Imperial College London
Department of Computing

End-to-End Multimodal Continuous Affect Recognition

Panagiotis Tzirakis

February 2022

Supervised by Prof. Björn Schuller & Prof. Stefanos Zafeiriou

Submitted in part fulfilment of the requirements for the degree of PhD in Computing and the Diploma of Imperial College London. This thesis is entirely my own work, and, except where otherwise indicated, describes my own research.

License

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC).

Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Statement of Originality

This thesis is my own work, and, except indicated appropriately by references, describes my own research.

Acknowledgements

First and foremost I would like to express my genuine gratitude to my supervisors Prof. Bjorn Schuller and Prof. Stefanos Zafeiriou for their support and guidance throughout the duration of my PhD. I would also like to thank Dr. Georgios Trigeorgis for our close collaboration in the early stages of my Ph.D. My thanks goes to Dr. Athanasios Papaioanou for his support and help at the late stages of my PhD. Moreover, I would like to thank my friend Stylianos Ploumpis for the great moments we had all these years both in UK and US. My sincere gratitude goes also to the The Engineering and Physical Sciences Research Council (EPSRC) for providing the funding for my thesis.

Finally, I would like to express my utmost gratefulness to my parents, Emmanouil and Christina, and my brother, Frangkiskos, for their continuous and unconditional support. This thesis is dedicated to them.

Abstract

Automatic continuous affect recognition is a vital component towards a complete and natural interaction between human and machine. However, the task is challenging as human emotions lack temporal boundaries and are expressed through various modalities and in various ways among different individuals and cultures. On top of that, the perception of emotion is highly subjective which makes the task even more challenging. The recent years, several advancements have been accomplished in determining the emotional states with the use of Deep Neural Networks (DNNs). Nevertheless, most of these methods fail to consider multiple modalities to model affect, or do not fully exploit the strong representation capabilities of DNNs as they also rely on hand-engineered features.

In this thesis, we aim to provide deep learning methodologies and models for the multimodal continuous affect recognition task, trained in an end-to-end manner, i. e. utilising the raw input signal. In particular, we show that utilising multiple modalities in an end-to-end deep learning model enhances the performance of the model than using a single modality when trained under controlled or naturalistic environmental conditions. Although using multiple modalities can be beneficial to the model's performance, it is important how the information from multiple modalities is fused to the model. To this end, we investigate the integration of attention methods to fuse multiple modalities in a deep neural network architecture, and present the performance gains these methods provide compared to conventional approaches. Finally, observing the high performance of end-to-end DNN models, we provide a toolkit that provides end-to-end learning capabilities.

Contents

License **3**

Originality **5**

Acknowledgements **7**

Abstract **9**

1 Introduction **15**

1.1 Problem Scope and Challenges 15

1.2 Objectives 17

1.3 Contributions 19

1.4 Impact and Applications 21

1.5 Publications 22

1.5.1 Related Publications 22

1.5.2 Other Publications 24

1.6 Thesis Outline 25

2 Literature Review **27**

2.1 Introduction 27

2.2 Theories of Emotion 28

2.2.1 Speech Signals 28

2.2.2 Quantifying Affect 29

2.3 Affective Computing 30

2.4 Databases . . . 31
2.5 Objective Function . . . 33
2.6 Unimodal Emotion Recognition . . . 34
2.6.1 Visual Emotion Recognition . . . 36
2.6.2 Speech Emotion Recognition . . . 37
2.6.3 Text Emotion Recognition . . . 38
2.7 Multimodal Emotion Recognition . . . 38
2.7.1 Fusion Strategies . . . 40
2.7.2 Related Research . . . 41

3 Multimodal Affect Recognition in Controlled Environments 45
3.1 Motivation . . . 45
3.2 Proposed Method . . . 49
3.2.1 Visual Network . . . 49
3.2.2 Speech Network . . . 51
3.2.3 Multimodal Network . . . 54
3.3 Experiments & Results . . . 55
3.3.1 Database . . . 55
3.3.2 Experimental Setup . . . 56
3.3.3 Training Process . . . 57
3.3.4 Ablation study . . . 58
3.3.5 Speech Modality . . . 60
3.3.6 Visual Modality . . . 63
3.3.7 Multimodal Analysis . . . 65
3.4 Conclusions . . . 66

4 Multimodal Affect Recognition in Naturalistic Conditions 68
4.1 Motivation . . . 68
4.2 Notation . . . 71

4.3 Proposed Method . . . 71
4.3.1 Unimodal Networks . . . 72
4.3.2 Attentional Fusion Strategies . . . 76
4.3.3 Multimodal Model . . . 82
4.4 Experiments & Results . . . 82
4.4.1 Database . . . 82
4.4.2 Training Process . . . 83
4.4.3 Experimental Setup . . . 84
4.4.4 Unimodal Ablation Study Results . . . 85
4.4.5 Unimodal Results . . . 87
4.4.6 Fusion Strategy Results . . . 89
4.4.7 Multimodal Results . . . 90
4.4.8 Cross-dataset Evaluation . . . 91
4.4.9 Hyper-parameter Optimisation . . . 92
4.5 Conclusions . . . 93

5 End-to-End Multimodal Profiling Applications 94
5.1 Motivation . . . 94
5.2 End2You Toolkit . . . 98
5.2.1 System Overview . . . 99
5.3 Experiments . . . 100
5.3.1 Multimodal Emotion Recognition . . . 101
5.3.2 Facial Emotion Recognition . . . 102
5.3.3 Interspeech Computational Paralinguistics ChallengE (ComParE) . . . 103
5.3.4 Multimodal Sentiment Analysis (MuSe) . . . 109
5.3.5 Eating Analysis and Tracking (EAT) . . . 112
5.4 Conclusions . . . 115

6 Conclusion **116**

6.1 Contributions 117

6.2 Future Work 118

List of Tables **121**

List of Figures **125**

Bibliography **129**

Chapter 1

Introduction

1.1 Problem Scope and Challenges

Human-to-Human Interaction (HHI) is prevailed by the emotional behaviour of individuals. Several research fields study the human behaviour including, but not limited to, cognitive science, linguistics, psychology, and computer science. The branch of computer science that studies the human emotions from an intelligent system's point of view, such as robots and virtual humans, is called affective computing. Its goal is to augment intelligent systems with the capability to detect, understand, and process emotions of humans. Essentially, affective computing tries to integrate emotional intelligence to machines.

To achieve this goal it is important for intelligent systems to consider multiple cues (modalities) as human emotions are presented with many different modalities including, but not limited to, visual [40], acoustic [182], body gestures [50], text [3], and physiological signals such as Electrocardiogram (ECG) [164], Electroencephalogram (EEG) [2], and Electromyogram (EMG) [78].

However, the task of intelligent systems to recognise emotions is challenging. First, human emotions have fuzzy temporal boundaries. Difficulties arise into specifying

the start or the completion of an emotion. In addition, emotions are expressed differently for each individual, and one utterance may contain more than one emotion [5].

Affective computing field has been studied for more than 25 years, with traditional approaches to try to extract from the raw signal hand-engineered (or hand-crafted) features which were built upon knowledge gained in decades of research and have shown to be robust for many domains. Examples of such features, for instance in the audio field are Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP) coefficients, and supra-segmental features. After the feature extraction step, a machine learning algorithm is used to find appropriate for the task at-hand patterns.

With the advent of Deep Neural Networks (DNN) in the last decade, a number of groundbreaking improvements have been observed in several established pattern recognition areas such as object, speech and speaker recognition, as well as in combined problem solving approaches, e. g., in audiovisual recognition, and in the rather recent field of paralinguistics. DNN can be broadly grouped in two categories depending on the representation information they extract: *(i)* spatial DNN like Convolution Neural Networks (CNN), that extract spatially features, and *(ii)* temporal DNN like Long Short-Term Memory (LSTM), that find temporal patterns in the signal. These network variants can model the inherent structure contained in the modality signals, and can derive a representation of the input signal directly from *raw* (end-to-end), unprocessed data utilising as little human a-priori knowledge as possible.

Although several advancements have been observed in the affective computing field, a number of challenges remain unsolved, which limit the applicability of current systems. First, several systems have mainly focused on recognising emotions under constrained (laboratory) conditions meaning human and environmental constraints. These conditions do not allow the models to generalise to naturalistic settings where

realistic behavioural (i. e. unconstrained) conditions of the subject and the environment are entailed. In addition, it is not straightforward how to exploit fusing multiple modalities in a single network. Temporal variations in the signals due to different sampling rate enhance this problem. Finally, the literature lacks an end-to-end baseline models completely entailed by DNN that can focus on either unimodal or multimodal models.

1.2 Objectives

In this thesis we aim to improve (in terms of performance criteria) the emotional Artificial Intelligence (AI) behaviour analysis systems. The methods will be applied to controlled and naturalistic data, which focus on capitalising on the power of deep learning methodologies. In particular, the following questions are to be addressed:

1. *Does integration of multiple modalities in a single system have better performance over unimodal systems when deep learning models are trained and evaluated in an end-to-end manner for continuous emotion recognition?*

2. *How do end-to-end deep learning methodologies perform when trained and evaluated under spontaneous conditions for continuous emotion recognition?*

3. *Can attention methods improve the performance of a multimodal model when trained under naturalistic conditions in an end-to-end manner?*

4. *Can we provide a toolkit to the affective computing community that provides profiling end-to-end learning capabilities?*

Objective 1. Studies in the literature have indicated that, with the use of hand-crafted features, multimodal emotion recognition systems can outperform the unimodal ones [115, 73] because they can exploit the complementary information in the input signals. However, at the time of this thesis there was no study showing the

results of end-to-end deep neural network models. For our purposes, we use (at the time) well-established end-to-end deep learning architectures to show the benefits of combining multiple input signals. The raw information is utilised as input to the model and thus performing end-to-end learning. This answers Question 1.

Objective 2. Several deep learning models have been proposed in the literature for continuous emotion recognition [56, 69]. However, there is no study to show the evaluation performance of each model. On top of that, the proposed studies do not show how the models are affected from the different hyperparameters. In this thesis, we study the performance of deep learning models by analysing spontaneous and naturalistic behaviour that was collected from web-cameras and microphones. For our purposes, we leverage constrained and naturalistic continuous emotion recognition datasets that provide multiple modalities, namely, the recently proposed REmote COLlab-orative and Affective (RECOLA), and the Sentiment Analysis in the Wild database (SEWA) datasets.

Objective 3. Combining the information from different modalities is vital for the multimodal system's performance. Traditional approaches concatenate the per modality information before further processing. This approach assumes the unimodal information of each modality to be of equal importance. However, this can be problematic when for example a sensor fails, and hence its information becomes redundant. Motivated by this, we study the problem of fusing the multiple input signals in a unified representation using attention-based methods. This answers Question 3.

Objective 4. Currently, highly hand-engineered features (e. g., Low Level Descriptors (LLD) for the audio modality, and geometric for the visual one) are utilised for different types of tasks which are unique to each modality. These features are extracted from well-established toolkits (e. g. OpenSmile [43]), and are heavily used to provide baseline results for several emotion recognition challenges such

as the Audio/Visual Emotion Challenge and Workshop (AVEC) [116, 115]. On the other hand, there is no toolkit that provides end-to-end deep learning approaches for the affective computing community. Recently proposed DNN-based toolkits utilise hand-crafted features (e. g. spectrograms) as input. To this end, we develop End2You, an end-to-end learning toolkit that provides multimodal profiling capabilities. This answers Question 4.

1.3 Contributions

In this section, the main contributions of this Ph.D. thesis are described in more detail and are related to the aforementioned objectives of Sec. 1.2.

- **Multimodal Affect Recognition in Controlled Environments** (chapter 3)

  Emotion understanding is an imperative component of intelligent systems in order to achieve a natural interaction with humans. For numerous years hand-engineered features that were built upon knowledge gained after decades of research, have been used in the affective computing field However, the last ten years a shift has been observed in the community towards utilising deep neural networks (DNN). With the advent of DNN, a number of groundbreaking improvements (performance-wise) have been observed in the emotion recognition domain. In this chapter, we study automatic affect sensing using both speech and visual information in an end-to-end manner. In more detail we propose and train a newly proposed end-to-end speech emotion recognition model and train an end-to-end visual emotion recognition model using a residual network of 50 layers, and show their corresponding performances. Finally, we integrate the information extracted by the speech and visual networks in a single audiovisual emotion recognition system. The models are trained using the RECOLA dataset, in which videos were captured under controlled (or posed

or laboratory) environmental conditions. This chapter provides solution to Objective 1 and (partially) 2 of Sec. 1.2.

- **Multimodal Affect Recognition in Naturalistic Conditions** (chapter 4)

  The affective computing community has wildly focused on controlled environmental setting, mainly due to its simplicity. However, these conditions do not apply to naturalistic (i. e. "in-the-wild") environments where unconstrained conditions of visual and speech are entailed [13]. In this chapter, we study the automatic continuous affect recognition task in naturalistic conditions. In particular, we utilise text, audio, and visual modalities to develop our in-the-wild multimodal system. We propose a transformer-based [163] architecture to extract information from the text modality, and use state-of-the-art speech- and visual-emotion recognition networks. The choice of the best performing audio and visual models is conducted based on an extensive evaluation of current state-of-the-art approaches. In addition, we present novel attention-based fusion strategies to combine the information extracted from the unimodal models. Using attention methods provides us with the capability to visualise which modality has the highest contribution to each frame, and the words that contribute the most to the final prediction. This chapter provides solutions to Objectives 3 and (partially) 2 of Sec. 1.2.

- **End-to-End Multimodal Profiling Applications** (chapter 5)

  Although DNNs have greatly improved several different fields of artificial intelligence including the affective computing field, still several hand-engineered features are utilised, especially in the audio domain that have long been studied and provide human a-priori knowledge in the field. However, DNNs provide the capability to be trained in an end-to-end manner, i. e., using *raw* information, with the least possible human a-priori knowledge. In this chapter, we propose the End2You toolkit – the Imperial College London toolkit for mul-

timodal profiling. The toolkit provides network architectures, training and evaluation capabilities with either unimodal or multimodal input which can be audio, visual, audiovisual, or any other kind of modality or sensor data such as physiological data. In addition, it provides a number of multimodal fusion strategies using attention mechanisms. Finally, we extensively experimented with our toolkit in a number of tasks where we compare its performance with other well-established toolkits. This chapter provides solution to Objectives 4 of Sec. 1.2.

1.4 Impact and Applications

As already discussed end-to-end multimodal emotion recognition models are trained without any human priors (i. e. hand-engineered features). This is very powerful as we are getting away from severely engineered features to ones that are highly tailored to a specific task and we are obtaining from weak assumptions that the hand-engineered features make. This highlights the impact of developing end-to-end learning models.

The impact of the proposed research direction is immense, spanning a wide, diverse range of applications, namely:

Advertisements. Advertisement systems can benefit from emotional intelligence as they can deliver an enhanced user experience by creating more targeted advertisements for the consumers.

Video Games. Currently lacks emotional understanding of individuals. An emotion-aware game can adapt its difficulty based on the player's responses.

Online Tutoring. The computer system interacting with humans should understand their emotion state in order to react and "behave" properly. The online

tutoring systems can adapt the teaching process based on the emotional state of individuals and achieve better learning outcome.

Additionally, the proposed ideas of this Ph.D. thesis, along with the provided open-source implementations, have the potential to accelerate research in other disciplines, such as healthcare, education, and robotics. This is possible by leveraging the multimodal end-to-end DNN models.

1.5 Publications

In this section, we provide a list of publications that were authored during the course of this Ph.D. thesis. We split these publications in two categories: (i) those that are related to the contents of this thesis 1.5.1 and (ii) other publications that are not directly relevant 1.5.2.

1.5.1 Related Publications

The work presented in this thesis, is directly related to the following publications:

- **Tzirakis, P.**, Trigeorgis, G., Nicolaou, M.A., Schuller, B.W. and Zafeiriou, S.. *End-to-end multimodal emotion recognition using deep neural networks*. IEEE Journal of Selected Topics in Signal Processing, 11(8), pp.1301-1309, 2017.

- **Tzirakis, P.**, Zafeiriou, S. and Schuller, B. *Real-world automatic continuous affect recognition from audiovisual signals*. In Multimodal Behavior Analysis in the Wild (pp. 387-406). Academic Press, 2019.

- **Tzirakis, P.***, Zhang, J.* and Schuller, B.. *End-to-end speech emotion recognition using deep neural networks*. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089-5093, 2018.

- **Tzirakis, P.***, Chen, J.*, Zafeiriou, S., Schuller, B., *End-to-end multimodal affect recognition in real-world environments*, Information Fusion, Volume 68, pp. 46-53, 2021.

- **Tzirakis, P.**, Zafeiriou, S. and Schuller, B.. *End2You–The Imperial Toolkit for Multimodal Profiling by End-to-End Learning.* arXiv preprint arXiv:1802.01115, 2018 (submitted to JMLR).

- **Tzirakis, P.**, Boros, D., Hajiyev, E., and Schuller, B.. *Facial Emotion Recognition using Deep Residual Networks in Real-World Environments*, 2021.

- Stappen, L., Baird, A., Rizos, G., **Tzirakis, P.**, Du, X., Hafner, F., Schumann, L., Mallol-Ragolta, A., Schuller, B., Lefter, I., Cambria, E., and Kompatsiaris, I.. *MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media: Emotional Car Reviews in-the-wild.* In Proc. of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (MuSe'20). In Proc. ACM Multimedia, pp. 35–44, 2020.

- Schuller, B., Steidl, S., Batliner, A., Marschik, P.B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F.B., Rathner, E., Bartl-Pokorny, K.D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., **Tzirakis, P.**, Zafeiriou, S. *The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats.* In Proc. Interspeech, pp. 122-126, 2018.

- Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A.S., Hidalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., Trigeorgis, G., **Tzirakis, P.**, Zafeiriou, S. *The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold &*

Snoring. In Proc. Interspeech, pp. 3442-3446, 2017.

- Schuller, B. W and Batliner, A. and Bergler, C. and Mascolo, C. and Han, J. and Lefter, I. and Kaya, H. and Amiriparian, S. and Baird, A. and Stappen, L. and Ottl, S. and Gerczuk, M. and **Tzirakis, P.** and Brown, C. and Chauhan, J. and Grammenos, A. and Hasthanasombat, A. and Spathis, D. and Xia, T. and Cicuta, P. and Rothkrantz, J.M. L. and Zwerts, J. and Treep, J. and Kaandorp, C. *The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, escalation & primates*, arXiv:2102.13468, 2021.

- Tzirakis, P., Boros, D., Hajiyev, E., and Schuller, B.W.. *Facial Emotion Recognition using Deep Residual Networks in Real-World Environments*, arXiv:2111.02717, 2021.

1.5.2 Other Publications

- **Tzirakis, P.**, Papaioannou, A., Lattas, A., Tarasiou, M., Schuller, B. and Zafeiriou, S., *Synthesising 3D Facial Motion from" In-the-Wild" Speech*, In Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp. 627-634, 2020. **Oral**

- **Tzirakis, P.***, Shiarella, A.*, Ewers, R. and Schuller, B. *Computer Audition for Continuous Rainforest Occupancy Monitoring: The Case of Bornean Gibbons' Call Detection*, In Proc. Interspeech, pp.1211-1215, 2020.

- **Tzirakis, P.**, Nicolaou, M.A., Schuller, B. and Zafeiriou, S., 2019, May. *Time-series clustering with jointly learning deep representations, clusters and temporal boundaries*, In Proc. IEEE International Conference on Automatic Face & Gesture Recognition (FG) pp. 1-5, 2019.

- Kollias, D., **Tzirakis, P.**, Nicolaou, M.A., Papaioannou, A., Zhao, G., Schuller, B., Kotsia, I. and Zafeiriou, S., 2019. *Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond*, International Journal of Computer Vision, 127 (6-7), pp. 907-929, 2019.

- Özbay, A.G., Laizet, S., **Tzirakis, P**., Rizos, G. and Schuller, B.. *Poisson cnn: Convolutional neural networks for the solution of the poisson equation with varying meshes and dirichlet boundary conditions*, arXiv preprint arXiv:1910.08613, 2019.

- Hantke, S., Schmitt, M., **Tzirakis, P.** and Schuller, B.. *EAT- The ICMI 2018 Eating Analysis and Tracking Challenge.* In Proc. ACM International Conference on Multimodal Interaction. pp. 559-563, 2018.

