteurs de locuteurs avec les représentations acoustiques-textuelles conjointes –W2V2-RoBERTa– donne les meilleurs résultats parmi les scénarios testés, bien qu'elle soit marginalement meilleure que les représentations acoustiques et textuelles conjointes sans les représentations du locuteur. Il a également été démontré qu'en utilisant des représentations profondes telles que Whisper, qui ont été pré-entraînées directement pour RAP et indirectement pour la reconnaissance du locuteur, il n'est peut-être pas nécessaire de fusionner davantage les représentations textuelles et les représentations du locuteur, pour obtenir des performances comparables aux représentations conjointes W2V2-RoBERTa tenant compte du locuteur pour la RAE actée.

4. Une méthode basée sur MTL a été proposée qui utilise des représentations acoustiques profondes, un modèle GRU partagé pour tous les ensembles de données, et des classificateurs linéaires dédiés à chaque ensemble de données. La méthode proposée a ensuite été évaluée à l'aide de quatre ensembles de données actées provenant de CaFE, EmoDB, GEMEP et RAVDESS, qui utilisent tous différents ensembles d'étiquettes d'émotions. Les résultats suggèrent que la méthode proposée peut produire une représentation latente des émotions qui peut améliorer les performances de RAE dans les ensembles de données avec différents ensembles d'étiquettes d'émotions.

5. La méthode basée sur la MTL expliquée ci-dessus a été évaluée plus avant, sur les ensembles de données IEMOCAP (actées) et CMU-MOSEI (naturelles), et pour les représentations acoustiques profondes –W2V2-XLSR-56– et textuelles –RoBERTa–. Les résultats suggèrent qu'en utilisant la méthode proposée basée sur MTL, la performance RAE de l'ensemble de

données CMU-MOSEI peut être améliorée en utilisant des représentations textuelles ou des représentations acoustiques-textuelles conjointes, mais pas avec des représentations acoustiques seules. Cela suggère que les représentations acoustiques profondes ne sont pas encore assez robustes pour la reconnaissance naturelle des émotions, alors que l'utilisation d'un RAP pour obtenir un aperçu des mots prononcés peut améliorer la performance RAE naturelle.

Bibliography

Abend, K. (2022). How convolutional neural networks defy the curse of dimensionality: Deep learning explained. *TechRxiv. Preprint: 18316439*. 25

Adoma, A. F., Henry, N.-M., and Chen, W. (2020). Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121, Chengdu, China. IEEE. 63

Aggarwal, C. C. and Zhai, C. (2012). A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer. 22

Akhtar, M. S., Chauhan, D. S., Ghosal, D., Poria, S., Ekbal, A., and Bhattacharyya, P. (2019). Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*. 47

AlBadawy, E. A. and Kim, Y. (2018). Joint discrete and continuous emotion prediction using ensemble and end-to-end approaches. In *Proceedings of the 20th International Conference on Multimodal Interaction (ICMI'18)*, pages 366–375, Boulder (CO), USA. ACM. 17, 51, 81

Alisamir, S. and Ringeval, F. (2021). On the evolution of speech representations for affective computing: A brief history and critical overview. *IEEE Signal Processing Magazine*, 38(6):12–21. 6, 7, 23, 38, 49

Alisamir, S., Ringeval, F., and Portet, F. (2022a). Cross-domain voice activity detection with self-supervised representations. *arXiv preprint arXiv:2209.11061*. 148

Alisamir, S., Ringeval, F., and Portet, F. (2022b). Dynamic time-alignment of dimensional annotations of emotion using recurrent neural networks. *arXiv preprint arXiv:2209.10223*. 18

Alisamir, S., Ringeval, F., and Portet, F. (2022c). Multi-corpus affect recognition with emotion embeddings and self-supervised representations of speech. In

2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1–8, Nara, Japan. IEEE. 26, 112

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., and Schuller, B. (2017). Snore sound classification using image-based deep spectrum features. In *Proc. Interspeech 2017*, pages 3512–3516, Stockholm, Sweden. ISCA. 32

Assunção, G., Menezes, P., and Perdigão, F. (2020). Speaker awareness for speech emotion recognition. *Int. J. Online Biomed. Eng.*, 16(4):15–22. 46, 98

Atmaja, B. T., Sasou, A., and Akagi, M. (2022). Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication*, 140:11–28. 7, 49, 88

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*. 62

Baevski, A., Schneider, S., and Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. *arXiv preprint arXiv:1910.05453*. 39

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*. 38, 39, 40, 62, 75

Bänziger, T., Mortillaro, M., and Scherer, K. R. (2012). Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion*, 12(5):1161–1179. 16, 59

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Pan Macmillan. 14

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828. 6, 23

Bertero, D. and Fung, P. (2017). A first look into a convolutional neural network for speech emotion detection. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5115–5119, New Orleans, Louisiana, USA. IEEE. 31

Brans, K. and Verduyn, P. (2014). Intensity and duration of negative emotions: Comparing the role of appraisals and regulation strategies. *PLoS One*, 9(3):e92410. 21

Braunschweiler, N., Doddipatla, R., Keizer, S., and Stoyanchev, S. (2021). A study on cross-corpus speech emotion recognition and data augmentation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 24–30, Cartagena, Colombia. IEEE. 48

Brosch, T., Scherer, K. R., Grandjean, D. M., and Sander, D. (2013). The impact of emotion on perception, attention, memory, and decision-making. *Swiss medical weekly*, 143:w13786. 1

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Proc. Interspeech 2005*, volume 5, pages 1517–1520, Lisbon, Portugal. 58

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359. 59

Cahyani, D. E. and Patasik, I. (2021). Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, 10(5):2780–2788. 42

Cai, X., Yuan, J., Zheng, R., Huang, L., and Church, K. (2021). Speech emotion recognition with multi-task learning. In *Proc. Interspeech 2021*, pages 4508–4512, Brno, Czech Republic. 87

Caruana, R. (1998). *Multitask learning*. Springer. 47

Chao, L., Tao, J., Yang, M., Li, Y., and Wen, Z. (2015). Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72. 26

Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A., and Narayanan, S. (2019). Data augmentation using gans for speech emotion recognition. In *Proc. Interspeech 2019*, pages 171–175, Graz, Austria. 37

Chaudhari, S., Mithal, V., Polatkan, G., and Ramanath, R. (2021). An attentive survey of attention models. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–32. 29

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1724–1734, Doha, Qatar. Association for Computational Linguistics. 26

Chung, Y.-A. and Glass, J. (2018). Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In *Proc. Interspeech 2018*, pages 811–815, Hyderabad, India. 39, 45

Chung, Y.-A. and Glass, J. (2020). Generative pre-training for speech with autoregressive predictive coding. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3497–3501, Barcelona, Spain. IEEE. 39, 40

Chung, Y.-A., Hsu, W.-N., Tang, H., and Glass, J. (2019). An Unsupervised Autoregressive Model for Speech Representation Learning. In *Proc. Interspeech 2019*, pages 146–150, Graz, Austria. 39, 40

Chung, Y.-A., Tang, H., and Glass, J. (2020). Vector-quantized autoregressive predictive coding. In *Proc. Interspeech 2020*, pages 3760–3764, Online, China. 39

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*, pages 2426–2430, Brno, Czech Republic. 62

Costanzi, M., Cianfanelli, B., Saraulli, D., Lasaponara, S., Doricchi, F., Cestari, V., and Rossi-Arnaud, C. (2019). The effect of emotional valence and arousal on visuo-spatial working memory: Incidental emotional learning and memory for object-location. *Frontiers in Psychology*, 10:2587. 17

da Silva, R., Valter Filho, M., and Souza, M. (2020). Interaffection of multiple datasets with neural networks in speech emotion recognition. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, pages 342–353. SBC. 5, 48

Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain. 1, 149

Deng, J., Xu, X., Zhang, Z., Frühholz, S., and Schuller, B. (2017). Semisupervised autoencoders for speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(1):31–43. 35

Deng, J., Zhang, Z., Eyben, F., and Schuller, B. (2014). Autoencoder-based unsupervised domain adaptation for speech emotion recognition. *IEEE Signal Processing Letters*, 21(9):1068–1072. 33

Denisov, P. and Vu, N. T. (2020). Pretrained semantic speech embeddings for end-to-end spoken language understanding via cross-modal teacher-student learning. In *Proc. Interspeech 2020*, pages 881–885, Online, China. 44

Deschamps-Berger, T., Lamel, L., and Devillers, L. (2021). End-to-end speech emotion recognition: challenges of real-life emergency call centers data recordings. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Cambridge, United Kingdom. IEEE. 31

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics. 43, 44, 63

Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., and Shanahan, M. (2016). Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*. 36

Dissanayake, V., Zhang, H., Billinghurst, M., and Nanayakkara, S. (2020). Speech emotion recognition "in the wild" using an autoencoder. *Proc. Interspeech 2020*, pages 526–530. 34