- **Tzirakis, P.**, Kumar, A., and Donley, J., *Multi-Channel Speech Enhancement using Graph Neural Networks.* In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

- **Tzirakis, P.**, Nguyen, A., and Zafeiriou, S., Schuller, B. *Speech Emotion Recognition using Semantic Information.* In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021.

- Akman, A., Coppock, H., Gaskell, A., **Tzirakis, P.**, Jones, L. and Schuller, B.W., *Evaluating the COVID-19 Identification ResNet (CIdeR) on the INTERSPEECH COVID-19 from Audio Challenges.* preprint arXiv: 2107.14549, 2021.

1.6 Thesis Outline

The present study is structured into six main chapters as follows.

Chapter 1 primarily presents an introductory motivation and the raised objectives of the thesis, followed by a structure of the present study.

Chapter 2 covers the theoretical background of emotion along with the current state of affective computing field. In particular, recent databases, and unimodal and multimodal emotion recognition systems are described.

Chapter 3 presents a multimodal emotion recognition system in posed environments, i. e., with environmental constraints. In addition, a methodology on how to built speech emotion recognition models is described.

Chapter 4 presents an affective computing system that can operate on naturalistic environments. More specifically, a novel text architecture is presented, along with attention-based fusion strategies.

Chapter 5 presents an end-to-end toolkit that provides capabilities for training unimodal and multimodal systems. The toolkit has been used in a number of challenges that are described in the Chapter.

Chapter 6 summarises the contributions of this thesis along with the weaknesses of the proposed approaches. Furthermore, future research directions are outlined.

Chapter 2

Literature Review

Contents

1.1 Problem Scope and Challenges 15

1.2 Objectives 17

1.3 Contributions 19

1.4 Impact and Applications 21

1.5 Publications 22

1.6 Thesis Outline 25

2.1 Introduction

In this chapter, our first aim is to present the audiovisual cues that convey emotions and how emotions can be modelled. Our second purpose is to describe the advances of the affective computing community towards continuous emotion recognition facing posed and naturalistic conditions.

The chapter is organised as follows. The first part (Sec. 2.2) describes how emotions can be perceived from audio and visual information, the main categories emotions can be quantified in, and the key differences between naturalistic and laboratory

settings. The second part is focused on the affective computing field (Sec. 2.3) by first describing the main publicly available databases (Sec. 2.4) along with the function that is commonly used for training and evaluation of deep learning models (Sec. 2.5). Then, we discuss the progress achieved using only one modality (Sec. 2.6), and in particular the audio, visual, and text ones that are more relevant to this thesis. Finally, models using multiple modalities are presented (Sec. 2.7), focusing on the audiovisual domain.

2.2 Theories of Emotion

Emotions are inherent biological states of humans. They affect several aspects of our lives including mental/physical health, work, and social interactions [39]. Humans can automatically and subconsciously recognise emotions [168]. To do so, the (human) emotion recognition system takes advantage of the multiple cues and signals (or modalities) that are available including but not limited to visual, acoustic, body gestures and others [19]. In this section we present the signals that convey emotion in the speech modality.

2.2.1 Speech Signals

An important modality for emotion recognition is speech, which conveys both linguistic and acoustic information [129]. Although linguistic information can be helpful in speech emotion recognition, it can be misleading, as the words can be misinterpreted. For instance, sarcasm is difficult to be understand from spoken words. Furthermore, spoken language depends on the individual, meaning that different words are used from different humans for exactly the same thing [125].

On the other hand, acoustically encoded information is often very significant for affect recognition. For example, the intonation, intensity, and speaking rate of the voice convey vital information for the arousal of the speaker. More particularly, com-

mon findings in the literature suggest that the mean of the fundamental frequency (F0), the speech rate, and the mean energy of the speech signal are positively correlated with the arousal of an individual [20, 71]. However, the reality of today's engines is much more complex than that with several tens to even thousands of characteristics being used to model speech acoustics for emotion recognition.

2.2.2 Quantifying Affect

Quantifying emotions has long been in the interest of various research areas like psychology and neuroscience. According to Grandjean et al. [51], there are three major approaches to characterise emotional models: categorical, dimensional, and appraisal-based. These are the ones mainly used in technical approaches these days, whereas in particular categorical and dimensional approaches are pre-dominating.

Categorical Emotions

The categorical approach indicates that there are a few emotions that are basic, and can be recognised universally independent of the race. This theory has been extensively studied in the literature for many years. The most famous model is the model of Ekman [38], which suggests that there are six basic emotions that can be recognised universally, namely, surprise, fear, disgust, anger, joy, and sadness. However, many emotion researchers believe that the affective states are a lot more complex, and that the emotions are not independent from one another, but related in a systematic manner [119]. Hence, a small number of states cannot capture the complexity of the emotions.

Dimensional Emotions

The dimensional approach suggests that emotions are not independent, but related to one another. Russell [119] proposed to characterise emotions with a small number

of latent dimensions. To this end, he described emotions with a circular configuration called Circumplex of Affect. This model represents the affective states in a two dimensional bipolar space. These dimensions are the arousal, which indicates how apathetic (i. e., boredom) or active (i. e., excited) the subject is, and valence, which indicates the degree of pleasant or unpleasant feelings. Figure 2.1 depicts this model.

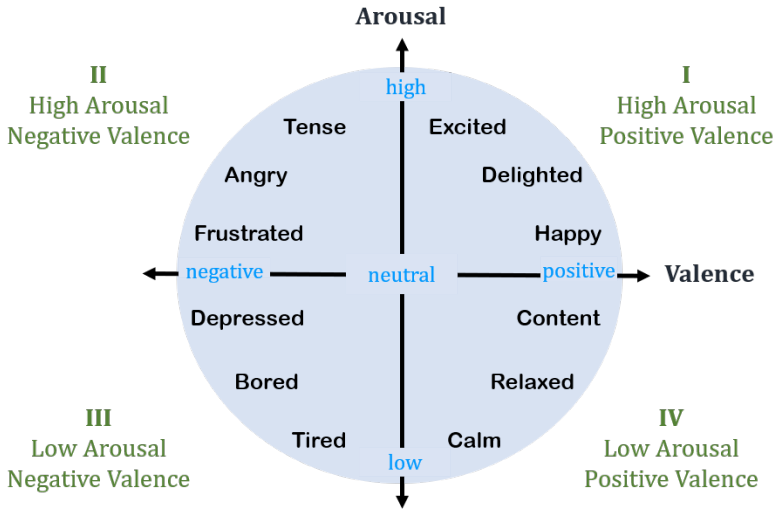


Figure 2.1: Circumplex of Affect. The horizontal axis represents the valence and the vertical the arousal. [119]

Appraisal-based Emotions

A relatively newer approach that models affect was introduced by Scherer, and is called Component Process Model (CPM) [122]. It is based on the appraisal-based approach [45, 51]. In this method, the emotions are assumed to be generated through the evaluation of events from the outside world and the internal state of the individual (e. g., motivation, physiological reactions).

2.3 Affective Computing

Affective computing is the field of science that tries to recognise, analyse, and predict the affective (syn. *emotional*) state of humans. It has been studied for more than 20 years and it is an important step towards a complete and natural Human-

Computer Interaction (HCI). Applications can be found in social psychology, video conferencing, user profiling, image retrieval, psychological area, e-learning, health care, gaming and others.

However, recognising emotion is a very challenging task. First, the duration of human emotions vary significantly and depends on the person and the situation. In addition, emotions are expressed differently among different individuals [5] and cultures. The difficulty of affect recognition is further escalated in naturalistic environments where uncontrolled conditions are entailed.

In the next sections, we focus on presenting the progress achieved to day in the affective computing community. First, we describe common databases that are used in the field and are relevant to this thesis. Thereafter, after we introduce the Concordance Correlation Coefficient (CCC) measure and show how it can be used as a loss function, we shortly describe the main unimodal emotion recognition literature focusing on audio, visual and textual information. Finally, after presenting the main multimodal fusion strategies, we discuss recent studies for the multimodal emotion recognition domain.

2.4 Databases

Databases are vital to train and evaluate models for affect recognition. The methods of collecting affect data can be categorised as follows [171]: (i) posed (or laboratory), (ii) naturalistic (or induced or in-the-wild), and (iii) spontaneous (or real-world). The posed method indicates that an actor was used to express different emotions. In the naturalistic class, a subject's emotions are triggered by, for example, watching TV or listening to music or imagining a story, etc. Finally, the real-world databases are collected when a spontaneous emotion is expressed by a subject when he/she does not know that he/she is recorded.

Before continuing the discussion on well-established databases, it is important to

further clarify the differences between naturalistic and laboratory settings, as both these settings are provided by the datasets that are used for the experiments. The main differences of these two settings are the following:

- **Environmental conditions.** Laboratory settings consider controlled environmental conditions, i. e., without noise or reverberation. However, in naturalistic settings the environmental conditions are not controlled. In addition, there may be different kinds of occlusions, which do not occur in a controlled environment.
- **Subject movements and linguistics.** Restrictions may be applied to the subject in terms of bodily movements, either moving head/hands or changing its position, in the laboratory settings. However, no such constrains apply in naturalistic situations. The subject can freely make any movement he/she desires.
- **Subject's expressions.** Posed settings may limit the expressions the participants can perform. On the other hand, in a naturalistic environment humans do not restraint themselves and can perform any kind of expression. Similarly, the verbal content may be prompted or predifend in lab settings and thus limited in some way. Obviously, this would be different in naturalistic conditions.

Since the discussion of this thesis is focused on naturalistic continuous emotion recognition, we present well known databases focusing on in-the-wild conditions and continuous affect prediction for audio-visual purposes. For further study the interested reader is forwarded to [171].

Numerous databases have been proposed in the literature for continuous emotion recognition. One of the first databases for continuous affect recognition was the Humaine database [35], which includes 50 recordings with multiple modalities, namely

speech, language, gesture, face and physiological. Another database is the Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression Dataset (SEMAINE) [99], which contains 959 conversations from 150 participants where each participant is interacting with an automatic agent. The modalities it supports are audiovisual. The Belfast induced natural emotion database [136] also includes audiovisual clips for various tasks. However, few of them are for continuous emotion recognition. Another database is the MAHNOB mimicry dataset [16], which was the first natural human-to-human interaction and consists of 54 audiovisual recordings (11 h) from 12 confederates and 48 counterparts. As a final example, the Acted Facial Expressions in the Wild (AFEW) [33] dataset is a facial expression data corpus. In a late edition, it is comprised of 1 426 video clips extracted from movies and annotated into six emotions plus a neutral: anger, disgust, fear, sadness, happiness, surprise, and neutral. Static frames were extracted from the video clips of AFEW, and the Static Facial Expressions in the Wild (SFEW) was created.

2.5 Objective Function

The Concordance Correlation Coefficient [87] (CCC) is frequently-used measure for emotion regression. It evaluates the agreement level between the predictions of the network and the gold-standard derived from the annotations. It has extensively been used the last 10 years in several publications [113, 116] as evaluation metric. Mathematically, given two time series x and y, the CCC is calculated as follows:

$$CCC = \frac{2\sigma_{xy}^2}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{2.1}$$

where $\mu_x = E(x)$, $\mu_y = E(y)$, $\sigma_x^2 = var(x)$, $\sigma_y^2 = var(y)$ and $\sigma_{xy}^2 = cov(x, y)$.

Several studies in the literature used to minimise the Mean Squared Error (MSE) during the training of the networks, and evaluated the models with respect to

CCC [113, 116]. In our studies, we propose to include the metric used to evaluate the performance in the objective function ($\mathcal{L}_c$) used to train the networks. Since the objective function is a cost function, we define $\mathcal{L}_c$ as follows:

$$\mathcal{L}_c = 1 - CCC = 1 - 2\sigma_{xy}^2\psi^{-1}, \tag{2.2}$$

where $\psi = \sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2$. The gradient of the loss to be propagated from the last layer with respect to the weights is

$$\frac{\partial \mathcal{L}_c}{\partial x} \propto 2\frac{\sigma_{xy}^2(x - \mu_y)}{\psi^2} + \frac{\mu_y - y}{\psi}, \tag{2.3}$$

where all vector operations are done element-wise.

For our purposes we train our networks to regress N output, and as such we define the loss function as:

$$\mathcal{L} = \frac{\sum_{i=1}^{N} \mathcal{L}_c^i}{N} \tag{2.4}$$

where $\mathcal{L}_c^i$ is the concordance loss for output i.

2.6 Unimodal Emotion Recognition

A number of approaches have been proposed the last 20 years in the literature exist that try to predict the emotional states of individuals using unimodal information such as audio, visual, or text [6, 14, 15, 27, 41, 126, 19, 37]. The pipeline of traditional machine learning approaches [158, 148], first, involved extracting features from the unimodal signal. Examples of extracted features for the audio modality

include Mel-Frequency Cepstral Coefficients (MFCCs), and Linear Prediction Cepstral Coefficient (LPCCs) in the audio domain, for the visual modality are the Histogram of Oriented Gradients (HOG), Local Binary Pattern from Three Orthogonal Planes (LBP-TOP), Local Phase Quantisation (LPQ), and Scale-Invariant Feature Transform (SIFT) for the visual modality, and for the text modality Term Frequency–Inverse Document Frequency (TF-IDF), and Bag of Words (BoW). These features are then passed to a machine learning algorithm. Examples of such algorithms include K-Nearest Neighbour (K-NN), Random Forest, and Support Vector Machines (SVM).

Although these algorithms had widely been used, they now become more and more obsolete due to the advent of Deep Neural Networks (DNN) [88] which outperform traditional approaches in several areas including computer vision, and speech analysis. These networks learn hierarchies of features where the higher levels are formed by the composition of lower level features. They are composed of multiple layers stacked on top of each other, with each layer learns representations from the previous layer. The layers are composed of non-linear modules that transform the representation of the previous layer. Stacking multiple layers in this manner complex functions can be learned. A series of neural network architectures have been revitalised, such as Autoencoder networks, Convolutional Neural Networks (CNNs), Deep Belief Networks (DBNs) or memory enhanced neural network models such as Long Short-Term Memory (LSTM) models.

In this section, we focus on providing a small review of the most recent and well-known emotion recognition studies in the literature that utilise DNN models either in an end-to-end manner, meaning using the raw input representation, or having a feature extraction step before using a DNN.

2.6.1 Visual Emotion Recognition

Visual emotion recognition is an essential component in HCI that captures non-verbal cues from facial emotions. Applications can be found in mental state identification, security, automatic counseling systems, lie detection, and automated tutoring systems. Recently several advancements have been accomplished towards a complete Facial Emotion Recognition (FER) system, mainly due to the latest CNN architectures that have been proposed in the computer vision domain such as Residual Networks (ResNet) [62], VGG-Face [107], MobileNet [64], and others. Providing sufficient data for the training, these networks have the capability to compute robust features, and as a consequence have high emotion recognition performance.

The high performance of these models is obtained given that the face is the input to these models. This can be accomplished with either a face detector such as the Multi-task Cascaded Convolutional Networks [180], or with a tracker algorithm such as Multi-Domain Convolutional Neural Network Tracker (MDNet) [103], which takes the bounding box of the face in the first frame of the video and tracks it in all frames.

Numerous papers have been proposed the last years that utilise the visual modality for the emotion recognition task. For example, Huang et al. [69] proposed a transductive learning framework for image-based emotion recognition by combining DNNs and hypergraphs. More particularly, after the DNN was trained for the emotion classification task, each node in the last fully connected layer was considered as an attribute and used to form a hyperedge in a hypergraph. In another study, Ebrahimi et al. [36] combined CNNs and RNNs to recognise categorical emotions in videos. A CNN was first trained to classify static images containing emotion. Then, the extracted features from the CNN were used to train an RNN to produce an emotion for the whole video. Yang et al. [174] proposed De-expression Residue

Learning (DeRL) where a generative model is used to extract the neutral face image, and then the method learns the residue (emotion) that remains in the generative model. In another study, [100] propose a deep learning approach based on attentional convolutional network. In a similar study, [83] uses a feature extractor based on VGG-Face [107] and a 2-layer Recurrent Neural Network (RNN) to account for the temporal information in the data, before getting the final prediction. For more affect recognition studies that exploit the visual information, the interested reader is referenced to Li et al. [90].

2.6.2 Speech Emotion Recognition

The speech signal conveys the paralinguistic information of the speaker which is an important factor to recognise emotions. Applications can be found in mobile services [179], call centres [54], and physiological assessment [95]. The Speech Emotion Recognition (SER) field has several breakthroughs with the advent of DNNs. Examples of such networks can be found in Zheng et al. [183], where extracted a log-scaled spectrogram from the raw waveform, before using CNNs for the final prediction. In another study, Xu et al. [172] proposed head fusion, a multi-head self-attention method. In a more recent studies, Neumann et al. [105] proposed an attentive convolutional neural network (ACNN) that combines CNNs and attention with a multi-view learning objective function. Xu et al. [173] proposed a hierarchical grained and feature model where features are extracted from a frame- and utterance-level structure before feeding them to a RNN for the final prediction.

Other studies use the raw waveform to model affect. For example, Trigeorgis et al. [151] proposed a deep convolutional recurrent neural network architecture utilising the raw waveform in an end-to-end training. In a more recent study, Li et al. [91] proposed the use of self-attention and global windowing in the transformer model for the task at hand. In yet another study, Triantafyllopoulos et al. [150] investigate

the usefulness of speech enhancement models such as deep residual networks for the speech emotion recognition task.

2.6.3 Text Emotion Recognition

Affect recognition via textual information, also known as sentiment analysis, has become a fundamental research area in natural language processing. Applications can be found in proliferation of user-generated content including but not limited to understanding customer feedback, and user comments in social media. As with the visual and speech modalities, DNN have achieved state-of-the-art performance on various sentiment tasks [79, 7]. Early studies focused on modelling text using RNN such as LSTM and GRU [137, 138]. Thereafter, CNN started to be used for the modelling in combination with RNN. One of the first studies was performed by Kim et al. [79] which proposed a character-aware neural language model that first learns word representations at character-level using a CNN, and then models language tasks at word-level utilising LSTM. Most recent studies rely heavily on attention mechanism methods which provide state-of-the-art performance [94, 184, 177]. Other studies focused on Multiple Instance Learning (MIL) [7], and domain adaptation [121]. The interested reader is referenced to [101] for more studies that can be found in the literature exploiting text modality for sentiment analysis.

2.7 Multimodal Emotion Recognition

Human-to-Human (HHI) interactions, and as a consequence affective behaviour, is multimodal; we see faces, hear voices, read text. Although the affective community focused on building unimodal emotion recognition systems for the past 20 years, the last 10 years there is a shift in the research field towards building multimodal systems by incorporating several modalities. The main reason is that complementary information between the modalities, and hence can further boost the performance

of the trained models. For example, it is easier to identify that someone is laughing by *looking* at their face rather than their speech.

However, several challenges exist on combining the unimodal cues [12]:

1. It is not trivial on how to represent the unimodal information such that the complementarity and redundancy of multiple modalities is exploited.

2. Long-range dependencies can be exhibited in two or more modalities, and hence an alignment is required. For example, text, which is by nature 'unaligned' with the visual modality, needs to take into account long range dependencies.

3. Fusing the modalities together is yet another challenge. Different modalities can have different predictive power for the learning task. For example, it may be the case that for some frames one modality's data are missing due to sensor failure.

4. A fourth challenge is how to transfer knowledge between the unimodal representations/features. This is important factor especially when one of the modalities has limited resources.

These challenges have been addressed by the machine learning community [12, 152, 80, 82] as are common problems that occur in different fields such as audiovisual lip reading [44], audiovisual speech recognition [109, 147], audiovisual speech enhancement [1] and others. For the purposes of this thesis the first three challenges arise.

Before diving into the current research on multimodal emotion recognition, we first briefly discuss the main fusion strategies that are being used to combine different

modalities together. Then, we present the most recent studies on multimodal emotion recognition, and give a short description on the competitions that have emerged for continuous affect recognition that are relevant to this thesis.

2.7.1 Fusion Strategies

As mentioned earlier, one common challenge in multimodal machine learning is fusing the information that is extracted from different modalities (i. e., audio, visual, physiology, or alike). For this purpose, there are three common fusion strategies that exist in the literature, namely, feature-level fusion, decision- or score-level fusion, and model-level fusion. Fig. 2.2 shows these techniques.