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200. 4, 14, 15, 54

Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712. 14

El Seknedy, M. and Fawzi, S. (2021). Speech emotion recognition system for human interaction applications. In *2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 361–368, Cairo, Egypt. IEEE. 118

Evain, S., Nguyen, H., Le, H., Boito, M. Z., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021a). Lebenchmark: A reproducible framework for assessing self-supervised representation learning from speech. In *Proc. Interspeech 2021*, pages 1439–1443, Brno, Czech Republic. 31, 38, 51, 62, 75

Evain, S., Nguyen, M. H., Le, H., Zanon Boito, M., Mdhaffar, S., Alisamir, S., Tong, Z., Tomashenko, N., Dinarelli, M., Parcollet, T., Allauzen, A., Estève, Y., Lecouteux, B., Portet, F., Rossato, S., Ringeval, F., Schwab, D., and Besacier, L. (2021b). Task agnostic and task specific self-supervised learning from speech with lebenchmark. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) 2021*, NeurIPS 2021 Datasets and Benchmarks Track, Online, USA. 7, 23, 26, 73, 75, 86

Evangelopoulos, N. E. (2013). Latent semantic analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 4(6):683–692. 23

Eyben, F. (2015). *Real-time speech and music classification by large audio feature space extraction*. Springer. 67, 148

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., et al. (2015). The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202. 21

Ezz-Eldin, M., Khalaf, A. A., Hamed, H. F., and Hussein, A. I. (2021). Efficient feature-aware hybrid model of deep learning architectures for speech emotion recognition. *IEEE Access*, 9:19999–20011. 23

Fridenson-Hayo, S., Berggren, S., Lassalle, A., Tal, S., Pigat, D., Meir-Goren, N., O'Reilly, H., Ben-Zur, S., Bölte, S., Baron-Cohen, S., et al. (2017). 'emotiplay': a serious game for learning about emotions in children with autism: results of a cross-cultural evaluation. *European child & adolescent psychiatry*, 26(8):979–992. 2

Gatopoulos, I. and Tomczak, J. M. (2021). Self-supervised variational auto-encoders. *Entropy*, 23(6):747. 38

Gendron, M. and Feldman Barrett, L. (2009). Reconstructing the past: A century of ideas about emotion in psychology. *Emotion review*, 1(4):316–339. 14

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press. 23, 26, 65

Gosztolya, G. (2020). Using the fisher vector representation for audio-based emotion recognition. *Acta Polytechnica Hungarica*, 17(6):7–23. 21

Goudbeek, M. and Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128(3):1322–1336. 92, 132

Gournay, P., Lahaie, O., and Lefebvre, R. (2018). A canadian french emotional speech dataset. In *Proceedings of the 9th ACM multimedia systems conference*, pages 399–402, New York, New York, USA. 4, 16, 58

Gross, R. (2020). *Psychology: The science of mind and behaviour 8th edition*. Hodder Education. 15

He, L., Jiang, D., Yang, L., Pei, E., Wu, P., and Sahli, H. (2015). Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 73–80. 26

Hermans, J. R., Spanakis, G., and Möckel, R. (2017). Accumulated gradient normalization. In *Asian Conference on Machine Learning*, pages 439–454, Seoul, Korea. PMLR. 64

Hernández-García, A. and König, P. (2018). Data augmentation instead of explicit regularization. *arXiv preprint arXiv:1806.03852*. 37

Heusser, V., Freymuth, N., Constantin, S., and Waibel, A. (2019). Bimodal speech emotion recognition using pre-trained language models. *arXiv preprint arXiv:1912.02610*. 7, 45, 74, 95, 108, 131, 154

Ho, N.-H., Yang, H.-J., Kim, S.-H., and Lee, G. (2020). Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network. *IEEE Access*, 8:61672–61686. 7, 87, 90, 91

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780. 26

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, volume 4, pages 9–56, Christchurch, New Zealand. 22

Huang, Z., Dong, M., Mao, Q., and Zhan, Y. (2014). Speech emotion recognition using cnn. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 801–804, Faro, Portugal. 34

Huang, Z. and Epps, J. (2017). A pllr and multi-stage staircase regression framework for speech-based emotion prediction. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5145–5149, New Orleans, Louisiana, USA. IEEE. 45

Huang, Z. and Epps, J. (2020). An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech. *IEEE Transactions on Affective Computing*, 11(4):653–668. 45