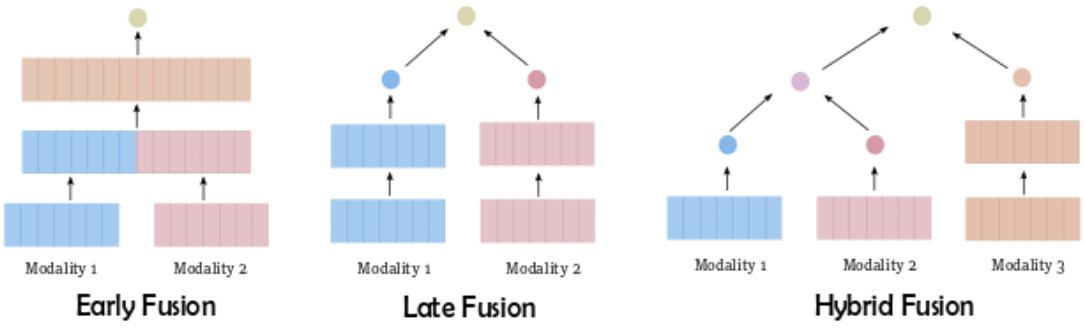


Figure 2.2: The three main fusion strategies: (a) early fusion, (b) model-level fusion, and (c) late fusion.

Early Fusion

The early (or feature-level) fusion combines the features extracted from different modalities by concatenating them in a single feature vector before passing it to a single prediction algorithm [34]. The process of this method is similar to the human 'mechanism'. That is, humans first extract features from modalities such as audio or visual cues, and then map these features to the same space before their prediction. This method requires large training datasets as more information is captured than using a single modality. An additional requirement for this method is that the

modalities should be in correspondence. The main advantage of this strategy is that a prediction will be available even if one modality's data are missing.

Late Fusion

In the late (or decision-level) fusion, each modality prediction is utilised independently, and then the scores of each modality are combined. The main disadvantage of this method is that if the data of one modality are missing, then its full capabilities cannot be exploited [34]. For classification tasks, the fusion can be performed with a simple majority vote (albeit in more sophisticated versions, weights can be introduced and even learnt), while for regression tasks, a linear regressor can be trained using the prediction for each modality and its weights can be used for the fusion.

Hybrid-level Fusion

The hybrid-level fusion combines the advantages of both the early and late fusion strategies [9]. An example of a model-level fusion is to perform a feature-level fusion of 'some' modalities and combine its prediction with scores from different modalities that were processed independently. A recent example of such a method is proposed by Wang et al. [167].

2.7.2 Related Research

As mentioned earlier, combining several modalities boost the performance of the emotion recognition models due to the complementary information that exists between them. Several approaches have been proposed across the years that exploit two or more modalities such as audio, visual and text, in a unified system [21, 159, 8, 53, 61, 152, 110, 160, 23, 149, 77, 142, 178]. The common pipeline is to extract features from the unimodal data, and use a per-modality model to find

relevant to the task at hand patterns, spatially and/or temporally, before the final prediction.

Earlier DNN studies used to extract features from the raw input modality, and use these features to train DNN models, such as Deep Belief Networks (DBF), autoencoders, or RNNs such as LSTM, to get higher level of abstractions, that have proven to have more predictive capabilities than conventional features [80]. As in traditional approaches these features were, then, input to conventional learning algorithms like SVM (or SVR for regression tasks) [80, 76, 146].

More recent studies utilise multitask learning [8], others combine fuzzy logic with DNN [23], some investigates local and global interactions between unimodal features [97], and others tried sentic blending [21], that enables the continuous interpretation of semantics, and the conceptual and affective information, in a natural language setting. Few works have been proposed for real-time analysis [149, 11]. It is worth noting of transfer learning methods [159, 181] where a network is pre-trained on large datasets and fine-tuned for the task at hand.

Of notable mention are studies that utilise attention mechanism to fuse (feature-level fusion) the modalities into a unified representation. They have gained much attention the recent years due to the high performing models it can produce, without the need of human a-priori knowledge. Attention fuses the modalities by utilising a weighted average of their features. Recent studies use attention to alleviate from the misalignment that exists among modalities because of variable sampling rates [152, 156, 60, 112]. Several attention methods have been proposed such as global/local attention [96] and transformer networks [152].

Affect Recognition Competitions

The importance of recognising emotions motivated the creation of a number of affect recognition competitions whose datasets and submitted approaches are used in

this thesis. The first ever Interspeech 2009 Emotion Challenge which was run for speech only, which featured classes, and was followed by the first ever dimensional Interspeech 2010 Paralinguistics Challenge, which had a continuous interest recognition task[1]. These were followed by another first of its kind Audio/Visual Emotion Challenge and Workshop (AVEC), which was first held in 2011 [132]. Its goal is to compare the audio, visual, and later also physiological modalities to recognise continuous emotions, as well as to estimate the severity of depression. Relevant to this thesis are the studies submitted to the challenge. To give just but a bit of insight into participants' contributions, Huang et al. [70] proposed to use variants of a Relevance Vector Machine (RVM) for modeling audio, video and audiovisual data. In another work, a model by Weber et al. [169] used high-level geometry features for predicting dimensional features. Brady et al. [18] also used low- and high-level features for modeling emotions. In a different study, Povolny et al. [111] complemented original baseline features for both audio and video to perform emotion recognition. Somandepalli et al. [139] also used additional features, but only for the audio modality.

The importance of predicting emotion in naturalistic environments led the AVEC 2017 emotion sub-challenge [115] to use the SEWA dataset which is comprised by three modalities: text, audio, and visual. The baseline model [115] used hand-crafted features in all of the modalities and a Support Vector Regressor (SVR) for the final prediction. The winning model of the challenge was proposed by Chen et al. [24] and use a multi-task learning method exploiting several deep convolutional neural networks. More particularly, they extracted visual features using the VGG-Face [107] and the DenseNet [66] networks, audio features using IS10 [128] and the SoundNet network [10]. Finally, textual features were extracted using pre-trained German Word2Vec embeddings. Their final model comprised of passing all the unimodal extracted features to a LSTM network before the final prediction. In a similar study,

[1] cf. www.compare.openaudio.eu

Huang et al. [68] utilizes several hand-crafted features in all modalities. More particularly, in audio they fine-tune on SEWA a network that was pre-trained on 300 h of speech recognition and use its features along with most of the features provided by the challenge. Similarly, in the visual modality they use geometric, appearance and deep visual features, and finally, for the text they use a Bag-of-Text-Words (BoTW). Another interesting study was submitted by Dang et al. [28] where they proposed a fusion strategy using probabilistic predictions. As the previous studies they use hand-crafted features in all the modalities; for audio they extract Bag-of-Audio-Words and eGeMAPS, for visual they exploit appearence features and for text BoTW. They form different feature sets which are passed to several regressors and create 7 subsystems that fuse them together for the final prediction.

Another well known challenge that followed the example is the Emotion Recognition in the Wild (EmotiW) [31] challenge. The first challenge was organised in 2013 [32] and its main purpose is to challenge the community on 'in-the-wild' data for emotion recognition. In 2016, two sub-challenges were proposed [31]. The first used audio-visual data for emotion recognition (VReco), and the second consisted of images with groups of people and the task was to infer the emotional state of the whole group (GReco). The winner of the GReco challenge used a CNN to extract useful information from the faces in the image and then a RNN to combine the information from the different extracted features for group emotion prediction. The winner of the VReco subchallenge used (i) a CNN with a RNN on top, and (ii) a 3D CNN to extract spatiotemporal features from the visual information of the video. These networks were trained separately with a linear classifier. In addition, they used the openSMILE [43] toolkit which has served as baseline toolkit across AVEC and EmotiW (and further challenges) to extract features from the audio and train a SVM. In the final step, they combined their predictions in a score-level fusion to obtain their final prediction.

Chapter 3

Multimodal Affect Recognition in Controlled Environments

Contents

2.1 Introduction 27
2.2 Theories of Emotion 28
2.3 Affective Computing 30
2.4 Databases 31
2.5 Objective Function 33
2.6 Unimodal Emotion Recognition 34
2.7 Multimodal Emotion Recognition 38

3.1 Motivation

As discussed in the previous chapter, human-to-human interaction (HHI) is to a large extent prevailed by the emotions, which are automatically and subconsciously understood by humans. To truly achieve a *natural* human-machine interaction, it is important for machines to be able to predict the emotional state of humans. A

multitude of applications can be benefited from automatic emotion recognition approaches such as automatic tutoring systems, where the system adjusts the level of the tutorial depending on the user's affective state, such as excitement or boredom [72].

The last ten years with the advent of deep neural networks (DNNs), a number of groundbreaking improvements have been observed not only in the emotion recognition domain but also in several established pattern recognition areas such as object, speech and speaker recognition, as well as in combined problem solving approaches, e. g., in audio-visual recognition [104]. A series of neural network architectures have been revitalised, such as autoencoder networks, Convolutional Neural Networks (CNNs), Deep Belief Networks (DBNs) or memory enhanced neural network models such as Long Short-Term Memory (LSTM) models.

Numerous studies have shown the favourable property of these network variants to model inherent structure contained in the audio and/or visual signal [63]. For example, Wöllmer et al. [170] were one of the first to propose a DNN architecture for affective computing which comprised of a three layer LSTM and was trained based on functionals of acoustic Low-Level Descriptors (LLDs). In another study, Stuhlsatz et al. [145] used Restricted Boltzmann Machines (RBM) to extract discriminative features from the raw signal and proposed a Generalized Discriminant Analysis (GerDA).

Although these studies had significant impact on the community, they used handcrafted features to train their DNN architectures. The last five years, however, a trend in the pattern recognition community has emerged towards deriving a representation of the input signal directly from *raw*, unprocessed data, attempting an *end-to-end* optimisation scheme, utilising as little human a-priori knowledge as possible [52]. The motivation behind this idea is that, ultimately, the network automatically learns an intermediate representation of the raw input signal that better

suits the task at-hand and hence leads to improved performance.

In this chapter, we study automatic affect sensing using both speech and visual information in an end-to-end manner. Features are extracted from the speech signal using a CNN architecture designed for the audio channel. In particular, we propose: *(i)* a "light" network inspired by Trigeorgis et al. [151] original publication, and *(ii)* a methodology to build speech networks inspired from the computation of conventional speech features like MFCCs, i. e., use small stride size in the convolution operation, and a large kernel size with no overlap in the max-pooling operation. To extract features from the visual modality we use a residual network (ResNet) architecture of 50 layers [62]. The output of these networks are concatenated to feed a LSTM before the final prediction. Contrary to the practices, where each network is trained individually and the results are simply fed to a subsequent classifier, our system is trained in an end-to-end manner. To our knowledge this is the first study in the literature that applies such an end-to-end model for the audio-visual emotion recognition task. On top of that, we suggest using explicit minimisation of the *concordance correlation coefficient* (CCC) [87] (see Se. 2.5) in our model and show that this improves performance in terms of emotion prediction compared to optimising the mean square error objective, which is traditionally used.

To show the benefits of our proposed multimodal model, we evaluate it in the REmote COLlaborative and Affective (RECOLA) database (see Sec. 2.4). A part of this database was used for the Audio/Visual Emotion Challenge and Workshop (AVEC) 2016. Our model is trained and tested using the whole database. Results show that the multimodal model benefits from the two modalities by producing equal results for arousal and valence as the (unimodal) speech and visual networks, respectively. We compare the unimodal and the multimodal models using results obtained from the AVEC 2016 challenge (see Sec. 2.7.2). Only the papers that used the audio, visual or audio-visual modalities are considered. In order to perform a fair comparison

we apply the proposed method on the test set of the AVEC challenge. As shown by our experiments, our unimodal models produce the best results for both the speech and visual modalities.

The contributions of this chapter can be summarised as follows:

- propose two speech networks to process the raw waveform. To create one of the networks we present a methodology to decide the kernel size of a convolution layer and a max-pooling one.
- suggest an end-to-end audio-visual emotion recognition system, where we utilise our proposed speech network with a ResNet-50 model to process the visual signal.
- propose to train and evaluate our unimodal and multimodal networks using the Concordance Correlation Coefficient (CCC) instead of the MSE loss that was widely used.

The content of this chapter is based around the following publications:

- **Tzirakis, P.**, Trigeorgis, G., Nicolaou, M.A., Schuller, B.W. and Zafeiriou, S., 2017. *End-to-end multimodal emotion recognition using deep neural networks.* IEEE Journal of Selected Topics in Signal Processing, 11(8), pp. 1301-1309.
- **Tzirakis, P.**, Zafeiriou, S. and Schuller, B., 2019. Real-world automatic continuous affect recognition from audiovisual signals. In Multimodal Behavior Analysis in the Wild, pp. 387-406. Academic Press.
- **Tzirakis, P.***, Zhang, J.* and Schuller, B.W., 2018, April. *End-to-end speech emotion recognition using deep neural networks.* In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5089-5093.

The remainder of the chapter is organised as follows. Section 3.2 introduces the end-to-end models, i. e., the two speech networks, the visual, and the multimodal ones. Section 3.3 presents the experiments performed and reports the results for the unimodal and the multimodal networks. Finally, Section 3.4 concludes this chapter.

3.2 Proposed Method

A key component of our model is the convolution operation. For the audio and visual signals, 1-d and 2-d convolution is used, respectively. The 2-d convolution is defined as follows:

$$(f \star h)(i,j) = \sum_{k=-T}^{T} \sum_{m=-T}^{T} f(k,m) \cdot h(i-k, j-m), \tag{3.1}$$

where $f(x)$ is a kernel function whose parameters are learnt from the data of the task at hand. After the spatial-modeling of the signals, which removes background noise and enhances specific parts of the signals, we model the temporal structure of both speech and video using a recurrent network with LSTM cells. We use LSTM for *(i)* simplicity, and *(ii)* to fairly compare against existing approaches which concentrated in the combination of hand-engineered features and LSTM networks. Finally, our model is subsequently trained with backpropagation by maximising the concordance correlation loss (Sec. 2.5).

3.2.1 Visual Network

Deep convolutional networks have been widely used in the computer vision domain, providing state-of-the-art results in several fields, including object detection, and face recognition. For our purposes, we utilise a deep residual network (ResNet) of 50 layers [62] to extract features from the face of individuals as it was shown to provide great results for spontaneous environmental conditions [13]. As input to the

network we used the pixel intensities from the cropped faces of the subject's video. Deep residual networks adopt residual learning by stacking building blocks of the form:

$$\mathbf{y}_k = \mathcal{F}(\mathbf{x}_k, \{W_k\}) + \mathbf{h}(x_k), \tag{3.2}$$

where $\mathbf{x}$ and $\mathbf{y}$ are the input and output of layer k, $\mathcal{F}(\mathbf{x}_k, \{W_k\})$ is the residual function to be learned and $h(\mathbf{x}_k)$ can be either an identity mapping or a linear projection to match the dimensions of the function $\mathcal{F}$ and the input $\mathbf{x}$.

The first layer of ResNet-50 is a 7×7 convolutional layer with 64 feature maps, followed by a max pooling layer of size 3×3. The rest of the network comprises of 4 bottleneck architectures, where after these architectures a shortcut connection is added. These architectures contain 3 convolutional layers of sizes 1×1, 3×3, and 1×1, for each residual function. Table 3.1 shows the replication and the sizes of the feature maps for each bottleneck architecture. After the last bottleneck architecture, an average pooling layer is inserted.

Bottleneck layer	**Replication**	**Number of feature maps** (1×1, 3×3, and 1×1)
1st	3	64, 64, 256
2nd	4	128, 128, 512
3rd	6	256, 256, 1024
4th	3	512, 512, 2048

Table 3.1: The replication of each bottleneck architecture of the ResNet-50 along with the size of the features maps of the convolutions.

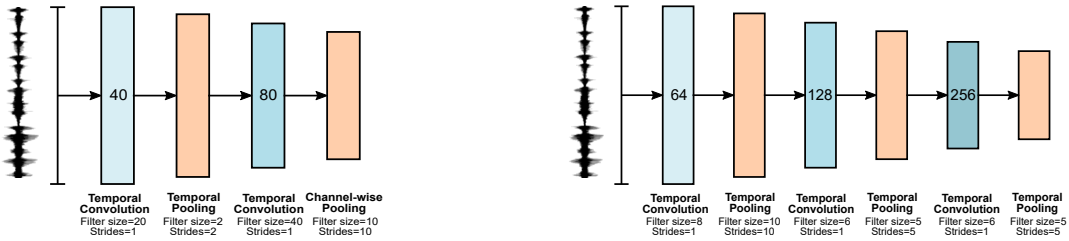
3.2.2 Speech Network

Our speech network is aimed to extract features from the raw waveform and learn the regression task in one jointly trained model for predicting the emotional state of individuals. The first attempt was performed by Trigeorgis et al. [151] and it was a nearly-blind search of the hyperparameters of the convolution neural network, which outperformed the at the time state-of-the-art models. We simplify their model by using smaller kernel size of the convolutions that results in faster training and higher performance. Our decision to search for a smaller kernel size stems from the process of extracting conventional features, where small kernel size is used. We call our model *Emo-16*.

In addition, we propose another speech network architecture that is inspired by the the way conventional speech features, like MFCCs, are computed. To the best of our knowledge, none of the studies in the literature mention a methodology on how the pooling size should be selected based on the kernel size of the convolution layer. In most cases, these are selected experimentally. For our purposes, we propose a simple methodology that computes the kernel size and stride of the convolution and the max-pooling operations. More specifically, we define the rate of overlap (R) between kernel size (K) and pooling size (P) as follows:

$$R = \frac{K - 1}{K + P - 1}. \tag{3.3}$$

The overlapping ratio R is less than 1 and its common value for creating hand-crafted features is considered most of the times around 0.5. If we use stride instead of max-pooling to reduce the dimensionality of the signal, we need to keep the rate of overlap around 0.5. However, we found that using convolution provides worse performance than using max-pooling. An intuition behind this finding is that first we extract the most important information and discard the futile, while it takes



(a) The Emo-16 architecture (b) The Emo-18 architecture

Figure 3.1: Depicting the (a) Emo-16, and (b) Emo-18 architectures. Emo-16 is comprised of two blocks of convolution and max-pooling with the second's block max-pooling to be applied in the channels. On the other hand, Emo-18 is comprised of three blocks of convolution and max-pooling, all of which operate in the time domain.

all information into consideration when using strides. At this time, we need to keep R less than 0.5 since we do not want it to extract the similar information for successive frames. To this end, between convolution and max-pooling layers we consider $R < 0.5$ (i.e. $R \approx 0.4$) for all layers. We call the derived model *Emo-18*. Table 3.2 depicts the architecture of the two models, which are also described in the two following sections. Fig. 3.1 depicts the two networks.

Emo-16

Input. The raw waveform is segmented to 6 s long sequences after it is preprocessed to have 0 dB reference level. At 16 kHz sampling rate, this corresponds to a 96000-dimensional vector, which is the input to the speech network.

Temporal Convolution. We use $F = 40$ space time finite impulse filters with a 5 ms window (size of 80) in order to extract finer-scale spectral information from the high sampling rate signal (16 kHz).

Pooling across time. The impulse response of each filter is passed through a half-wave rectifier (analogous to the cochlear transduction step in the human ear) and then downsampled to 8 kHz by pooling each impulse response with a pool size = 2.

Speech Network	Layer	Kernel/Stride	Channels	Activation
Emo-16	Convolution	20/1	40	ReLU
	Max-pooling	2/2	—	—
	Convolution	40/1	40	ReLU
	Max-pooling (*channel-wise*)	10/10	—	—
Emo-18	Convolution	8/1	64	ReLU
	Max-pooling	10/10	—	—
	Convolution	6/1	128	ReLU
	Max-pooling	8/8	—	—
	Convolution	6/1	256	ReLU
	Max-pooling	8/8	—	—

Table 3.2: Network architecture for Emo-16, and Emo-18 speech networks.

Temporal Convolution. In total, $M = 40$ space time finite impulse filters of 500 ms window (size of 4000). These are used to extract more long-term characteristics of the speech and the roughness (i. e., irregularities) of the speech signal.

Max pooling across channels. Max-pooling across the channel domain is performed with a pool size of 10. This reduces the dimensionality of the signal while preserving the necessary statistics of the convolved signal.

Dropout. Due to the large number of parameters compared to the number of training examples, some regularisation is needed in order for the model not to overfit on the training data. To this end, a dropout with a probability of 0.5 was used.

Emo-18

Input. After we preprocess the raw signal to have zero mean and unit variance, we segment it to 20 s sequences long and use them as input. At 16 kHz this corresponds to 320 000-dimensional input vector.