Ibrahim, H., Loo, C. K., and Alnajjar, F. (2021). Speech emotion recognition by late fusion for bidirectional reservoir computing with random projection. *IEEE Access*, 9:122855–122871. 118

Jiang, D., Lei, X., Li, W., Luo, N., Hu, Y., Zou, W., and Li, X. (2019). Improving transformer-based speech recognition using unsupervised pre-training. *arXiv preprint arXiv:1910.09932*. 39

Jianqiang, Z., Xiaolin, G., and Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. *IEEE access*, 6:23253–23260. 23

Juremi, N. R. M., Zulkifley, M. A., Hussain, A., and Zaki, W. M. D. W. (2017). Inter-rater reliability of actual tagged emotion categories validation using cohen's kappa coefficient. *Journal of Theoretical and Applied Information Technology*, 95(2):259–264. 16

Kensinger, E. A. and Schacter, D. L. (2006). Processing emotional pictures and words: Effects of valence and arousal. *Cognitive, Affective, & Behavioral Neuroscience*, 6(2):110–126. 17

Khandelwal, S., Lecouteux, B., and Besacier, L. (2016). *Comparing GRU and LSTM for automatic speech recognition*. PhD thesis, LIG. 52

Khorram, S., McInnis, M. G., and Provost, E. M. (2021). Jointly aligning and predicting continuous emotion annotations. *IEEE Transactions on Affective Computing*, 12(4):1069–1083. 17, 18

Kim, E., Song, H., and Shin, J. W. (2020a). Affective latent representation of acoustic and lexical features for emotion recognition. *Sensors*, 20(9):2614. 45

Kim, J. Y., Liu, C., Calvo, R. A., McCabe, K., Taylor, S. C., Schuller, B. W., and Wu, K. (2019). A comparison of online automatic speech recognition systems and the nonverbal responses to unintelligible speech. *arXiv preprint arXiv:1904.12403*. 7, 45

Kim, M., Tack, J., and Hwang, S. J. (2020b). Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:2983–2994. 38

Kim, N. K., Lee, J., Ha, H. K., Lee, G. W., Lee, J. H., and Kim, H. K. (2017). Speech emotion recognition based on multi-task learning using a convolutional

neural network. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 704–707, Kuala Lumpur, Malaysia. IEEE. 47, 111

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR (Poster)*. 64

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *International Conference on Neural Information Processing Systems*, volume 2 of *NIPS'14*, page 3581–3589. MIT Press. 36

Kollias, D. and Zafeiriou, S. (2018). Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*. 126

Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B., Star, K., Hajiyev, E., and Pantic, M. (2021). SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1022–1040. 2, 17, 56, 127, 149

Krishna, D. and Patil, A. (2020). Multimodal emotion recognition using cross-modal attention and 1d convolutional neural networks. In *Proc. Interspeech 2020*, pages 4243–4247, Online, China. 43

Kuppens, P., Tuerlinckx, F., Russell, J. A., and Barrett, L. F. (2013). The relation between valence and arousal in subjective experience. *Psychological bulletin*, 139(4):917. 4, 18

Kuppens, P., Tuerlinckx, F., Yik, M., Koval, P., Coosemans, J., Zeng, K. J., and Russell, J. A. (2017). The relation between valence and arousal in subjective experience varies with personality and culture. *Journal of personality*, 85(4):530–542. 4, 18

Lang, P. J., Davis, M., and Öhman, A. (2000). Fear and anxiety: animal models and human cognitive psychophysiology. *Journal of affective disorders*, 61(3):137–159. 119

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., and Schuller, B. W. (2020). Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378*. 6, 21, 23, 31, 73

Latif, S., Rana, R., Qadir, J., and Epps, J. (2018). Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study. In *Proc. Interspeech 2018*, pages 3107–3111, Hyderabad, India. 37

Le, D., Aldeneh, Z., and Provost, E. M. (2017). Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. In *Proceedings INTERSPEECH 2017, 18th Annual Conference of the International Speech Communication Association*, pages 1108–1112, Stockholm, Sweden. 26

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. 6, 23

Lee, C.-C., Sridhar, K., Li, J.-L., Lin, W.-C., Su, B.-H., and Busso, C. (2021). Deep representation learning for affective speech signal analysis and processing: Preventing unwanted signal disparities. *IEEE Signal Processing Magazine*, 38(6):22–38. 8

Lee, S.-w. (2019). The generalization effect for multilingual speech emotion recognition across heterogeneous languages. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5881–5885, Brighton, United Kingdom. IEEE. 48