Temporal Convolution. We use 64 time impulse filters with kernel size of 8 to extract information from the raw signal in the first convolution layer.

Max-pooling. Based on the previous kernel size (8), we apply max-pooling with a size of 10 to decrease the frame rate of the signal and keep the most descriptive features. The stride is also 10.

Temporal Convolution. Going deeper we want to extract a larger number of high-level abstractions. For this purpose, we convolve in the time domain with a kernel size of 6 and with a channel size of 128.

Max-pooling. To keep the overlap rate below 1, we pool across the time domain with a size of 8, and stride also 8.

Temporal Convolution. The last convolution layer provides us with an even higher level of abstractions. We keep the kernel size to 6 and increase the filter size to 256.

Max-pooling. Considering the kernel size of the previous convolution layer, we perform a max-pooling across time with a size of 8, and stride also 8.

As in the Emo-16 network we use dropout regularisation after each pooling layer, with a probability of 0.5.

3.2.3 Multimodal Network

Our multimodal network is comprised of the multimodal feature extraction part, which contains the speech and visual networks, and the temporal feature extraction part, which contains the RNN. Fig. 3.2 shows the multimodal network.

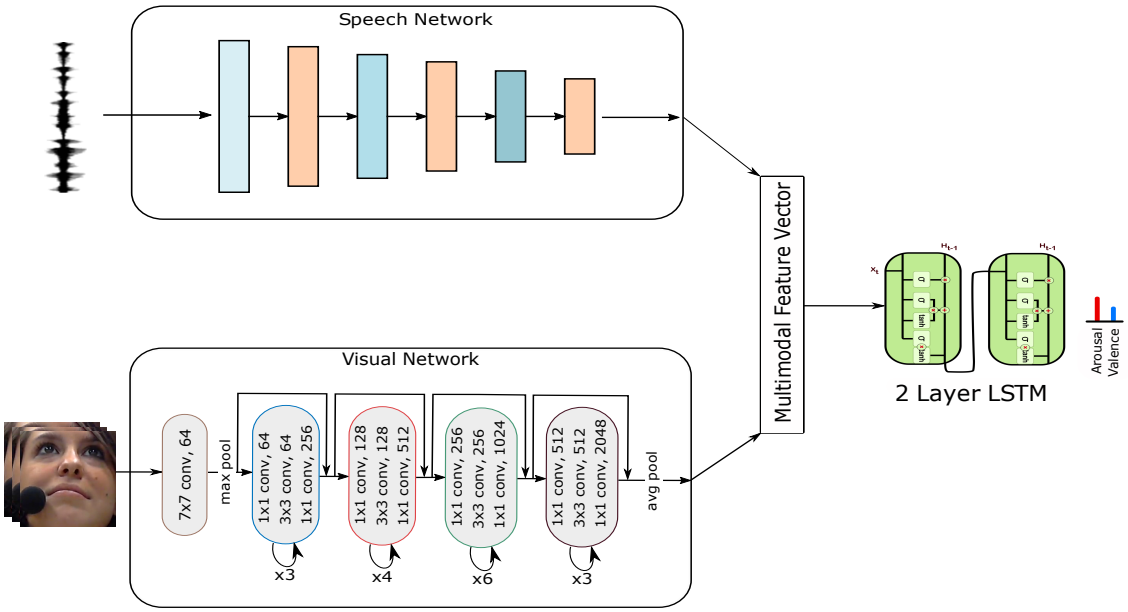


Figure 3.2: The network comprises of two parts: the multimodal feature extraction part and the RNN part. The multimodal part extracts features from raw speech and visual signals. The extracted features are concatenated and used to feed a 2-layer LSTM network. These are used to capture the temporal information in the data.

3.3 Experiments & Results

3.3.1 Database

To show the beneficial properties of our approach we utilised the REmote COLlaborative and Affective (RECOLA) database, introduced by Ringeval et al. [117], for time-continuous prediction of spontaneous and natural emotions (arousal and valence). Four modalities are included in the corpus: audio, visual, electro-cardiogram (ECG) and electro-dermal activity (EDA). The emotions investigated in this database are spontaneous and natural. In total, 9.5 h of multimodal recordings from 46 French-speaking participants were recorded and annotated for 5 minutes each, performing a collaboration task in dyads during a video conference. Among the participants, 17 were French, three German and three Italian. The database is split into three partitions – train (16 subjects), validation (15 subjects) and test (15 subjects)

– by stratifying (i. e., balancing) the gender and the age of the speakers. Finally, 6 French-speaking annotators (three male, three female) annotated all the recordings. Table 3.3 contains more details for each split of the database. The original labels of the RECOLA are re-sampled at constant frame rate of 40 ms. The data is then averaged over all raters by considering the inter-evaluator agreement, to provide a 'gold standard'. Each record is 300 s audio data with a sampling rate of 16 kHz.

Feature	**Train**	**Valid**	**Test**
Female/Male	10/6	9/6	8/7
French	11	11	11
Italian	3	2	3
German	2	1	1
Portuguese	0	1	0
age $\mu(\sigma)$	22.3(3.4)	21.6(2.1)	21.2(2.0)

Table 3.3: The partitioning of the RECOLA database.

3.3.2 Experimental Setup

For training the models, we utilise the Adam optimisation method [81], and a fixed learning rate of 10^{-4} throughout all experiments. For the audio models, we use a mini-batch of 25 samples. In addition, for regularisation of the network, we utilise dropout [141] with $p = 0.5$ for all layers except the recurrent ones. This step is important, as our models have a large amount of parameters and not regularising the network makes it prone on overfitting on the training data.

For the video model, the image size used is 96×96 with a mini-batch of size 3. A small mini-batch is selected because of hardware limitations. The data were augmented by resizing the image to size 110×110 and randomly cropping it to equal its original size. This produces a scale invariant model. In addition, colour augmentation is used by introducing random brightness and saturation to the image.

3.3.3 Training Process

Before training the multimodal network, each modality-specific network is trained separately to speed up the training procedure. Details for each network's training are found below.

Visual Network. For the visual network, we chose to fine-tune the pre-trained ResNet-50 on the database used in this work. This model was trained on the ImageNet [118] classification dataset that consists of 1 000 classes. The pretrained model was preferred over training the network from scratch in order to be benefited by the features already learned by the model. To train the network, a 2-layer LSTM with 256 cells each is stack on top of it to capture temporal dynamics in the signal.

Speech Networks. The CNN networks operate on the raw signal to extract features relevant to the task at-hand. In order to consider the temporal structure of speech, we use 2-layer LSTM with 256 cells each on top of the CNN.

Multimodal Network. After training the visual and speech networks, the LSTM layers are discarded, and only the extracted features are considered. The speech networks extract 1 280-dimensional (Emo-16), and 256-dimensional (Emo-18), features while the visual network extracts 2 048-dimensional features. These are concatenated to form for the Emo-16 a 3328-dimensional feature vector, and for the Emo-18 a 2 304 feature vector, which are then fed to a 2-layer LSTM with 256 cells each. The weights of the LSTM layers are initialised following Glorot based initialisation [49], and the visual and speech networks are initialised utilising the weights of the unimodal models. Finally, the whole network is trained in an end-to-end manner.

The goal for each unimodal and the multimodal network is to minimise the average Concordance Correlation Coefficient (CCC):

$$\mathcal{L}_c = \frac{\mathcal{L}_c^a + \mathcal{L}_c^v}{2}, \tag{3.4}$$

where $\mathcal{L}_c^a$ and $\mathcal{L}_c^v$ are the CCC of the arousal and valence, respectively.

3.3.4 Ablation study

Due to memory and training instability concerns [25] it is not always optimal to use very large sequences in recurrent networks. The justification for this can be either the over-blowing of gradients or the very deep unrolled graph which makes training of such big networks harder.

In order to choose the best sequence length to feed our LSTM layers, we conducted experiments using sequence lengths of 75, 150, and 300 for both the speech and the visual models. We have not included higher sequence length for memory constraints from the visual modality. In addition, we show the models performance when they are trained using one emotion dimension as output, and using both of them. Table 3.4 shows the results on the development set.

For the visual network we expect to get the highest value in the valence dimension, while for the speech model in the arousal dimension. Results indicate that the best value for the speech and visual models is 150, either training the models with one or multiple outputs. We also observe that training the models with one emotion dimension provides slightly better results that training than training the model with both dimensions. However, this comes with a cost of more training time and resources as we need to train two networks separately for the same modality, and also we double the number of parameters. In addition, after performing the wilcoxon test, we concluded that the results are not statistically significant. To this end, for the rest of our experiments we choose to train our models with 150 sequence length, and both emotion dimensions.

Network	Sequence Length	Arousal	Valence	Both	
				Arousal	Valence
Visual network	75	.264	.513	.278	.501
	150	.299	.531	.296	.524
	300	.291	.519	.287	.517
Speech network (Emo-16)	75	.651	.312	.646	.305
	150	.676	.335	.672	.332
	300	.667	.321	.661	.327
Speech network (Emo-18)	75	.706	.385	.698	.373
	150	.731	.408	.723	.397
	300	.716	.394	.709	.385

Table 3.4: Results (in terms of CCC) on arousal and valence after 60 epochs when varying the sequence length for the speech and visual networks.

In order to compensate for the discontinuities that the LSTM introduces in the beginning of the processing of the sequence, a chain of post-processing is applied by a number of studies in the literature [116, 115, 56] to the predictions obtained on the development set: *(i)* median filtering (with the size of the window ranging from 0.4 s to 20 s) [116], *(ii)* centring (by computing the bias between the gold-standard and the prediction) [75], *(iii)* scaling (using the ratio of standard-deviation of the gold-standard and the prediction as scaling factor) [75] and *(iv)* time-shifting (by shifting the prediction forward in time with values ranging from 0.04 s to 10 s), to compensate for delays in the ratings [98]. Any of these post-processing steps is kept when an improvement is observed on the CCC of the validation set, and applied then with the same configuration on the test partition.

In our proposed approach, we study end-to-end learning, meaning from the raw input getting the final prediction without any intermediate steps. In order to compensate for the discontinuities introduced by the LSTM and free ourselves from using the

post-processing method, we evaluate the model using the whole input sequence, though it was trained on a smaller input sequence. We use the model that was trained with 150 input sequence length (Table 3.4) as suggested from our previous experiments. Table 3.5 depicts the results. As expected, evaluating the model in a smaller sequence produces the worse results. On the other hand, evaluating the model in the whole input sequence the results are improved significantly but not as much as using the post-processing approach which produces the best results. As we are studying end-to-end learning methodologies, in what follows, we do not use the post-processing method, rather we evaluate the results on the whole input sequence.

Network	**Eval Method**	**Arousal**	**Valence**
Visual network	NO	.296	.524
	PP	.371	.637
	WH	.364	.621
Speech network (Emo-16)	NO	.672	.332
	PP	.786	.428
	WH	.772	.407
Speech network (Emo-18)	NO	.723	.397
	PP	.815	.502
	WH	.808	.493

Table 3.5: Results (in terms of CCC) on arousal and valence for the speech and visual networks when no post-processing is introduced and the same sequence length as in training is used (**NO**), when post-processing is applied (**PP**), and when the whole sequence is evaluated at once without any intermediate processing methods (**WH**).

3.3.5 Speech Modality

We first compare our speech networks with traditional approaches, i. e., using hand-engineered features extracted from the raw waveform. Results obtained for the Emo-16 and Emo-18, using all 46 participants of the RECOLA database are shown in

Table 3.6. In all of the experiments, our models outperform the designed features in terms of CCC. One may note, however, that the eGeMAPS [42] feature set provides close performance on valence, which is much more difficult to predict from speech compared to arousal. Furthermore, we show that by incorporating CCC directly in the optimisation function of all networks allows us to optimise the models on the metric CCC on which we evaluate the models. This provides us with *i)* a more elegant way to optimise models, and *ii)* gives consistently better results across all test-runs as seen in Table 3.6.

Predictor	**Features**	**Arousal**	**Valence**
a. Mean squared error objective			
SVR	eGeMAPS	.318 (.489)	.169 (.210)
SVR	ComParE	.366 (.491)	.180 (.178)
BLSTM	eGeMAPS	.300 (.404)	.192 (.187)
BLSTM	ComParE	.132 (.221)	.117 (.152)
Proposed (Emo-16)	raw signal	.684 (.728)	.249 (.312)
Proposed (Emo-18)	raw signal	.754 (.786)	.334 (.396)
b. Concordance correlation coefficient objective			
BLSTM	eGeMAPS	.316 (.445)	.195 (.190)
BLSTM	ComParE	.382 (.478)	.187 (.246)
Proposed (Emo-16)	raw signal	**.702** (.786)	**.365** (.428)
Proposed (Emo-18)	raw signal	**.778** (.815)	**.502** (.493)

Table 3.6: RECOLA database results (in terms of CCC) for prediction of arousal and valence. In parenthesis are the performances obtained on the development set. In *a)* we optimised the models w. r. t. MSE whereas in *b)* w. r. t. CCC.

In addition, we compare the performance of our models with more recent approaches that existed in the literature. Most of them have been submitted to the AVEC 2016 challenge, with 27 participants (Table 3.7). For fair comparison, we test our model on the same test set. In case performance on the test or validation set was not reported

in the paper, a dash is inserted. Results show that our model outperforms the other models in the test set when predicting the arousal dimension. It is important to notice that although our model gets a lower CCC on the arousal dimension for the validation set compared to the baseline of the challenge, its performance is better on the test set, indicating that the tailored to the task at hand features learned by our models have high generalisable capabilities compared to the hand-crafted features of the other approaches.

Comparing our two proposed models, Emo-18 performs better in both the arousal and valence dimensions than Emo-16. We believe the extra convolution layer in the Emo-18 network provides higher level of abstractions, and as such the model is able to learn more complex functions. In addition, the Emo-16 network performs its second max-pooling along the channel dimension, which seems not to be able to preserve the necessary information extracted in the time-domain, as is the case for the Emo-18.

Predictor	**Features**	**Arousal**	**Valence**
Baseline [162]	eGeMAPS	.648 (.796)	.375 (.455)
RVM [70]	eGeMAPS	– (.750)	– (.396)
Povolny et al. [111]	Mixed	– (.833)	– (.503)
Brady et al. [18]	MFCC	– (.846)	– (.450)
Weber et al. [169]	eGeMAPS	– (.793)	– (.456)
Somandepalli et al. [139]	Mixed	– (.800)	– (.448)
Han et al. [55]	13 LLDs	.666 (.755)	.364 (.476)
Proposed (Emo-16)	raw signal	**.702** (.772)	**.365** (.397)
Proposed (Emo-18)	raw signal	**.778** (.808)	**.436** (.493)

Table 3.7: RECOLA database results (in terms of CCC) for the prediction of arousal and valence using the speech networks. In parenthesis are the performances obtained on the development set. A dash is inserted if the results were not reported in the original papers.

We now compare our best performing model (Emo-18) with (at the time) state-

Predictor	Arousal	Valence
Han et al. [57]	.729 (.785)	.309 (.364)
Han et al. [56]	.744 (.774)	.377 (.412)
Trigeorgis et al. [151]	.686 (.741)	.261 (.325)
Schmitt et al. [123]	.753 (.793)	.430 (.550)
Proposed (Emo-16)	**.702** (.772)	**.365** (.397)
Proposed (Emo-18)	**.778** (.808)	**.436** (.493)

Table 3.8: Performance comparison (w. r. t. CCC) between the proposed method and other state-of-the-art methods. In parenthesis are the performance obtained on the validation set.

of-the-art approaches. The results are depicted in Table 3.8. The proposed Emo-18 network outperforms the other approaches in all cases on the test set for both arousal and valence, indicating the effectiveness of our model. Particularly, our model achieves .778 of CCC for arousal and at .429 of CCC for valence. We can observe that the improvement on the arousal is higher than that on valence. This was expected as it has been shown in the literature [115, 56] that the audio models can more easily predict the arousal dimension.

In more detail, all studies, except the Trigeoris et al. [151], use hand-crafted features in their models. We believe this is the main reason our model outperforms these approaches. Although, Trigeoris et al. [151] model uses the raw waveform as input, the architecture is not well designed as the proposed one, nor is as deep making it perform worse even when compared with the other methods that use hand-crafted features.

3.3.6 Visual Modality

The visual modality has been shown to be efficient in predicting the valence dimension rather than the arousal [116, 115].Table 3.9 presents the best results for

the valence dimension submitted to the AVEC 2016 challenge, along with the work from Han et al. [55] which was not submitted to the challenge. The features used for all of the models are appearance and geometric. For appearance, Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) features were extracted, whereas facial landmarks were extracted for the geometric features. The input to our network are the raw pixel intensities from the face extracted from the frames of the videos using the Multi-Domain Convolutional Neural Network Tracker (MDNet) [103] tracking algorithm. This algorithm takes the bounding box of the face in the first frame of the video and tracks it in all frames.

Predictor	**Features**	**Arousal**	**Valence**
Baseline [162]	Appearance	.343 (.483)	.486 (.474)
Baseline [162]	Geometric	.272 (.379)	.507 (.612)
RVM [70]	Geometric	– (.467)	– (.571)
RVM [70]	Appearance	– (.615)	– (.530)
Video CNN-L4 [111]	Mixed	– (.595)	– (.497)
Brady et al [18]	Appearance	– (.346)	– (.511)
Weber et al. [169]	Geometric	– (.476)	– (.683)
Weber et al. [169]	Appearance	– (.594)	– (.506)
Somandepalli et al. [139]	Geometric	– (.297)	– (.612)
Somandepalli et al. [139]	Appearance	– (.481)	– (.474)
Han et al. [55]	Mixed	.265 (.292)	.394 (.592)
Proposed	raw signal	**.425** (.364)	**.608** (.621)

Table 3.9: RECOLA database results (in terms of CCC) for prediction of arousal and valence using the Visual Network. In parenthesis are the performances obtained on the development set. A dash is inserted if the results were not reported in the original papers.

From Table 3.9 we observe that the visual modality benefits the models in the valence dimension rather than the arousal as was expected [116, 115]. Our model provides the highest performing model on the test set compared with all the other

studies, on both the arousal and valence dimensions. This is due to the fact all other approaches utilise hand-crafted features such as histogram of gradients (HOG) and landmarks. On the other hand, ResNet-50 can exploit its deep architecture and learn complex functions tailored to the task at hand.

3.3.7 Multimodal Analysis

We show the results for the multimodal model and compare with the (at the time) state-of-the-art results. Only two other models are found in the literature to use both speech and visual modalities on the full RECOLA database. These are the Output-Associative Relevance Vector Machine Staircase Regression (OA RVM-SR) [70] and the strength modelling system proposed by Han et al. [55]. To have a fair comparison with the other methods, we utilised both training and validation sets in the training process. We stopped the training of our model when we reached twice the number of epochs required to train the model when only the training set was used. We did this, as the number of both validation and training examples is twice the number of training examples.

Results are shown in Table 3.10. Our model outperforms the other two models in both the arousal and valence dimensions; especially for the valence dimension with high magnitude. Thus, we can conclude that the geometric/appearance features and mixtures of visual features cannot capture as adequately the emotional response of individuals as the proposed system. We should also mention that our system operates directly on the raw pixel domain, while the other two systems made use of a number of geometric features (e. g., 2D/3D facial landmarks etc.) which require the presence of an accurate facial landmark tracking methodology (ours was applied on the results of a conventional face detector only).

To show, however, that our model can also benefit from these features, we incorporate them into our model. In particular, we first extract facial landmarks using the

face alignment method by Deng et al [29], and perform Procrustes alignment (i. e., removing scale, rotation, translation). Then, we concatenate these features to the feature vector that is input to the recurrent model and we train only the recurrent network while keeping the speech and visual network parameters fixed. Results are depicted in Table 3.10. As expected, these features benefit our model to predict the valence dimension while the arousal is unchanged.

For the arousal dimension we can observe the high generalisable properties of our model compared to the other two; the test results are higher than the validation one. Thus, we can see that for the task-at-hand the learned characteristic of the neural network are more appropriate than well-established hand-crafted features such as LLDs or MFCCs.

Comparing the performance of the multimodal network with the unimodal ones, we can see that the valence dimension is more benefited rather than the arousal one. We believe this is because of the larger number of features extracted by the visual network (2048) compared to the speech network (1280 for Emo16 and 1024 for Emo18), and hence the multimodal information is governed (mostly) by the visual modality.

3.4 Conclusions

In this chapter, we proposed a multimodal system that operates on the raw speech and visual signals that perform end-to-end learning for the spontaneous emotion prediction task. On top of that, we proposed a novel speech neural network architecture. Our models take into consideration the temporal information on the signal with the use of recurrent networks (LSTM). To speed up the training of the multimodal model, we first pretrained the speech and visual networks, separately. For our experiments we used the RECOLA database that is a large-scale database and can allow our deep model to learn complex functions. First, our ablation study showed

Predictor	Audio Features	Visual Features	Arousal	Valence
OA RVM-SR [106]	eGeMAPS ComParE	Geometric Appearance	.770 (.855)	.545 (.642)
Han et al. [56]	13 LLDs	Mixed	.610 (.728)	.463 (.544)
Proposed (Emo-16)	raw signal	raw signal	.776 (.718)	.679 (.483)
Proposed (Emo-16)	raw signal	raw & geometric	.776 (.721)	.717 (.503)
Proposed (Emo-18)	raw signal	raw signal	**.788** (.736)	**.707** (.526)
Proposed (Emo-18)	raw signal	raw & geometric	**.789** (.736)	**.746** (.546)

Table 3.10: RECOLA database results (in terms of CCC) for the prediction of arousal and valence using the proposed multimodal emotion recognition network. In parenthesis are the performance obtained on the development set.

that training a model to predict one emotion dimensions performs slightly better than predicting both simultaneously without statistically difference. Thereafter, we observed that the unimodal models achieve remarkably better performance on the test set in comparison to other at the time state-of-the-art models. In addition, our multimodal model greatly outperforms in both the valence and arousal dimensions the other models.

Although our results outperformed the at-the-time state-of-the-art it has a number of limitations. First, we should point out that our multimodal model consists of simply concatenating the audio and visual features, before passing them to the RNN, making the model prone to misuse of information if for example a sensor fails. Second, the models were trained in controlled conditions where the face is visible and without any noisy conditions both for audio and visual modalities. We try to compenstate from these limitations in the next chapter. In particular, we try to better exploit multiple input signals and make the multimodal model more robust under naturalistic conditions we propose different attention-based methods.

Chapter 4

Multimodal Affect Recognition in Naturalistic Conditions

Contents

3.1 Motivation 45

3.2 Proposed Method 49

3.3 Experiments & Results 55

3.4 Conclusions 66

4.1 Motivation

In the previous chapter, we discussed the laboratory (or controlled) emotion recognition setting using audiovisual information. Although the affective computing community had wildly focused on laboratory settings for over twenty years mainly due to its simplicity, the following drawbacks are posed:

- **Enviromental restrictions**. They contain controlled visual and speech conditions (e. g. no audiovisual noise, predefined number of lights, and illumination).

- **Human-level restrictions.** They restrict the subject in many aspects including but not limited to body placement, head movement, and the level of expression is allowed.

The aforementioned limitations however do not apply to naturalistic conditions. From a subject-wise point of view, the subject can freely make any desired head or body movements, and perform any kind of expressions, and hence show all possible emotional states. From an environmental point of view, any kind of occlusions may exist (e. g. a hand occluding the face of the subject). In addition, non-uniform and noisy conditions are entailed (e. g. lighting and audio recording).

In this chapter, we study the automatic continuous emotion recognition task in naturalistic conditions. In particular, we develop a multimodal system that utilises text, audio and visual information in an *end-to-end* manner. We propose a transformer-based [163] architecture to extract information from the text modality with the word representation network to be chosen based on an ablation study we performed. Thereafter, the architecture of the network is comprised of position-wise convolution to extract N-grams, with a transformer (multi-linear projection) with RNN architecture to extract multiple views from the sentence. The extracted information is passed to an attention mechanism to extract a weighted average feature representation.

Our audio and visual networks are selected from the current state-of-the-art models in the affective computing literature after performing an ablation study (Sec. 4.4.4). The general structure of the networks is comprised from a spatial feature extractor for per-frame feature representation, and a temporal feature extractor to capture the temporal dynamics in the signals. Our audio network is inspired by an earlier publication [161], and as visual network we use a variant of a high resolution network (HRNet) architecture [62]. The variation in the network stems from the attention mechanism we are using instead of the last average pooling layer, and show that the

model's performance can be benefited.

After the extraction of the unimodal features, we propose attention-based fusion strategies (feature-level fusion) (Sec. 2.7.1) to fuse the modalities together, before passing them to a RNN network that captures the temporal characteristic in the signals. Using attention-based methods to fuse the modalities has two beneficial effects: (i) improves the generalisation capability of the model, and (ii) allows us to visualise the findings, such as identifying the modality the models attends to at a particular frame.

The contributions of this chapter can be summarised as follows:

- develop a transformer-based architecture in our text model and capture the semantics of a sentence by extracting context-aware features. To the best of our knowledge, this is the first time a transformer mechanism has been used in the sentiment analysis domain.
- propose novel attention-based fusion strategies to combine the features extracted from the raw speech, text and visual modalities.
- visualise (i) the attention mechanism in our text model and indicate the words that contribute the most to the final prediction, and (ii) the attention-based fusion strategies that indicate which modality has the highest contribution per frame.
- provide state-of-the-art results for the visual, text and multimodal modalities, and the second highest performing model in the audio modality, when comparing with the winning papers from the challenge that use several hand-crafted and deep features. On top of that, we perform an extensive experimentation of current state-of-the-art unimodal architectures in visual, and audio modalities.

The content of this chapter is based around the following publication:

- **Panagiotis Tzirakis***, Jiaxin Chen*, Stefanos Zafeiriou, Björn Schuller, *End-to-end multimodal affect recognition in real-world environments*, Information Fusion, Volume 68, pp. 46-53, 2021.

The rest of the chapter is organised as follows. After we introduce our notation in Sec. 4.2, we present the proposed approaches, namely, unimodal and multimodal networks with the attention-based fusion strategies, in Sec. 4.3. Finally, experimental studies and results are presented and discussed in Sec. 4.4, along with visualisation of the proposed attention methods, before concluding this chapter in Sec. 4.5.

4.2 Notation

Before describing our model, we define our notation. Matrices are defined as uppercase bold letters as $\mathbf{X}$, vectors as lowercase bold letters as $\mathbf{z}$, and scalars as non-bold letters. The i-th row of a matrix is defined as $\mathbf{X}[i,:]$, the j-th column as $\mathbf{X}[:,j]$, and the element at position $[i,j]$ as $\mathbf{X}[i,j]$. The k-th element of a vector is specified as $\mathbf{z}[k]$. The row-wise concatenation of the vectors $\{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m\}$ is denoted as $[\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_m]$.

4.3 Proposed Method

We introduce a multimodal system that is comprised of three DNNs utilised to extract (spatial) features from the raw text, audio and visual cues. The features are combined to form a unified representation using attention mechanism. Temporal information is vital in our task and as such we utilise LSTM networks to compute temporal patterns in the extracted features, before the final prediction.

4.3.1 Unimodal Networks

Text Network

Our proposed text network is depicted in Fig. 4.1 and its comprised of six different parts, which are summarised below, along with the input representation.

Input. The most fundamental step in text processing is the choice of the input representation. For our word-level representation, we adopt the pre-trained XLNet [176] of 768 dimensions with enriched subword information such that a sentence is represented as the matrix $\mathbf{S} \in \mathbb{R}^{300 \times L}$, where L is the maximum words in a sentence of our corpus.

Position-wise N-Grams. We use a CNN to capture position-wise N-grams of the input matrix $\mathbf{S}$ for each frame. In particular, each word vector $\mathbf{S}[:, l]$ is processed by several convolutional layers, and depending on their filter width, i.e. $n = 2, 3, 4, ..., N$, they are used to capture N-grams producing a feature map as follows:

$$\mathbf{f}[l] = g(\tilde{\mathbf{S}}[:, l : l + n - 1] \circ \mathbf{H} + b) \tag{4.1}$$

where $\tilde{\mathbf{S}}$ is the matrix $\mathbf{S}$ padded, $g(\cdot)$ is an activation function (in our case $tanh$), n is the width of the filter $\mathbf{H} \in \mathbb{R}^{300 \times n}$, and $\circ$ denotes the hadamard multiplication. While in standard approaches only the extracted N-gram features are processed, in our model, we propose to augment the embedded matrix $\mathbf{S}$ by concatenating to it the extracted N_k feature maps to form a new matrix $\mathbf{C} = [\mathbf{S}, \mathbf{f}^1, ..., \mathbf{f}^{N_k}] \in \mathbb{R}^{D_e \times L}$, where $D_e = 300 + N_k$. We perform this operation because the word-embedding matrix was computed using a pre-trained network and as such valuable information is contained for each word.

Multihead Linear Projection. To improve the diversity of the input features and

inspired by [163], we propose to use an attention mechanism to enable each frame to discover multiple representative features in different context spaces. More particularly, given the input matrix $\mathbf{C}$, we first apply different linear projections on $\mathbf{C}$ to explore different perspective of context for the current frame, i. e., $\mathbf{V}_p = \mathbf{W}_p\mathbf{C}$, where $\mathbf{W}_p \in \mathbb{R}^{D_p \times D_e}$ is the p^{th} linear projection matrix and D_p is the dimension of the p^{th} context space.

Recurrent Context Generator. To capture semantic information in the different context spaces we utilise a 2-layer Bi-LSTM. Specifically, the sequential features of each position $\mathbf{V}_p[:, l]$ are processed as follows:

$$\begin{aligned} \mathbf{h}_p^{'l} &= Bi - LSTM(\mathbf{h}_p^{'l-1}, \mathbf{h}_p^{'l+1}, \mathbf{V}_p[l, :]) \\ \mathbf{h}_p^{l} &= Bi - LSTM(\mathbf{h}_p^{l-1}, \mathbf{h}_p^{l+1}, \mathbf{h}_p^{'l}) \end{aligned} \tag{4.2}$$

where $\mathbf{h}_p^l \in \mathbb{R}^{D_l}$ and $\mathbf{h}_p^{'l} \in \mathbb{R}^{D_l}$ are the context vectors at position l in the p^{th} context space and D_l is the output dimension. At position one $\mathbf{h}_p^0$ and $\mathbf{h}_p^{'0}$ are initialised to zero vectors.

Selector. The semantic-aware hidden states $\mathbf{h}_p^l$ of the Bi-LSTM are processed by an attention mechanism to enable the model to *select* the most useful context-aware features in each context space. The attention mechanism is applied in a scaled dot product way [163] as follows:

$$\begin{aligned} e_p^l &= \frac{\exp((\mathbf{h}_p^l)^T \mathbf{q}_p)}{\sum_{k=1}^{L} \exp((\mathbf{h}_p^k)^T \mathbf{q}_p)} \\ \alpha_p^l &= softmax(e_p^l / \sqrt{D_l}) \\ \mathbf{s}_p &= \sum_{l=1}^{L} \alpha_p^l \mathbf{h}_p^l \end{aligned} \tag{4.3}$$

where $\mathbf{q}_p \in \mathbb{R}^{D_v}$ is a learnable vector, α_p^l is our attention score on the context-aware

hidden states $\mathbf{h}_p^l$ at position l. The scaling term $\sqrt{D_l}$ allows more stable training by preventing peaked softmax issue as well as smaller gradient issue [163]. The final attentional feature in the p^{th} context space is obtained by $\mathbf{s}_p$.

Context Feature Concatenation. In this step we concatenate the context-aware features obtained in each projection space to form a single feature representation, i.e. $\mathbf{z} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_P] \in \mathbb{R}^{PD_l}$ where m is the total number of context space projections.

Highway Networks. Highway networks [140] are proven to have a strong empirical performance especially for deep networks. To this end we adopt highway networks by feeding them our frame-level feature $\mathbf{z}$, i.e.,

$$\tilde{\mathbf{z}} = \mathbf{o} \circ g(\mathbf{W}^z \mathbf{z} + b^z) + (\mathbf{1} - \mathbf{o}) \circ \mathbf{z} \tag{4.4}$$

where $\mathbf{o} = \sigma(\mathbf{W}^o \mathbf{z} + b^o)$ acts as a transformation gate which allows shortcut from input to output directly, and $g(\cdot)$ is an activation function (in our case *tanh*).

Audio Network

An important step in the field of paralinuistics is extracting features from the raw waveform. For our purposes, we performed an ablation study (see Sec. 4.3 to find the best audio feature extractor for continuous emotion recognition with three recently proposed architectures : (i) Trigeorgis et al. [151], (ii) Tzirakis et al. [161], and (iii) Li et al. [91].

The input to the network is the raw waveform, segmented into 160 s long sequences. At a sampling rate of 22050 Hz this corresponds to a $3,528,000$-dimensional vector. In addition, we utilise the interlocutor information provided with the SEWA database by concatenating to the initial vector an additional $3,528,000$-dimensional vector in each audio frame, similar to Huang et al. [68] - thus, the final input to the

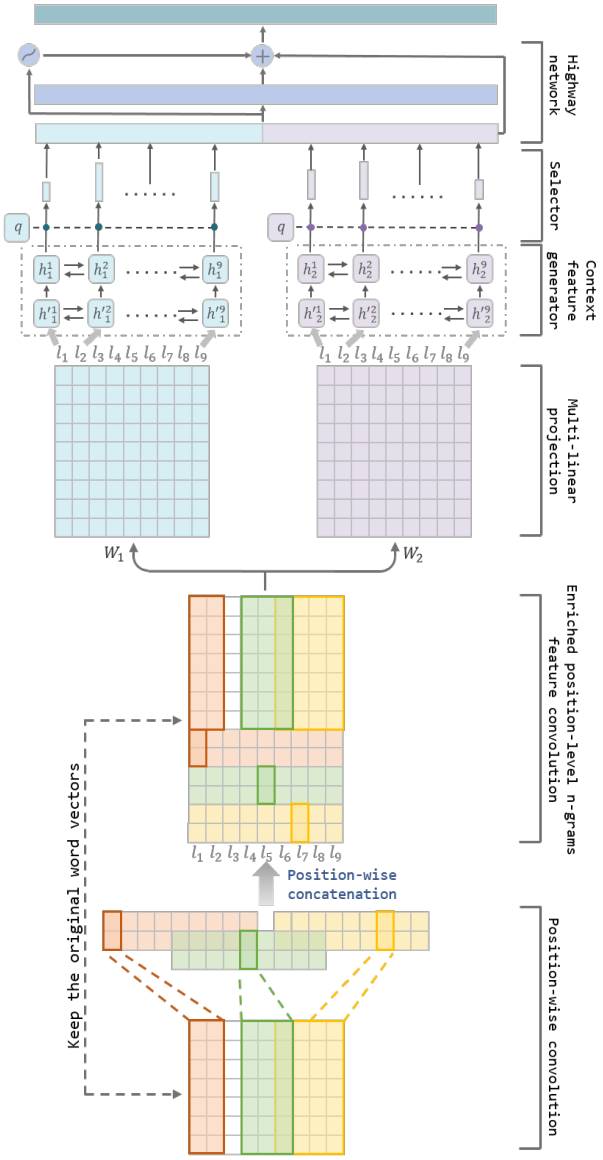


Figure 4.1: Text network architecture. The input is a word-embedding matrix that is passed to a N-gram feature extractor (padding is not shown), before being concatenated with the N-gram features. Then two linear projections are applied such that different contexts are extracted. Each context-matrix is processed by a 2-layer Bi-LSTM to get the semantics in the sentence before the selector that weights the outputs. Finally, the outputs of each recurrent network is concatenated and passed by a highway network to get the semantic-aware feature vector. (Best viewed in colour)

network is a $(2 * 3,528,000)$-dimensional vector. To use the interlocutor information, and be able to distinguish the target from the chatting partner in the contained dyadic conversations, in each audio frame half of the entries of the input vector are zeros depending on the speaker turn, and the other half is filled with the extracted features of the speaking subject.

Visual Network

To find the best visual feature extractor for the task at hand, we experiment with five network architectures: (i) deep residual network (ResNet) of 50 layers [62], (ii) densely connected convolutional network (DenseNet) [66], (iii) VGG Face [107], (iv) MobileNet [64], and (v) HRNet [165]. The choice for using these networks stems from the fact that are the most popular and best performing image recognition models.

Studies in the literature have shown the beneficial properties of using an attentional pooling layer [48] instead of the last average one in deep convolution networks. To this end, we adopt an hierarchical attention scheme, where we replace the last average pooling layer with an attention mechanism following the low-rank second-order pooling (top-down attention) [48]. On top of that, we propose to use another attention between the feature maps. Utilising this attention scheme speeds up the training of the network and increases its generalisable capabilities. Fig 4.2 shows the proposed pooling layer.

4.3.2 Attentional Fusion Strategies

To take advantage of the information extracted from text ($\mathbf{x}^t \in \mathbb{R}^{D_t}$), audio ($\mathbf{x}^a \in \mathbb{R}^{D_a}$) and visual ($\mathbf{x}^v \in \mathbb{R}^{D_v}$) modalities, we propose four different attention-based fusion strategies and compare them to the simple concatenation approach used in the previous chapter.

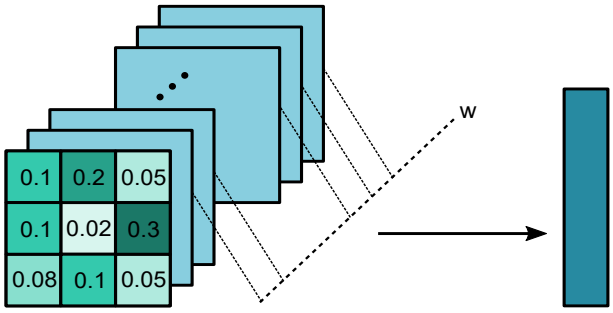


Figure 4.2: Attentional pooling layer. First, an attentional feature map is applied before taking a weighted average of these maps to compute the final feature representation.

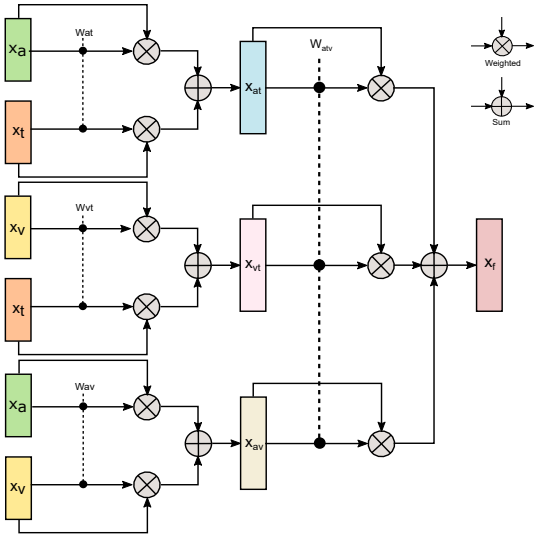


Figure 4.3: Hierarchical attention fusion methodology. Each modality features are combined using a weighted average of them, and the corresponding outputs are again combined in the same manner.

Concatenation

The first approach is a simple concatenation of the unimodal features, i. e., $\mathbf{x}^{fusion} = [\mathbf{x}^t, \mathbf{x}^a, \mathbf{x}^v]$. This is a standard feature-level fusion approach that has been used extensively in the literature.