Lewis, P. A., Critchley, H., Rotshtein, P., and Dolan, R. J. (2006). Neural correlates of processing valence and arousal in affective words. *Cerebral cortex*, 17(3):742–748. 17

Li, J.-L. and Lee, C.-C. (2019). Attentive to individual: A multimodal emotion recognition network with personalized attention profile. In *Proc. Interspeech 2019*, pages 211–215, Graz, Austria. 7, 83, 87, 90, 91

Li, L. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268. 65

Liang, H., Sun, X., Sun, Y., and Gao, Y. (2017). Text feature extraction based on deep learning: a review. *European Association for Signal Processing (EURASIP) journal on wireless communications and networking*, 2017(1):1–12. 22

Lieskovská, E., Jakubec, M., Jarina, R., and Chmulík, M. (2021). A review on speech emotion recognition using deep learning and attention mechanism. *Electronics*, 10(10):1163. 90

Lilleberg, J., Zhu, Y., and Zhang, Y. (2015). Support vector machines and word2vec for text classification with semantic features. In *2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC)*, pages 136–140, Beijing, China. IEEE. 42

Lindquist, K. A., MacCormack, J. K., and Shablack, H. (2015). The role of language in emotion: Predictions from psychological constructionism. *Frontiers in psychology*, 6:444. 81, 123

Ling, S., Liu, Y., Salazar, J., and Kirchhoff, K. (2020). Deep contextualized acoustic representations for semi-supervised speech recognition. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6429–6433, Barcelona, Spain. IEEE. 39

Liu, A. T., Li, S.-W., and Lee, H.-y. (2021). Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 29:2351–2366. 7

Liu, A. T., Yang, S.-w., Chi, P.-H., Hsu, P.-c., and Lee, H.-y. (2020). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423, Barcelona, Spain. IEEE. 39

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., and Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26. 23

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. 44, 63, 86

Livingstone, S. R. and Russo, F. A. (2018). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391. 59

Low, D. M., Bentley, K. H., and Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1):96–116. 2

Macary, M., Tahon, M., Estève, Y., and Rousseau, A. (2020). AlloSat: A new call center French corpus for satisfaction and frustration analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1590–1597, Marseille, France. European Language Resources Association. 2, 57, 76

Macary, M., Tahon, M., Estève, Y., and Rousseau, A. (2021). On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 373–380, Shenzhen, China. IEEE. 45, 81

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861. 119

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. 42

Moine, C. L., Obin, N., and Roebel, A. (2021). Speaker Attentive Speech Emotion Recognition. In *Proc. Interspeech 2021*, pages 2866–2870, Brno, Czech Republic. 8, 46, 100

Monkaresi, H., Bosch, N., Calvo, R. A., and D'Mello, S. K. (2016). Automated detection of engagement using video-based estimation of facial expressions and heart rate. *IEEE Transactions on Affective Computing*, 8(1):15–28. 2

Nandan, A. and Vepa, J. (2020). Language agnostic speech embeddings for emotion classification. *Workshop on Self-supervision in Audio and Speech at the 37th International Conference on Machine Learning*. 7

Naseem, U., Razzak, I., Khan, S. K., and Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35. 6

Neumann, M. et al. (2018). Cross-lingual and multilingual speech emotion recognition on english and french. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5769–5773, Calgary, AB, Canada. IEEE. 48

Neumann, M. and Vu, N. T. (2019). Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7390–7394, Brighton, United Kingdom. IEEE. 34

Nicolaou, M. A., Gunes, H., and Pantic, M. (2011). Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2(2):92–105. 18

Nicolaou, M. A., Gunes, H., and Pantic, M. (2012). Output-associative rvm regression for dimensional and continuous emotion prediction. *Image and Vision Computing*, 30(3):186–196. 18

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*. 38, 39

Palaz, D., Collobert, R., et al. (2015). Analysis of cnn-based speech recognition system using raw speech as input. Technical report, Idiap. 25, 31

Pandit, V. and Schuller, B. (2019). The many-to-many mapping between the concordance correlation coefficient and the mean square error. *arXiv preprint arXiv:1902.05180*. 65

Paninski, L. (2003). Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253. 22

Pantic, M., Pentland, A., Nijholt, A., and Huang, T. S. (2007). Human computing and machine understanding of human behavior: A survey. In *Artifical Intelligence for Human Computing*, pages 47–71. Springer. 18