Hierarchical Attention

We propose to perform hierarchical attention to the unimodal features so as to maximise the relevant information that is propagated forward to the network. More

particularly, we first perform attention to all paired unimodal features, i.e. audio-text, audio-visual, and visual-text, before using a higher level attention to the paired outputs. Fig 4.3 depicts the architecture of the proposed approach. Formally, given the features extracted from text $\mathbf{x}^t$, audio $\mathbf{x}^a$ and visual $\mathbf{x}^v$ modalities, we first perform a linear projection to each of these features such that they are projected to the same vector space (with the same number of dimensions D_u), namely,

$$\begin{aligned} \tilde{\mathbf{x}}^t &= \mathbf{W}^t\mathbf{x}^t + b^t \\ \tilde{\mathbf{x}}^a &= \mathbf{W}^a\mathbf{x}^a + b^a \\ \tilde{\mathbf{x}}^v &= \mathbf{W}^v\mathbf{x}^v + b^v \end{aligned} \tag{4.5}$$

where $\mathbf{W}^t \in \mathbb{R}^{D_u \times D_t}$, $\mathbf{W}^a \in \mathbb{R}^{D_u \times D_a}$, $\mathbf{W}^v \in \mathbb{R}^{D_u \times D_v}$ are projection matrices for text, audio and visual modalities, respectively, with dimension D_u. The projected features are combined pair-wise, and then are passed through a selective attention mechanism to obtain the final fusion feature, i.e.,

$$\begin{aligned} \alpha^i &= softmax(\frac{\tilde{\mathbf{x}}^i\mathbf{q}^i}{\sqrt{D_u}}) \\ Attention(\mathbf{x}^1 \cdots \mathbf{x}^M) &= \sum_{i=1}^{M} \alpha^i \tilde{\mathbf{x}}^i \end{aligned} \tag{4.6}$$

where $\mathbf{q}^i \in \mathbb{R}^{D_u}$ is a learnable vector that attends to different modality features, and $i \in [1, M]$ is an index denoting the M-th modality. We obtain final attentive fusion feature $\mathbf{x}^{fusion}$ as follows:

$$\begin{aligned} \mathbf{x}^{vt} &= Attention(\tilde{\mathbf{x}}^v, \tilde{\mathbf{x}}^t) \\ \mathbf{x}^{va} &= Attention(\tilde{\mathbf{x}}^v, \tilde{\mathbf{x}}^a) \\ \mathbf{x}^{ta} &= Attention(\tilde{\mathbf{x}}^t, \tilde{\mathbf{x}}^a) \\ \mathbf{x}^{fusion} &= Attention(\mathbf{x}^{vt}, \mathbf{x}^{va}, \mathbf{x}^{ta}) \end{aligned} \tag{4.7}$$

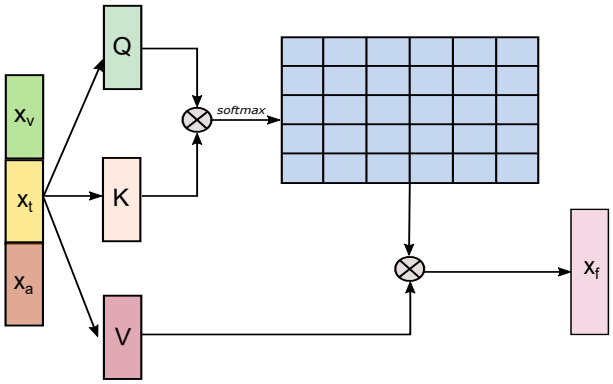


Figure 4.4: Self-attention fusion methodology. All modality-specific features are concatenated and passed through a self-attention layer to combine them in the final fusion vector.

Self-Attention

We also propose to apply self-attention fusion strategy, shown in Fig. 4.4, by enabling the extracted features to attend to each other. A similar approach (transformer-based) was proposed by Tsai et al. [153] but our method is simpler with less parameters. We start by first applying a linear projection on the unimodal features $\mathbf{x}^t$, $\mathbf{x}^a$ and $\mathbf{x}^v$ like in Eq. 4.5.

After we concatenate the three vectors together, i. e., $\mathbf{X}^c = [\tilde{\mathbf{x}}^t, \tilde{\mathbf{x}}^a, \tilde{\mathbf{x}}^v]^T \in \mathbb{R}^{3\times D_u}$, we apply a multi-head 3-way linear projection of the matrix [163], i. e., $\mathbf{Z}_i^j = \mathbf{X}^c\mathbf{W}_i^j \in \mathbb{R}^{3\times D_s}$ for $j = 1, 2, 3$, $i \in \mathbb{N}^+$ where $\mathbf{W}_i^j \in \mathbb{R}^{D_u\times D_s}$, D_s is the dimension of each projection space, and i is an index for operations in different projection space. Then, we utilise self-attention to explore complementary relationships among the different modalities. The attention is applied as follows:

$$
\begin{aligned}
\boldsymbol{A}_i &= softmax(\frac{\mathbf{Z}_i^1(\mathbf{Z}_i^2)^T}{\sqrt{D_s}})\mathbf{Z}_i^3 \\
\boldsymbol{\alpha}_i &= \boldsymbol{A}_i\mathbf{q}_i \\
\mathbf{v}_i &= \sum_{k=1}^{3} \boldsymbol{\alpha}_i[k]\boldsymbol{A}_i[k,:]
\end{aligned}
\tag{4.8}
$$

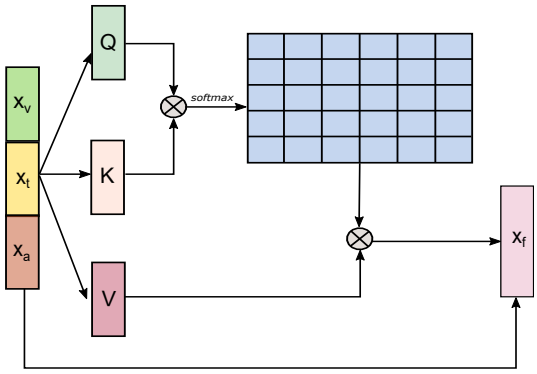


Figure 4.5: Residual self-attention fusion methodology. All modality-specific features are concatenated and passed through a self-attention layer, and a residual connection from the input to the output is added to combine them in the final fusion vector.

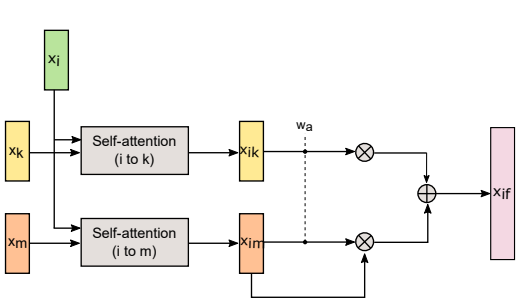
where $\mathbf{q}_i \in D_s$ here is again a learnable variable for self-attentive features and the subscript i denotes the i^{th} linear projection. The final fusion feature vector is the concatenation of all the projected features, i.e., $\mathbf{x}^{fusion} = [\mathbf{v}_1, ..., \mathbf{v}_p]$ where p is the number of projections applied.

Residual Self-Attention

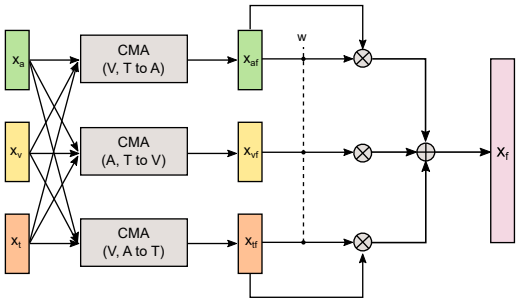
Inspired by the work on non-local neural networks [166] we extend the self-attention fusion strategy by incorporating a residual connection (see Fig. 4.5). To this end, the residual self-attention can be defined as follows:

$$\mathbf{z}_i = \mathbf{W}_z \mathbf{v}_i + \mathbf{X}^c, \tag{4.9}$$

where i represents the i-th linear projection (see Eq. 4.8), "$+\mathbf{X}^c$" denotes the residual connection, and $\mathbf{v}_i$ is defined in Eq. 4.8. The final fusion representation is computed by concatenating the projected features $\mathbf{z}_i$, i. e., $\mathbf{x}^{fusion} = [\mathbf{z}_1, ..., \mathbf{z}_p]$ where p is the number of projections applied.



(a) Cross-modal attention (CMA) method. Modality i is combined with the other other two modalities k, and m using self-attention, and a weighted average of the output features.



(b) Cross-modal self-attention fusion architecture. Audio, visual and text modalities are combined using CMA method, and a weighted average of the output features.

Figure 4.6: Showing the (a) Cross-modal attention (CMA) method, and (b) Cross-modal self-attention fusion architecture.

Cross-modal Hierarchical Self-Attention

We also propose a cross-modal hierarchical self-attention fusion, where each modality is combined with the rest using self-attention, and then the features per modality are fused in an attentional manner. Fig. 4.6 depicts the architecture. Mathematically, we define the cross-modal attention (CMA) (shown in Fig. 4.6a) as follows:

$$\mathbf{a}^{m,c} = \tilde{\mathbf{x}_1}^{m} softmax(\frac{\tilde{\mathbf{x}_2}^{m}(\tilde{\mathbf{x}}_1^{c})^T}{\sqrt{D_s}}), \tag{4.10}$$

where the modality m is fused with modality c, $\tilde{\mathbf{x}}_i^m$ is the i-th linear projection with dimensions D_s of modality $\mathbf{x}_i^m$, as defined in Eq. 4.5. The cross-modal attentional representation of modality m with the rest modalities c_i is defined as $\mathbf{v}_f^m = Attention(\mathbf{a}^{m,c_1}, \mathbf{a}^{m,c_2}, ...\mathbf{a}^{m,c_i})$. Finally, we fuse the M (one per modality) cross-modal representations using another attention, i. e.,

$$x^{fusion} = Attention(\mathbf{v}_f^{m_1}, \mathbf{v}_f^{m_2}, ..., \mathbf{v}_f^{m_M}). \tag{4.11}$$

4.3.3 Multimodal Model

Our multimodal model is an aggregation of four steps: (i) modality-specific feature extraction, where we use the unimodal networks to extract features from the raw signal, (ii) fusion strategies, where we use the proposed attention-based fusion methods to combine all modality-specific features and extract a single representation vector, (iii) a LSTM network to capture the temporal dynamics in the signal, and (iv) the prediction layer that performs the final prediction of the model. Fig. 4.7 depicts the overall architecture.

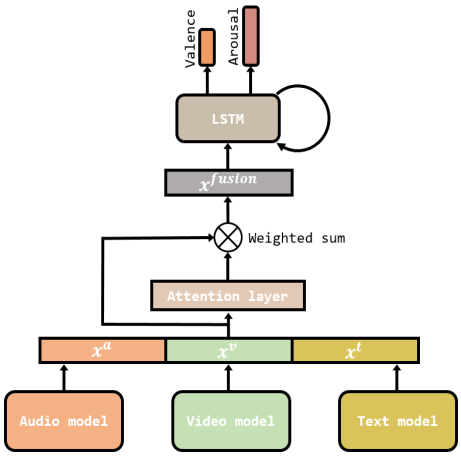


Figure 4.7: The multimodal network is comprised of the features extracted from the text ($\mathbf{x}^t$), audio ($\mathbf{x}^a$), and visual ($\mathbf{x}^v$) modalities, the attentional layer that fuses the extracted features ($\mathbf{x}^{fusion}$) using a weighted vector, and 1-layer LSTM that capture the temporal dynamics in the data before the final prediction. (Best viewed in colour)

4.4 Experiments & Results

4.4.1 Database

To show the beneficial properties of the proposed model we use a recently proposed database that focuses on multiple cultures and dyadic conversations: the Sentiment Analysis in the Wild (SEWA) [1]. In total, six cultures across Europe and

[1] https://sewaproject.eu/

Chinese are contained. It consists of audio-visual recordings that were collected from web-cameras and microphones and captured spontaneous and naturalistic behaviour (arousal and valence), i. e., under naturalistic conditions. In total, 64 subjects (age from 18 to 60 years old) are paired (i. e., 32 pairs) to watch a 90 second commercial video, and their task was to discuss the content with their partner for maximum 3 minutes. The database provides three modalities, namely, text, audio and visual, and it is split into 3 partitions: training (17 pairs), development (7 pairs), and test (8 pairs). In total, the training, development and test sets contain 54 813, 22 556, and 27 950 audiovisual frames, and 1 329, 604, and 629 sentences, respectively. Finally, 6 German-speaking annotators (3 female, 3 male) were used to annotate the database in continuous arousal and valence. Table 4.1 demonstrates a summary of the database.

Culture	**#Conversatinons**	**#Subjects**	**Durantion (min)**
Chinese	35	70	101
Hungarian	33	66	67
German	32	64	89
British	33	66	94
Serbian	36	72	98
Greek	28	56	81

Table 4.1: Summary of the SEWA database. Number of cultures, conversations, subjects and their corresponding duration.

4.4.2 Training Process

The first step in our training process is to train individually each unimodal network. To this end, we stack on top of each network a 1-layer LSTM of 256 hidden units before the predictions of the arousal and valence.

The weights of each trained unimodal network are used to initialise the weights in

Hyper-parameter	**Range**
Learning rate	$[10^{-4} - 10^{-6}]$
Batch size (Text/Audio)	$[10, 20, 30]$
Batch size (Visual/Multimodal)	$[2, 3]$
Dropout	$[0.1 - 0.6]$
Sequence length	$[75, 100, 200]$
Hidden units (LSTM)	$[64, 128, 256, 512]$
Num layers (LSTM)	$[1, 2, 3]$

Table 4.2: Range of values for hyper-parameter tuning.

the multimodal network. As in the unimodal cases, we use 1-layer LSTM of 256 hidden units to capture the temporal dynamics in the data, which was initialised by the Glorot based method [49].

4.4.3 Experimental Setup

Our networks have a number of parameters that need to be tuned. For our purposes, we consider tuning the following hyper-parameters: learning rate, batch size, dropout, and number of hidden units and number of layers for the LSTM. The range of the values for each hyper-parameter is shown in Table 4.2. We should point out that due to memory constrains the batch size for the visual and multimodal models is in the range $[2, 3]$, whereas for the text and audio modalities was set in the range $[10, 20, 30]$.

Due to the high number of configurations, we adopt the Hyperband algorithm [89] to tune the hyper-parameters. This is an iterative algorithm that starts by randomly sampling a number of hyper-parameter configurations in the search space, and discards some of them (in our case a 25%) during the training of the model.

After tuning the hyper-parameters, we use Adam optimiser [81] with initial learning

rate of 5×10^{-4} throughout all experiments. The length of a sequence in a batch is set to 100. The batch size for training the visual and multimodal networks was set to 3, while for the text and audio models it was set to 20. Additionally, we apply dropout [141] to the text and audio models so that our model would not overfit on the training data. In particular, for the audio model dropout is performed with a probability of 0.5 after each convolution layer. For the text model, we apply dropout of 0.1 for original word representations, 0.3 for enriched position-wise features, 0.3 for the output hidden states of context feature generator, and 0.2 for the output of the multi-linear projection matrices. Finally, each LSTM network we use in the training phase is trained with a dropout of 0.5.

4.4.4 Unimodal Ablation Study Results

We experiment with different well-established networks for each modality in order to choose the best performing one. All models were trained using 1-layer LSTM on top of the extracted features such that the temporal dynamics in the signal are captured. Using the visual modality we experiment with VGG [107], ResNet-50 [62], DenseNet [66], MobileNet [64], and HRNet [165]. Using the text modality we experiment with different word embedding networks, namely, Fasttext [17], BERT [30], AlBert [86], XLNet [176], RoBerta [93], ULMFit [65]. We should point out that we did not fine tune these models, we simply used the extracted word embeddings. Finally, we experiment with three end-to-end models using the audio modality, i. e, Trigeorgis et al. [151], Emo-18 (proposed in Ch. 3) [161], and Li et al. [91].

Table 4.3 depicts the results for all modalities. From the results we conclude that the best average CCC for the visual modality is HRNet. This is an interesting finding as HRNet model utilises different image resolutions in its architecture and we believe it can capture nuance facial expression characteristics compared to the other architectures. The best performing word embedding model for the proposed

Modality	Network	Arousal	Valence	Avg
Visual	Resnet-50 [62]	.641	.689	.665
	VGG-16 [107]	.624	.668	.646
	DenseNet [66]	.617	.672	.645
	MobileNet [64]	.621	.689	.655
	HRNet [165]	**.647**	**.695**	**.671**
Text	Fasttext [17]	.508	.554	.531
	BERT [30]	.537	.578	.558
	AlBert [86]	.532	.577	.555
	RoBerta [93]	.523	.561	.542
	ULMFit [65]	.527	.559	.543
	XLNet [176]	**.544**	**.581**	**.563**
Audio	Trigeorgis et al. [151]	.541	.488	.513
	Li et al. [91]	.552	.534	.543
	Emo-18 [161]	**.563**	**.532**	**.545**

Table 4.3: SEWA database development results (in terms of CCC) for the visual, text, and audio modalities, for the prediction of arousal and valence.

text architecture is found to be the XLNet. This was expected as XLNet was trained on multiple languages compared to Fasttext and the variations of the BERT. Lastly, for the audio modality our proposed model, Emo-18, (Sec. 3.2.2) outperforms the other two. We see that our model performs well with both constrain and naturalistic conditions. We believe our model performs better than the one proposed by Li et al. [91] because it does not need abundance of training data as opposed to that model that utilises a self-attention layer on top of a 2-layer Bi-LSTM. To what follows, we utilise the best performing models of each modality in our multimodal network.

4.4.5 Unimodal Results

Before showing the results of the multimodal network, we compare each of our unimodal networks to the baseline paper of the AVEC 2017 challenge along with the results of the best performing models of that year.

Text Modality

Most of the studies in the AVEC 2017 challenge use BoTW to extract features from the text. Only Chen et al. [24] uses word2vec features. Table 4.4 depicts the results in the test and development sets. Our model outperforms all the other methods in predicting both the arousal and valence dimensions with high margin. This was expected as our model is deeper and hence able to learn more complex functions. In addition, our model can better capture the contextual information in the sentences and can attend to the most relevant information with the use of attention. Finally, our model uses state-of-the-art word embeddings trained on large scale databases such as the Wikipedia one which increases its performance.

Predictor	**Arousal**	**Valence**
Baseline[115]	.375 (.373)	.425 (.390)
Dang et al. [28]	.320 (.441)	.394 (.499)
Huang et al. [68]	.483 (.489)	.520 (.523)
Chen et al. [24]	.463 (.478)	.515 (.532)
Proposed	**.532** (.544)	**.568** (.581)

Table 4.4: SEWA database test results (in terms of CCC) for the text modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set.

Audio Modality

As already discussed, the audio models are more effective at predicting the arousal dimension rather than valence. Table 4.5 shows the results of our model and the rest of the AVEC challenge papers. Our model has the second highest performance behind the Huang et al. [68] method. We should note, however, that the network they use was pre-trained on 300 hours of a spontaneous English speech recognition corpus before fine-tuning it to the SEWA database. In addition to the features of the network, they also utilise several hand-engineered features. The baseline and the Dang et al. [28] methods use hand-crafted features and as we discussed in the previous chapter the performance of such approaches is low. Our model performs better than Chen et al. [24], although they utilise both hand-crafted and pre-trained SoundNet [10] features. However, based on their ablation study the features extracted by SoundNet perform similarly as the hand-crafted ones, indicating the low generalisation ability of the network compared to our model.

Predictor	**Arousal**	**Valence**
Baseline [115]	.225 (.344)	.244 (.351)
Dang et al. [28]	.344 (.494)	.346 (.507)
Huang et al. [68]	**.583** (.584)	**.487** (.585)
Chen et al. [24]	.422 (.524)	.405 (.504)
Emo-18	.456 (.563)	.428 (.532)

Table 4.5: SEWA database test results (in terms of CCC) for the audio modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set.

Visual Modality

The prediction of the valence dimension can be benefited using the visual information. Table 4.6 depicts the results using the visual modality only. Dang et al. [28]

utilises hand-crafted features for the visual modality as such it produces the worst results among the three best performing approaches.Huang et al. [68] model produces better results but their model is not able to generalise well on the arousal dimension. A possible reason is the inefficiency of combining the multiple of visual features (Geometric, LGBP-TOP, HOG and Deep visual) that are used in their approach during training. For the Chen et al. [24] approach we were not able to get their results on the test set so we can not make a conclusion on the generalisability of their approach. The proposed variation of HR-Net provides state-of-the-art results compared to all the other methods in both arousal and valence dimensions. Finally, noticing the performance of the models on the development set, we can see that our model has better generalisation capabilities. We believe this is due to the attentional pooling layer we used, as has been shown on different studies that utilise attention methods [105, 48].

Predictor	**Arousal**	**Valence**
Baseline[115]	.308 (.466)	.455 (.400)
Dang et al. [28]	.390 (.518)	.496 (.583)
Huang et al. [68]	.531 (.682)	.670 (.720)
Chen et al. [24]	— (.675)	— (.693)
Proposed	**.608** (.647)	**.675** (.695)

Table 4.6: SEWA database test results (in terms of CCC) for the visual modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set. A dash is inserted if the results could not be obtained.