Pappagari, R., Wang, T., Villalba, J., Chen, N., and Dehak, N. (2020). x-vectors meet emotions: A study on dependencies between emotion and speaker recognition. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7169–7173, Barcelona, Spain. IEEE. 46, 74, 98, 109

Parthasarathy, S. and Busso, C. (2017). Jointly predicting arousal, valence and dominance with multi-task learning. In *Proc. Interspeech 2017*, pages 1103–1107, Stockholm, Sweden. 47

Parthasarathy, S. and Busso, C. (2018). Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes. *arXiv preprint arXiv:1804.10816*. 35

Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., and Bengio, Y. (2019). Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*. 39

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Lin, Z., Gimelshein, N., and Antiga, L. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the thirty-third Conference on Neural Information Processing Systems (NIPS)*, pages 8026–8037, Vancouver, BC, Canada. Neural Information Processing Systems Foundation. 70

Peng, Z., Lu, Y., Pan, S., and Liu, Y. (2021). Efficient speech emotion recognition using multi-scale cnn and attention. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3020–3024, Online, Canada. IEEE. 8, 45, 74, 95, 108, 131, 154

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. 42

Peri, R., Parthasarathy, S., Bradshaw, C., and Sundaram, S. (2021). Disentanglement for audio-visual emotion recognition using multitask setup. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348, Online, Canada. IEEE. 8, 46

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, volume 1, pages 2227–2237, New Orleans, Louisiana, USA. 43

Picard, R. W. (2003). Affective computing: challenges. *International Journal of Human-Computer Studies*, 59(1-2):55–64. 16

Povolny, F., Matejka, P., Hradis, M., Popková, A., Otrusina, L., Smrz, P., Wood, I., Robin, C., and Lamel, L. (2016). Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 75–82. 42

Prabhu, N. R., Carbajal, G., Lehmann-Willenbrock, N., and Gerkmann, T. (2022). End-To-End Label Uncertainty Modeling for Speech-based Arousal Recognition Using Bayesian Neural Networks. In *Proc. Interspeech 2022*, pages 151–155, Incheon, Korea. 51

Quitry, F. d. C., Tagliasacchi, M., and Roblek, D. (2019). Learning audio representations via phase prediction. *arXiv preprint arXiv:1910.11910*. 39

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. Technical report, Tech. Rep., OpenAI. 60, 103

Rana, R. (2016). Gated recurrent unit (gru) for emotion classification from noisy speech. *arXiv preprint arXiv:1612.07778*. 52

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. arXiv:2106.04624. 70

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., Ciftci, E., Gülec, H., Salah, A. A., and Pantic, M. (2018a). Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 8th*

International Workshop on Audio/Visual Emotion Challenge (AVEC'18), pages 3–13. ACM. 21, 23, 60

Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., et al. (2018b). Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pages 3–13. ACM. 31, 32

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., et al. (2019). Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12. 17, 32, 48

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE. 5, 17, 18, 50, 60, 76

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*. 24

Rudovic, O., Lee, J., Dai, M., Schuller, B., and Picard, R. W. (2018). Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19):eaao6760. 8, 46

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161. 4, 15, 54

Sahu, S., Gupta, R., Sivaraman, G., AbdAlmageed, W., and Espy-Wilson, C. (2018). Adversarial auto-encoders for speech based emotion recognition. *arXiv preprint arXiv:1806.02146*. 37

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2234–2242, Barcelona, Spain. Curran Associates Inc. 38

Sander, D., Grandjean, D., and Scherer, K. R. (2005). A systems approach to appraisal mechanisms in emotion. *Neural networks*, 18(4):317–352. 8, 14

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social science information*, 44(4):695–729. 3

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and emotion*, 23(7):1307–1351. 4, 8, 15

Schmitt, M., Ringeval, F., and Schuller, B. (2016). At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *Proc. Interspeech 2016*, pages 495–499, San Francisco, California, USA. ISCA. 21, 23

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469, Graz, Austria. 39

Schuller, B. W., Batliner, A., Bergler, C., Messner, E.-M., Hamilton, A., Amiriparian, S., Baird, A., Rizos, G., Schmitt, M., Stappen, L., Baumeister, H., MacIntyre, A. D., and Hantke, S. (2020). The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing Masks. In *Proc. Interspeech 2020*, pages 2042–2046, Online, China. 132

Sharma, G. and Dhall, A. (2021). A survey on automatic multimodal emotion recognition in the wild. In *Advances in Data Science: Methodologies and Applications*, pages 35–64. Springer. 49