4.4.6 Fusion Strategy Results

In this subsection we show the results of our different fusion strategies, namely, concatenation, hierarchical attention, self-attention, residual self-attention, and cross-modal hierarchical self-attention. Table 4.7 shows the results on the development

set. As expected the simple concatenation method produces the worst results. The hierarchical attention outperforms, in both arousal and valence dimensions, the self-attention and residual self-attention, but not the cross-modal hierarchical self-attention. We believe that the bottom-up approach of the hierarchical attention provides the network the capability to more efficiently attend to one of the modalities before combining them together. Compared to hierarchical attention, the cross-modal hierarchical self-attention provides better results. A possible explanation is that the latter approach combines each modality with the other two as a first step, and hence looking at the information of all modalities at once before combining them into a single feature vector, compared to the hierarchical attention method that fuses each modality with a different one.

Method	**Arousal**	**Valence**
Concatenation	.698	.737
Self-Attention	.729	.751
Hierarchical Attention	.785	.808
Residual Self-Attention	.748	.767
Cross-modal Hierarchical Self-Attention	**.797**	**.818**

Table 4.7: Results (in terms of CCC) on the development set of the SEWA database for the prediction of arousal and valence using five different fusion strategies, namely, concatenation, self-attention, hierarchical attention, residual self-attention, and cross-modal self-attention.

4.4.7 Multimodal Results

To have a fair comparison with the models used in the AVEC 2017 challenge, and the Singh et al. [135] and Khorram et al. [77] studies, our model was trained utilising all the modalities of the database, namely, text, audio, and visual, for the final predictions. The results are shown in Table 4.8. Our model is the highest performing model compared with the other models. We should note that Chen et al.'s [24]

model utilises several hand-engineered and deep features, while our model operates on the raw signal. For example, the visual features they use are extracted from the DenseNet [66] and the VGG [107], while for the audio the features are the IS10 [128] and the features extracted from the SoundNet [10].

Predictor	Arousal	Valence
Baseline [115]	.375 (.525)	.466 (.507)
Singh et al. [135]	.276 (.294)	.365 (.346)
Khorram et al. [77]	.412 (.530)	.379 (.542)
Dang et al. [28]	.523 (.657)	.540 (.602)
Huang et al. [68]	.599 (.721)	.721 (.728)
Chen et al. [24]	.672 (.823)	.756 (.796)
Proposed	**.690** (.797)	**.783** (.819)

Table 4.8: SEWA database results (in terms of CCC) for our multimodal network, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set.

4.4.8 Cross-dataset Evaluation

In this section, we investigate how the performance of the proposed end-to-end approach is affected in two scenarios: (i) when constrain conditions are entailed, and (ii) when the models are trained on constrained conditions and evaluated in naturalistic ones, and vise-versa. To this end, we leverage the RECOLA (Sec. 3.3.1) and SEWA datasets. We should point out that we fix the parameters of the network responsible for the spatial feature extraction part, i. e. the convolution networks, and train the recurrent network, and the output layer. We perform this approach to compensate for the timestamp difference in the annotations of each dataset. Table 4.9 shows the results.

We observe that with the current approach - visual and audiovisual, audio is the same - we obtain better results than the ones we had in the previous chapter (Ch. 3).

Thus, we see that the attention mechanism used to fuse the modalities improves the performance of the models. In addition, we see that training the models using the RECOLA dataset, and then fine-tune it on the SEWA we get worse results than training the model directly on the SEWA. On the other hand, training the model using SEWA, and fine-tuning it on the RECOLA provides slightly better results than training directly on the RECOLA. These results were expected as RECOLA recordings were captured under constrained conditions, whereas SEWA's under naturalistic. Finally, we should note the high decrease of the the audio model's performance when trained using RECOLA and fine-tuned on the SEWA, and the slight increase in performance of the model when the inverse takes place. This was expected as the SEWA dataset contains 6 different languages but RECOLA contains only one.

Train \ Evaluate	Modality	RECOLA		SEWA	
		Arousal	**Valence**	**Arousal**	**Valence**
RECOLA	Visual	.441 (.417)	.628 (.653)	.557 (.584)	.631 (.631)
	Audio	.778 (.808)	.436 (.493)	.415 (.523)	.402 (.501)
	Audiovisual	.797 (.736)	.721 (.614)	.594 (.711)	.713 (.743)
SEWA	Visual	.457 (.461)	.636 (.664)	.608 (.647)	.675 (.695)
	Audio	.781 (.809)	.447 (.502)	.456 (.563)	.428 (.532)
	Audiovisual	.807 (.751)	.728 (.658)	.626 (.742)	.751 (.782)

Table 4.9: Cross-dataset results (in terms of CCC) for the visual, audio, and audiovisual modalities, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set.

4.4.9 Hyper-parameter Optimisation

Each hyper-parameter has a different effect on the final multimodal network. In Table 4.10 we show, in terms of the average CCC, the effect of each hyper-parameter on the development set of the SEWA database. Results indicate that dropout has the highest effect on the model's performance with the highest value of standard

deviation on both arousal and valence. On the other hand, the batch size has the lowest effect with 0.01 standard deviation. In addition, the number of hidden units and number of LSTM layers seem also to have a small impact on the performance of the model.

Hyper-parameter	**Arousal**	**Valence**
Learning rate	$.765 \pm .03$	$.798 \pm .02$
Batch size	$.785 \pm .01$	$.805 \pm .01$
Dropout	$.742 \pm .05$	$.766 \pm .04$
Sequence length	$.783 \pm .03$	$.796 \pm .03$
Hidden units (LSTM)	$.783 \pm .02$	$.801 \pm .02$
Num layers (LSTM)	$.759 \pm .03$	$.792 \pm .03$

Table 4.10: SEWA database results (in terms of CCC) for the multimodal model on the development set for different hyper-parameters of the model.

4.5 Conclusions

In this chapter, we proposed a transformer-based text architecture which utilises multi-linear projection and a context-aware feature generator to capture the semantics of a sentence. Furthermore, we proposed attention-based fusion strategies to balance the relationship among different modalities when combined in a multimodal network, which achieves a higher affect recognition performance on naturalistic scenarios. In addition, our text, visual and multimodal models outperformed the (at the time) state-of-the-art methods, while our audio network achieved the second best performance when compared with models that utilise several hand-crafted and deep features. Finally, we performed a cross-dataset evaluation where we showed that models that are trained in more difficult conditions like naturalistic ones that SEWA provides, provide better results when they are fine-tuned on easier conditions like the ones provided by RECOLA. The opposite, however, i. e. training on controlled conditions and fine-tuning on naturalistic ones, is not true.

Chapter 5

End-to-End Multimodal Profiling Applications

Contents

4.1 Motivation 68

4.2 Notation 71

4.3 Proposed Method 71

4.4 Experiments & Results 82

4.5 Conclusions 93

5.1 Motivation

In the previous two chapters, we saw that end-to-end deep learning approaches using audio, visual, text, or multimodal input produced high performing models, compared with traditional approaches that use hand-engineered features. This is accomplished by providing features tailored to the emotion recognition task under both posed and naturalistic conditions. Our findings were aligned with the affective computing literature [24, 28, 92] where Deep Neural Networks (DNN) have

systematically outperform conventional approaches. However, there are still several methods, especially in the audio domain, that utilise hand-crafted features. Examples can be found in several competitions such as the Interspeech Computational Paralinguistics ChallengE (ComParE) [127, 128], and the Audio/Visual Emotion Challenge and Workshop (AVEC) [116, 115]. One of the reasons that drives the use of conventional approaches instead of DNN ones is that they are well-established methods with open-source toolkits (e. g. openSMILE [43] in the audio domain) that have been used for several years. On the other hand, there are recently several end-to-end learning approaches [155, 161, 151] that have not yet been established as baseline methods in the affective computing community.

To fill this gap in the community we propose the End2You toolkit – the Imperial College London toolkit for end-to-end multimodal profiling capabilities written in PyTorch [108]. The toolkit provides a number of models for one-, two-, or three-dimensional input signals for either unimodal or multimodal input such as audio, visual, audiovisual, or any other kind of modality or sensor data such as physiological data. The provided models produce strong results in a number of tasks including emotion recognition. Depending on the task at-hand, we provide the ability to the user to train a model from scratch, i. e., using random weight initialisation, or fine-tuned using pre-train weights such as the ones from the AffWild database [84].

In more detail, our models can perform spatial feature extraction, namely, extracting features from a single frame, using for example CNNs, or temporal feature extraction to extract feature on the time domain using RNNs (e. g. Long Short-Term Memory (LSTM) networks), or both spatial and temporal feature extraction such as the proposed approaches on the former two chapters. As discussed earlier, our toolkit can combine multiple input signals in a single network with different types of architectures that incorporate spatial and temporal feature extraction layers. On top of that, we provide the a number of multimodal fusion strategies using atten-

tion mechanisms that can balance the relationship among different modalities and produce well performing models, as shown in Chapter 4.

We show the favourable properties of our toolkit in a number of tasks and compare with several other approaches. First, we use the REmote COLlaborative and Affective (RECOLA) database as used in the AVEC 2015 Challenge [116] to model the affective state of individuals, and compare with the baseline methods of the challenge. In addition, we achieve state-of-the-art results on the RECOLA database when we use a pre-trained facial network trained on an industrial large scale dataset and provided by our toolkit.

Our toolkit has been used with great success providing strong baselines in a number of challenges: (i) the Interspeech ComParE 2017, 2018, 2021 [131, 127, 133], (ii) the ICMI 2018 Eating Analysis and Tracking (EAT) Challenge [58], and (iii) the ACM MM MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media: Emotional Car Reviews in-the-wild [142]. We show the results of our toolkit and compare it with other well-known and established toolkits such as openSMILE [43], OpenXBOW [124], DeepSpectrum [4], and AuDeep [46].

The DeepSpectrum [4] and AuDeep [46] toolkits are most related to ours, as they also use deep learning models. However, instead of using the raw input signal, they first extract spectrograms to feed their networks. DeepSpectrum then uses pre-trained on ImageNet visual networks such as VGG [107], or ResNet [62] to extract features that are used as input to a learnable predictor (e. g. Support Vector Machine (SVM)). On the other hand, AuDeep uses deep recurrent sequence-to-sequence architectures to extract features with the networks trained in an unsupervised manner. As with the DeepSpectrum, the extracted features feed a predictor for the final prediction.

The contributions of this chapter can be summarised as follows:

- present End2You, an end-to-end deep learning toolkit, providing unimodal and multimodal generic capabilities.
- extensive experimentation and comparisons of our toolkit with other approaches.

The content of this chapter is based around the following publications:

- **Tzirakis, P.**, Zafeiriou, S. and Schuller, B.W., 2018. *End2You–The Imperial Toolkit for Multimodal Profiling by End-to-End Learning.* arXiv preprint arXiv:1802.01115.
- **Tzirakis, P.**, Boros, D., Hajiyev, E., and Schuller, B.W.. *Facial Emotion Recognition using Deep Residual Networks in Real-World Environments*, 2021.
- Schuller, B. W and Batliner, A. and Bergler, C. and Mascolo, C. and Han, J. and Lefter, I. and Kaya, H. and Amiriparian, S. and Baird, A. and Stappen, L. and Ottl, S. and Gerczuk, M. and **Tzirakis, P.** and Brown, C. and Chauhan, J. and Grammenos, A. and Hasthanasombat, A. and Spathis, D. and Xia, T. and Cicuta, P. and Rothkrantz, J.M. L. and Zwerts, J. and Treep, J. and Kaandorp, C. *The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 cough, COVID-19 speech, escalation & primates*, arXiv:2102.13468, 2021.
- Schuller, B., Steidl, S., Batliner, A., Marschik, P.B., Baumeister, H., Dong, F., Hantke, S., Pokorny, F.B., Rathner, E., Bartl-Pokorny, K.D., Einspieler, C., Zhang, D., Baird, A., Amiriparian, S., Qian, K., Ren, Z., Schmitt, M., **Tzirakis, P.**, Zafeiriou, S. *The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical & Self-Assessed Affect, Crying & Heart Beats.* In Proc. Interspeech, pp. 122-126, 2018.
- Schuller, B., Steidl, S., Batliner, A., Bergelson, E., Krajewski, J., Janott, C., Amatuni, A., Casillas, M., Seidl, A., Soderstrom, M., Warlaumont, A.S., Hi-

dalgo, G., Schnieder, S., Heiser, C., Hohenhorst, W., Herzog, M., Schmitt, M., Qian, K., Zhang, Y., Trigeorgis, G., **Tzirakis, P.**, Zafeiriou, S. *The INTERSPEECH 2017 Computational Paralinguistics Challenge: Addressee, Cold & Snoring.* In Proc. Interspeech, pp. 3442-3446, 2017.

- Stappen, L., Baird, A., Rizos, G., Tzirakis, P., Du, X., Hafner, F., Schumann, L., Mallol-Ragolta, A., Schuller, B., Lefter, I., Cambria, E., and Kompatsiaris, I.. *MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-target Engagement and Trustworthiness Detection in Real-life Media: Emotional Car Reviews in-the-wild. In Proc. of the 1st International on Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop (MuSe'20).* In Proc. ACM Multimedia, pp. 35–44, 2020.

- Hantke, S., Schmitt, M., **Tzirakis, P.** and Schuller, B.. *EAT- The ICMI 2018 Eating Analysis and Tracking Challenge*, In Proc. ACM International Conference on Multimodal Interaction. pp. 559-563, 2018.

The remainder of the chapter is organised as follows. Sec. 5.2 introduces the End2You toolkit with a system overview. Sec. 5.3 presents the experiments performed and reports the results using both posed and real-word datasets. Finally, Sec. 5.4 concludes this chapter.

5.2 End2You Toolkit

Our toolkit provides a set of models to extract information from the raw 1D (e. g., audio, heart rate, EEG, etc.), 2D (e. g., spectrogram, images, etc.), or 3D (e. g., video) input signals. As mentioned earlier, the provided models combine a feature extractor, that computes spatial features, and a temporal extractor, that extract contextual information from the data. Depending on the nature of the data, a number of architectures can be defined.

One-dimensional. The models provided are from Tzirakis et al. [161] and Trigeorgis et al. [151]. Both models base on a similar architecture with a feature extractor and a RNN on top to capture the contextual information in the data. In particular, the feature extractor of [161] uses three blocks of convolution and max-pooling, and the one by [151] two such blocks. Both models have shown strong empirical results in a number of tasks [157].

Two-dimensional. A number of models are supported to process the two-dimensional, including ResNet [62], DenseNet [67], etc. An important feature of our toolkit is that three types of weight initialisation are provided for these models: random initialisation, a pre-trained model based on ImageNet [85], and a pre-trained model based on AffWild [84].

Multimodal. The 1D and 2D model extractor can be used to train a multimodal network by fusing their features. We provide several attention mechanisms to fuse the modalities, which can be used to visualise the importance of each modality per sample as shown in Chapter 4.

5.2.1 System Overview

End2You is implemented in Python, and can be used either with the Python API or using the command line interface. The code is publicly available on GitHub under a modification of the BSD-3 license[1]. The system workflow is shown in Fig. 5.1. First, the raw file formats are transformed to the *.hdf5* file format, for efficient loading from disk. Both 1D and 2D unimodal or multimodal data can be transformed. At this stage, the toolkit supports the capability to use the face detection model Multi-task Cascaded Convolutional Networks (MTCNN) [180] to extract the face from an image frame. After the data transformation, the training phase takes place. In particular, this phase is comprised of: (i) the data provider that reads the hdf5

[1] cf. https://github.com/end2you/end2you

files, (ii) the model to be trained, and (iii) the prediction task, which can be either classification or regression. A number of loss functions, optimisers, and metrics are provided in the toolkit.

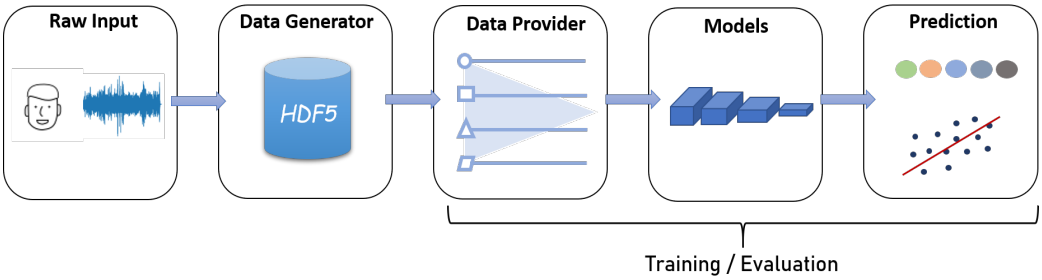


Figure 5.1: The End2You workflow is comprised of the data generator that transforms the raw file format to hdf5, the data provider that feeds the either static or dynamic data to the models, and finally, the prediction.

5.3 Experiments

We now compare our toolkit with several other toolkits, on a number of tasks. First, we use the RECOLA database as used in the AVEC 2015 Challenge [114], and compare with the baseline methods of the challenge (Sec. 5.3.1). Then, we perform facial emotion recognition with our toolkit and use a pre-trained network, and achieve state-of-the-art results.

Our toolkit has been used in a number of challenges across the years for different types of tasks including heart beat identification, snoring identification and others. Here we present its results and compare with other toolkits. In Sec. 5.3.3 we present the results for the Interspeech ComParE challenges, where different types of paralinguistic data have been used. In Sec. 5.3.4 we discuss the results for the Multimodal Sentiment Analysis (MuSe) challenge. Finally, in Sec. 5.3.5 we show the results for the Eating Analysis and Tracking (EAT) challenge. In all challenges, but the MuSe one which uses a large-scale dataset, we compare our toolkit with other well-established toolkits.

5.3.1 Multimodal Emotion Recognition

We first demonstrate the capabilities of End2You by creating both unimodal and multimodal models, and trained them for dimensional affect recognition. To this end, we utilise the REmote COLlaborative and Affective (RECOLA) database [117], as used in the Audio/Visual Emotion Challenge and Workshop (AVEC) repeatedly since 2016 [162]. It includes four modalities or sensor-signals, namely,audio, video, electro-cardiogram (ECG), and electro-dermal activity (EDA). The model needs to predict two time-continuous outputs of spontaneous and natural emotions (arousal and valence).

The unimodal input is comprised of audio, video, or physiological data (ECG), and the multimodal input combines all of them. For the audio and video models, a 2-layer gated recurrent unit (GRU) network with 64 units is stacked on top for the prediction. For the physiological modality, the network is a 1-layer GRU with 64 units. The multimodal model is comprised of the video and the audio model, and a fully connected network (FCN) of 256 units that combines the features extracted by the unimodal models and also the raw physiological information frames. On top of this model, we stack a 2-layer GRU network. Our loss and metric functions are based on the concordance correlation coefficient (CCC) as was used as competition measure in AVEC, and as optimiser, Adam is used.

We compare the results of End2You with the baseline and the winner paper in the AVEC 2016 emotion prediction sub-challenge. Results are depicted in Table 5.1 for both the unimodal (audio, video, physio) and the multimodal (using all modalities) cases. Our models perform better than the baseline in both the audio and video modalities, and close for the physiological one. As expected from the literature [159], the audio model performs best in the arousal domain and the video model in the valence one. Our multimodal model combines the best scores of the unimodal ones.

However, this is not enough to surpass the performance of the models from the winner of the challenge and the baseline. We believe that this performance gap is because these models use, among others, physiological hand-crafted features, such as heart rate and heart rate variability (HRHRV), which provide [162] more information for the emotion recognition task than the raw signals (ECG and EDA).

Modality	**Method**	**Arousal**	**Valence**
Audio	Baseline	.648 (.796)	.375 (.455)
	End2You	.657 (.753)	.278 (.361)
Video	Baseline	.272 (.379)	.507 (.612)
	End2You	.347 (.404)	.552 (.571)
Physio (ECG)	Baseline	.158 (.271)	.121 (.153)
	End2You	.154 (.264)	.052 (.291)
Mutlimodal	Baseline	.683 (.821)	.639 (.683)
	End2You	.672 (.762)	.521 (.493)
	Winner	.770 (.862)	.687 (.750)

Table 5.1: Results on the test set of the RECOLA database (w. r. t. CCC) for the unimodal (audio, video, and physio) and multimodal models. In parenthesis are the performances obtained on the development set.

5.3.2 Facial Emotion Recognition

Our toolkit provides a pre-trained deep residual network of 50 layers (ResNet-50) [62] that is trained in an end-to-end manner utilising an industrial dataset [154]. The dataset is comprised of more than a million face images and more than 2 500 thousand subjects for categorical emotion recognition, and in particular predicting eight emotions: Happy, Surprise, Confusion, Disgust, Contempt, Empathy, Other, and Neutral. The approach used to train the model as well as the dataset is out of the scope of this thesis. The interested reader is referred to [154] for more details.