Siirtola, P., Tamminen, S., Chandra, G., Ihalapathirana, A., and Röning, J. (2023). Predicting emotion with biosignals: A comparison of classification and regression models for estimating valence and arousal level using wearable sensors. *Sensors*, 23(3):1598. 16

Singh, V., Kumar, B., and Patnaik, T. (2013). Feature extraction techniques for handwritten text in various scripts: a survey. *International Journal of Soft Computing and Engineering (IJSCE)*, 3(1):238–241. 22

Siriwardhana, S., Reis, A., Weerasekera, R., and Nanayakkara, S. (2020). Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. In *Proc. Interspeech 2020*, pages 3755–3759, Online, China. 7, 23, 44, 45, 59, 63, 83, 84, 90, 91, 92, 93, 108, 154

Song, X., Wang, G., Huang, Y., Wu, Z., Su, D., and Meng, H. (2020). Speech-XLNet: Unsupervised Acoustic Model Pretraining for Self-Attention Networks. In *Proc. Interspeech 2020*, pages 3765–3769, Online, China. 39

Suleman, R. M. and Korkontzelos, I. (2021). Extending latent semantic analysis to manage its syntactic blindness. *Expert Systems with Applications*, 165:114130. 41

Sun, C., Huang, L., and Qiu, X. (2019). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 380–385, Minneapolis, Minnesota, USA. 44

Sun, Z., Sarma, P., Sethares, W., and Liang, Y. (2020). Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8992–8999, New York, New York, USA. 92, 93, 94

Ta, B. T., Nguyen, T. L., Dang, D. S., Le, N. M., et al. (2022). Improving speech emotion recognition via fine-tuning asr with speaker information. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6, Chiang Mai, Thailand. IEEE. 8, 46, 74, 98, 100

Tagliasacchi, M., Gfeller, B., Quitry, F. d. C., and Roblek, D. (2019). Self-supervised audio representation learning for mobile devices. *arXiv preprint arXiv:1905.11796*. 31, 39

Tahon, M., Macary, M., Luzzati, D., et al. (2021). Mutual impact of acoustic and linguistic representations for continuous emotion recognition in call-center conversations. *techrxiv.17104526.v1*. 81

Tarpin-Bernard, F., Fruitet, J., Vigne, J.-P., Constant, P., Chainay, H., Koenig, O., Ringeval, F., Bouchot, B., Bailly, G., Portet, F., Alisamir, S., Zhou, Y., et al. (2021). Theradia: Digital therapies augmented by artificial intelligence. In *Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2021 Virtual Conferences on Neuroergonomics and Cognitive Engineering, Industrial Cognitive Ergonomics and Engineering Psychology, and Cognitive Computing and Internet of Things*, pages 478–485, New York, New York, USA. Springer Nature. 135

Tits, N., El Haddad, K., and Dutoit, T. (2018). ASR-based features for emotion recognition: A transfer learning approach. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 48–52, Melbourne, Australia. Association for Computational Linguistics. 32

Triantafyllopoulos, A., Reichel, U., Liu, S., Huber, S., Eyben, F., and Schuller, B. W. (2023). Multistage linguistic conditioning of convolutional layers for speech emotion recognition. *Frontiers in Computer Science*, 5:1072479. 92, 132

Triantafyllopoulos, A. and Schuller, B. W. (2021). The role of task and acoustic similarity in audio transfer learning: Insights from the speech emotion recognition case. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7268–7272, Online, Canada. IEEE. 32

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204, Shanghai, China. IEEE. 6, 23, 25, 26, 31

Valstar, M., Gratch, J., Schuller, B., Ringeval, F., Lalanne, D., Torres Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). Avec 2016 – depression, mood, and emotion recognition workshop and challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge (AVEC'16)*, pages 3–10, Amsterdam, The Netherlands. ACM. 60

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017a). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, Long Beach, California, USA. 28, 29

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017b). Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*, pages 5998–6008, Long Beach, California, USA. 29, 43

Vazquez-Rodriguez, J. (2021). Using multimodal transformers in affective computing. In *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–5, Nara, Japan. IEEE. 29

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., and Schuller, B. W. (2022). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, (01):1–13. 29

Wang, W., Tang, Q., and Livescu, K. (2020). Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6889–6893, Barcelona, Spain. IEEE. 39

Wang, X. and Zheng, Q. (2013). Text emotion classification research based on improved latent semantic analysis algorithm. In *Conference of the 2nd*

International Conference on Computer Science and Electronics Engineering (ICCSEE 2013), pages 210–213, Hangzhou, China. Atlantis Press. 23

Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., Gao, S., Sun, Y., Ge, W., Zhang, W., and Zhang, W. (2022). A systematic review on affective computing: emotion models, databases, and recent advances. *Information Fusion*, 83-84:19–52. 49

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., and Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE Access*, 9:47795–47814. 49

Weninger, F., Ringeval, F., Marchi, E., and Schuller, B. (2016). Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2196–2202, New York, New York, USA. AAAI Press. 26

Wu, W., Zhang, C., and Woodland, P. C. (2021). Emotion recognition by fusing time synchronous and time asynchronous representations. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6269–6273, Online, Canada. IEEE. 8, 45, 74, 95, 108, 131, 154

Xi, Y., Li, P., Song, Y., Jiang, Y., and Dai, L. (2019). Speaker to emotion: Domain adaptation for speech emotion recognition with residual adapters. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 513–518, Lanzhou, China. IEEE. 8

Xia, R. and Liu, Y. (2015). A multi-task learning framework for emotion recognition using 2d continuous space. *IEEE Transactions on affective computing*, 8(1):3–14. 47, 111

Xu, H., Zhang, H., Han, K., Wang, Y., Peng, Y., and Li, X. (2019). Learning Alignment for Multimodal Emotion Recognition from Speech. In *Proc. Interspeech 2019*, pages 3569–3573, Graz, Austria. 43

Xu, X., Deng, J., Coutinho, E., Wu, C., Zhao, L., and Schuller, B. W. (2018). Connecting subspace learning and extreme learning machine in speech emotion recognition. *IEEE Transactions on Multimedia*, 21(3):795–808. 59, 118

Yang, K., Lee, D., Whang, T., Lee, S., and Lim, H. (2019). Emotionx-ku: Bert-max based contextual emotion classifier. *arXiv preprint arXiv:1906.11565*. 44

Yang, S., Yu, X., and Zhou, Y. (2020). Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International workshop on electronic communication and artificial intelligence (IWECAI)*, pages 98–101, Shanghai, China. IEEE. 52

Yoon, S., Byun, S., Dey, S., and Jung, K. (2019). Speech emotion recognition using multi-hop attention mechanism. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826, Brighton, United Kingdom. IEEE. 7, 45, 74, 95, 108, 131, 154

Zadeh, A. B., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. (2018). Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2236–2246, Melbourne, Australia. 58, 96

Zhang, C. and Xue, L. (2021). Autoencoder with emotion embedding for speech emotion recognition. *IEEE Access*, 9:51231–51241. 83

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. (2020a). Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems (NeurIPS)*, 33:15383–15393. 64

Zhang, R., Wu, H., Li, W., Jiang, D., Zou, W., and Li, X. (2020b). Transformer based unsupervised pre-training for acoustic representation learning. *arXiv preprint arXiv:2007.14602*. 34

Zhang, R., Wu, H., Li, W., Jiang, D., Zou, W., and Li, X. (2021). Transformer based unsupervised pre-training for acoustic representation learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6933–6937, Toronto, Ontario, Canada. IEEE. 40

Zhang, W., Fang, Y., and Ma, Z. (2017a). The effect of task similarity on deep transfer learning. In *International Conference on Neural Information Processing Systems*, pages 256–265, Long Beach, California, USA. Springer. 32

Zhang, Y., Liu, Y., Weninger, F., and Schuller, B. (2017b). Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4990–4994, New Orleans, Louisiana, USA. IEEE. 5, 47, 48, 111, 117

Zhang, Y., Weninger, F., Schuller, B., and Picard, R. W. (2022). Holistic affect recognition using panda: Paralinguistic non-metric dimensional analysis. *IEEE Transactions on Affective Computing*, 13(2):769–780. 5, 48, 117

Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., and Schuller, B. (2016). Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks. In *Proc. Interspeech 2016*, pages 3593–3597, San Francisco, California, USA. 81

Zhao, S., Jia, G., Yang, J., Ding, G., and Keutzer, K. (2021). Emotion recognition from multiple modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine*, 38(6):59–73. 49

Zhou, P., Feng, J., Ma, C., Xiong, C., Hoi, S. C. H., et al. (2020). Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21285–21296. 64

Zhu, Z. and Sato, Y. (2020). Reconciliation of multiple corpora for speech emotion recognition by multiple classifiers with an adversarial corpus discriminator. In *Proc. Interspeech 2020*, pages 2342–2346, Online, China. 5, 48, 111, 117