We fine-tune this model on the RECOLA database, and compare different initiali-

sation schemes: (i) from scratch, i. e. randomly using Xavier initialisation, (ii) using the weights of the model pre-trained on the ImageNet dataset, and (iii) the current state-of-the-art AffWildNet [84]. The same training framework was used to all models, i. e., a feature- and a contextual-extractor.

Table 5.2 shows the results with respect to CCC. Our model provides the best results compared to the other methods in both the arousal and valence dimensions, which indicates the superiority of the pre-trained model. We should also point out that our pre-trained model has faster convergence (i. e., the time to reach its best performance) of the optimisation than the other two models. In particular, our model convergence's is more than 2 times faster than the pre-trained one from ImageNet, and almost 3 times faster than the random initialisation. Finally, we should note that all the models perform better in the valence dimension rather than the arousal one, which has also been shown in other emotion recognition studies that utilise the face for dimensional emotion recognition [160, 159].

Initialisation Strategy	**Arousal**	**Valence**
Avg random	.319 (.342)	.481 (.548)
pre-trained on ImageNet	.324 (.330)	.492 (.554)
AffWildNet [84]	.273 (—)	.526 (—)
Proposed	.341 (.376)	.538 (.596)

Table 5.2: Performance on the RECOLA dataset (w. r. t. CCC) between the proposed model and the rest. In parenthesis are the performance obtained on the validation set. A dash is inserted if the results could not be obtained.

5.3.3 Interspeech Computational Paralinguistics ChallengE (ComParE)

The Interspeech Computational Paralinguistics ChallengE (ComParE) series is an open Challenge in the field of Computational Paralinguistics which annually takes

place at INTERSPEECH since 2009. The focus of the series varies each year dealing with different signal processing problems such as snoring classification and COVID-19 detection from cough or speech.

Our toolkit has been used three years in the ComParE challenges (2017 [127], 2018 [133], and 2021 [133]). To what follows we yearly present the datasets and the baselines used in the challenges. For detailed information, the interested reader is referred to the ComParE paper.

Datasets

Interspeech ComParE 2017: In this years challenge three different problems are addressed: (i) the *Addressee* classification problem where a speech signal produced by an adult is classified based on whether it is a directed towards another adult or towards a child, (ii) the *Cold* classification problem where the goal is to predict whether a person is cold or not based on his/her speech signal, and (iii) the *Snoring* classification problem which goal is to identify four different types of snoring. Table 5.3 demonstrates the number of instances per split of the datasets.

Interspeech ComParE 2018: In this years challenge four different problems are addressed: (i) the *Atypical Affect* classification problem where a speech signal produced by handicapped individuals is classified in four basic emotions, (ii) the *Self-Assessed Affect* classification problem where the goal is to predict three valence scores provided by the speakers themselves, (iii) the *Crying* classification problem which goal is to identify three types of infant vocalisations, and (iv) the Heart Beats classification problem where three different types of heart beats need to be identified. Table 5.4 depicts the number of instances per split of the datasets.

Interspeech ComParE 2021: In this years challenge four different problems are addressed: (i) the *COVID-19 Cough* classification problem where a cough signal is classified as COVID-19 positive or negative, (ii) the *COVID-19 Speech* classification

#	Train	Devel	Test	Σ
Addressee Corpus				
Child	2 302	2 182	2 148	6 632
Adult	1 440	1 368	1 446	4 254
Σ	3 742	3 550	3 594	10 886
Cold Corpus				
Cold	970	1 011	895	2 876
Non-cold	8 535	8 585	8 656	25 776
Σ	9 505	9 596	9 551	28 652
Snore Corpus				
Velum	168	161	155	484
Oropharyngeal	76	75	65	216
Tongue	8	15	16	39
Epiglottis	30	32	27	89
Σ	282	283	263	828

Table 5.3: ComParE 2017 databases with the number of instances per split.

problem where the goal is to predict whether a speech signal is classified as COVID-19 positive or negative, (iii) the *Escalation* classification problem which goal is to assess three types of the level of escalation in a dialogue, and (iv) the *Primates* classification problem where four different species vs background need to be identified. Table 5.5 shows a summary of the datasets.

Experiments

We now show the results of all other baseline approaches used in each year of the challenge, and compare our toolkit with them. In all of the challenges, a fusion method has also been used that combines all baselines to a single predictor. As ensemble methods are out of the scope of this thesis, we do not show these results.

Interspeech ComParE 2017: Three baselines were used in this years challenge, namely, our End2You toolkit, a SVM trained with the compare feature set [43] using functionals over low-level descriptor (LLD), and a bag-of-audio-words (BoAW) feature set extracted with openXBOW [124]. From the results we can see that our

#	Train	Devel	Test	Σ
Atypical Affect				
Angry	125	50	272	447
Happy	743	965	650	2 358
Neutral	2 287	2 842	2 024	7 153
Sad	187	329	153	669
Σ	3 342	4 186	3 099	10 627
Self-Assessed Affect				
Low	95	79	75	249
Medium	388	310	353	1 051
High	363	353	297	1 013
Σ	846	742	725	2 313
Heart Beats				
Normal	84	32	28	144
Mild	276	98	91	465
Moderate	142	50	44	236
Σ	502	180	163	845

Table 5.4: ComParE 2018 databases with the number of instances per split.

approach performs worse than the other two toolkits. This is because of two fold. First, the splits on these datasets are performed based on the results provided by the compare feature set with the SVM, and as such they are optimised on this approach, making it unfair to our toolkit. Second, our approach needs an abundance of data to be able to outperform well-established hand-engineered features, otherwise it can easily overfit on the datasets. However, we should note once again, as indicated also in the previous chapters, the generalisation capabilities of our models, as the validation and test performance is similar. Table 5.6 depicts the best results with respect to Unweighted Average Recall (UAR), given by each model.

Interspeech ComParE 2018: In this years challenge the same baselines as in 2017 were used with the addition of AuDeep toolkit [46], which extracts features based on a recurrent sequence to sequence autoencoder approach implemented with RNNs and using as input spectrograms. From Table 5.7 we observe that our approach performs worse than the other approaches on all datasets. It was expected for the

#	Train	Dev	Test	Σ
COVID-19 Cough				
Negative	215	183	169	567
Positive	71	48	39	158
Σ	286	231	208	725
COVID-19 Speech				
Negative	243	153	189	585
Positive	72	142	94	308
Σ	315	295	283	893
Escalation				
L	156	69	260	485
M	75	34	191	300
H	64	16	50	130
Σ	295	119	501	915
Primate Vocalisations				
C	2 217	2 217	2 218	6 652
M	874	874	875	2 623
R	208	209	210	627
G	158	159	159	476
Background	3 458	3 459	3 461	10 378
Σ	6 915	6 918	6 923	20 756

Table 5.5: ComParE 2021 databases with the number of instances per split.

UAR [%]	**Addressee**		**Cold**		**Snoring**	
	Dev	Test	Dev	Test	Dev	Test
ComParE functionals + SVM	60.5	67.7	64.0	70.2	31.1	51.4
ComParE BoAW + SVM	62.4	68.2	63.9	66.7	44.2	51.2
End2You	60.9	59.1	58.6	59.6	40.3	40.3

Table 5.6: ComParE 2017 baseline results (w. r. t. UAR) for the Addressee, Cold, and Snoring sub-challenges.

openSMILE, and OpenXBOW methods as described in the previous challenge, but it is surprisingly to see that AuDeep performed higher than our approach as both toolkits rely on deep learning approaches. One explanation for this may be the use of spectrograms by the AuDeep toolkit where the frequency characteristics seem to

UAR [%]	**Atypical**		**Self-Ass.**		**Crying**		**Heart**	
	Dev	Test	Dev	Test	Loso	Test	Dev	Test
ComParE functionals + SVM	37.8	43.1	54.2	61.2	75.6	71.9	44.5	45.6
ComParE BoAW + SVM	40.5	36.5	55.8	59.1	75.9	71.8	42.3	47.2
auDeep RNN + SVM	39.1	34.9	44.8	54.4	73.2	71.1	34.1	42.1
End2You CNN+LSTM	41.8	28.0	49.7	45.8	–	63.5	41.2	37.7

Table 5.7: ComParE 2018 baseline results (w. r. t. UAR) for the Atypical, Self-Ass, Crying, and Heart sub-challenges.

provide useful information for the final prediction. Table 5.7 depicts the best results with respect to Unweighted Average Recall (UAR), given by each model.

Interspeech ComParE 2021: In this years challenge the same baselines as in 2017 were used with the addition of DeepSpectrum [4] which performs feature extraction from audio data with pre-trained Image Convolutional Neural Networks (CNNs) such as VGG [107]. In addition, DiFE is used which extracts Contextual word embeddings for one of this years dataset using a frozen and fine-tuned Dutch Bidirectional Language Transformer (BERT).

From Table 5.8 we observe that as in the previous challenges openSMILE and OpenXBOW provide the best results (w. r. t. UAR) compared to the other approaches, including ours. This was expected as explained earlier with the ComParE 2017 challenge. Comparing our toolkit with the DeepSpectrum, we can see that our toolkit provides better results with the COVID-19 datasets and comparable results with the Escalation one. However, DeepSpectrum's performance is better on the Primates dataset. We can conclude that our toolkit performs better when processing human vocalisation signals but DeepSpectrum is better to model Primates signal. We believe that the frequency characteristic (represented as spectrogram) on these types of signals provide important patterns that the visual networks can capture.

UAR [%]	COVID-19 Cough		COVID-19 Speech		Escalation		Primates	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
ComParE functionals + SVM	61.4	65.5	57.9	72.1	70.5	58.6	82.4	82.2
ComParE BoAW + SVM	64.7	72.9	66.3	68.7	70.6	59.8	–	–
DeepSpectrum + SVM	41.8	28.0	49.7	45.8	–	63.5	41.2	37.7
auDeep RNN + SVM	67.6	67.6	59.4	53.3	42.0	44.3	81.6	84.1
End2You CNN+LSTM	61.8	64.7	70.5	68.7	64.1	54.0	72.7	70.8

Table 5.8: ComParE 2021 baseline results (w. r. t. UAR) for the Covid-19 Cough and Speech, Escalation, and Primates sub-challenges.

Similar observations can be performed when comparing our approach with AuDeep. Our toolkit provides better results with the COVID-19 databases and the Escalation one, but worse for the Primates, indicating that our tookit can better model the human vocalisation signals on these datasets than AuDeep, but lacks this capability for the Primate's one. As with DeepSpectrum, we believe that the use of spectrograms boost the performance of the toolkit.

5.3.4 Multimodal Sentiment Analysis (MuSe)

The Multimodal Sentiment Analysis (MuSe) 2020 challenge [142] tries to bridge the gap between emotion recognition and sentiment analysis by comparing deep learning approaches using large volumes of audiovisual and textual data captured under naturalistic conditions. Three tasks were introduced in this years challenge: (i) the *MuSe-Wild* which task is the prediction of continuous emotions (arousal and valence), (ii) the *MuSe-Topic* which task is to recognise three classes (low, medium, and high) of valence and arousal, that are the mean value of the temporally aggregated continuous labels of MuSe-Wild that correspond to a specific topic, and (iii) the *MuSe-Trust* where the proposed trustworthiness dimension is predicted. In what follows, we describe the dataset and the baseline experiments of the challenge.

Partition	No.	MuSe-Wild	MuSe-Topic	MuSe-Trust
Train	166	22 :16 :43	22 :35 :55	22 :45 :52
Devel.	62	06 :48 :58	06 :49 :46	06 :52 :22
Test	64	06 :02 :20	06:14 :08	06 :12 :53
$\sum$	291	35 :08 :01	35 :39 :49	35 :51 :07

Table 5.9: Partitioning of the *MuSe-Car* dataset, applied for each of the tasks. Reported are the number of unique videos, and the duration for each task hh :mm ss.

Dataset

For all tasks the MuSe-Car dataset [143] is used. The dataset is a large in-the-wild dataset providing multiple modalities: audio, visual, and transcripts, making a perfect tool for multimodal emotion recognition and sentiment analysis. It is a collection of YouTube videos captured inside or outside of a car under naturalistic conditions. It provides annotations of the continuous dimensions: arousal, valence, and the newly proposed trustworthiness. For more information, the interested reader is referred to the author's paper [143]. Table 5.9 shows the splits of the different tasks of the dataset.

Experiments

A number of hand-engineered and deep features were used as baseline methods. In particular, for the audio modality openSMILE and DeepSpectrum features were extracted. For the visual modality VGGface [107], XCeption [26], GoCar [144], MTCNN [180], and OpenPose [22] were used. Finally, for the text modality Fast-Text [17] features were extracted.

In total four baselines approaches were utilised in this challenge. End2You was used as the end-to-end learning approach. The other baseline was comprised of two parallel LSTM networks with a self-attention layer that used different types

System	Features	Muse-Wild			Muse-Trust
		Valence	Arousal	Combined	Trustworthiness
		devel / test	devel / test	devel / test	devel / test
LSTM + Self-ATT	LLD	.0711 / .0349	.3078 / .2834	.1894 / .1592	.2560 / .1343
LSTM + Self-ATT	DS	.0165 / .0024	.1585 / .1723	.0875 / .0874	.2019 / .1701
LSTM + Self-ATT	Ge	.0435 / -.0097	.1090 / .0827	.0762 / .0365	.1576 / .1385
LSTM + Self-ATT	FT	.1273 / .1816	.0959 / .1074	.1116 / .1445	.2278 / .2549
LSTM + Self-ATT	Ge + FT	.0520 / .0361	.1375 / .1018	.0947 / .0690	.2296 / .2054
LSTM + Self-ATT	X	.0499 / .0426	.0776 / .0683	.0638 / .0555	.1178 / .1664
LSTM + Self-ATT	aV	.0098 / .0272	.1598 / .1227	.0848 / .0749	.1167 / .1378
LSTM + Self-ATT	Ge + FT + V	.0393 / .0654	.1809 / .0865	.1101 / .0760	.1245 / .1695
End2You	FT + VG + RA	.1506 / .2431	.2587 / .2706	.2047 / .2568	.3198 / .4128

Table 5.10: Showing the results for the *MuSe-Wild* and *MuSe-Trust* tasks (w. r. t. CCC). As feature sets FastText (FT), eGeMAPS (eG), DeepSpectrum (DS), VGGface (VG), and Xception (X) and all visual (aV) are fed into the models. Furthermore, the raw audio signal (RA) is used by End2You, and LLD are utilised for *MuSe-Trust*.

of features. Finally, the Multimodal Transformer (MMT) was used to model the non-sequential predictions of *MuSe-Topic*.

Results (w. r. t. CCC) for the *MuSe-Wild* and *MuSe-Trust* tasks are depicted in Table 5.10. We observe that End2You provides the best results compared to the two parallel LSTMs for both *MuSe-Wild* and *MuSe-Trust* tasks. This was expected as *MuSe-Car* is a large scale dataset and as such the End2You models can find features tailored to the task at-hand and outperform other approaches that use hand-engineered features, as we have shown in the previous two chapters of this thesis.

Results for the *MuSe-Topic* task are depicted in Table 5.11. We can see that our toolkit provides worse results when compared with the other approaches. At first this can be surprising as these classes are the mean value of the temporally aggregated continuous labels of *MuSe-Wild*, where our model provided the best results. However, because of the nature of the data in this task where conversations take place, our model is not able to capture the dynamics in the dyadic data. On the

System	Features	c-Valence			c-Arousal			Combined
		F1	UAR	Combined	F1	UAR	Combined	
		devel / test	devel / test	devel / test	devel / test	devel / test	devel / test	devel / test
LSTM + Self-ATT	eG + FT + aV	36.06 / 37.14	35.20 / 37.14	35.77 / 37.14	39.92 / 35.16	40.44 / 34.76	40.10 / 35.02	37.93 / 36.08
MMT	DS + eG + VG	39.40 / 32.54	38.08 / 32.40	38.95 / 32.49	45.77 / 41.03	44.66 / 40.63	45.39 / 40.89	42.17 / 36.69
End2You	FT + eG + X	37.19 / 33.54	35.70 / 33.18	36.68 / 33.42	42.76 / 32.45	42.67 / 33.34	42.73 / 32.75	39.70 / 33.08

Table 5.11: *MuSe-Topic* Results using the Unweighted Average Recall (UAR), F1, and Combined ($0.66 \cdot F1 + 0.34 \cdot UAR$). As feature sets FastText (FT), FastText (FT), eGeMAPS (eG), DeepSpectrum (DS), Xception (X), VGGface (VG), and all visual feature set (aV) are used.

other hand, the other approaches with the use of self-attention mechanism are able to do so.

5.3.5 Eating Analysis and Tracking (EAT)

The goal of the Eating Analysis and Tracking (EAT) challenge is to monitor automatically eating patterns using multiple input signals such as audio and visual. Three tasks were introduced in the challenge: (i) the *Food Type*, where the task is to find the food type of the speaker, (ii) the *Likability* sub-challenge, where the goal is to recognise the speaker's likability of the food, and (iii) chewing and speaking *Difficulty* sub-challenge, where the task is to recognise the level of difficulty of a speaker. In the next subsections, we describe the dataset along with the two baselines that were used in the challenge.

Dataset

The EAT dataset [59] is used in this challenge. This is an audio-visual dataset which was also partly featured as a Sub-Challenge of the Interspeech 2015 ComParE Challenge [130]. It is comprised of 30 speakers (15 male and 15 female) that were recorded under controlled environmental conditions (quiet and low reverberant office room) eating a specific food type or not. Self-annotations ('Neutral', 'Like') are provided on how much the individuals like each sort of food, which food type are eating, and how difficult it is to speak while eating. Table 5.12 shows the split

#	Train	Test	Σ
No Food	140	70	210
Apple	140	56	196
Nectarine	133	63	196
Banana	140	70	210
Crisp	140	70	210
Biscuit	133	70	203
Gummi bear	119	70	189
Σ	945	469	1 414

Table 5.12: Number of instances per class in the train/test split used for the EAT challenge.

statistics of the dataset for different food types.

Experiments

Two baselines were used for this challenge, namely, the openXBOW and the End2You. Table 5.13 depicts the results of the baseline results across all sub-challenges. For the food type and likability sub-challenges, the UAR is used are metric, whereas for the difficulty sub-challenge the CCC is used. We see that our toolkit has worse performance in all the sub-challenges compared to openXBOW. In particular, for the food type and difficulty sub-challenges our model performs significantly worse using the audio modality and audiovisual modalities, and similar when visual modality is used. On the other hand, for the likability sub-challenge our model performs similar for the audio and visual modalities, but better for the audiovisual case.

We can see that our model performs better when human vocalisations exist in the input signal, as is the case in the likability sub-challenge. In addition, as observed in the previous sections our toolkit need large-scale data in order to be able to find the necessary characteristics in the input signal (independent of modality),

and outperform hand-engineered features. Finally, it is interesting to note that openXBOW performs better in the unimodal case rather than the audiovisual one, indicating that it cannot cope with different input streams. On the other hand, our toolkit provides better results when both audio and visual signals are used in all sub-challenges but the difficulty one where with audio as input performs better. Hence, we can conclude that our toolkit can better cope with different modality signals compared to openXBOW.

Modality	openXBOW		End2You	
	LOSO	Test	Dev	Test
Food Type				
Audio-only	64.3	67.2	35.2	32.8
Video-only	28.0	27.0	27.1	24.5
Audio+Video	63.9	67.0	34.8	33.6
Likability				
Audio-only	66.5	54.2	55.1	53.7
Video-only	55.9	58.3	52.9	50.9
Audio+Video	65.5	51.8	55.1	54.2
Difficulty				
Audio-only	.470	.506	.342	.323
Video-only	.246	.252	.264	.220
Audio+Video	.481	.501	.345	.311

Table 5.13: Baseline results for audio, visual, and audiovisual modalities. Results for Food Type (7-classes) and *Likability* (2-classes) are in terms of UAR, and results for *Difficulty* (regression) are in terms of CCC.

5.4 Conclusions

In this chapter, we have introduced End2You – the Imperial toolkit for end-to-end multimodal profiling written in Python and using PyTorch. It provides capabilities for training and evaluating models with either unimodal or multimodal input, and several implemented architectures are provided for 1D, 2D, or 3D inputs such as audio, visual, audiovisual, and physiological data. We have shown the capabilities of our toolkit in a number of tasks including emotion recognition, snoring and cold identification, and others, and we have performed an extensive comparison with other well-established toolkits.

From our experiments we can conclude that our toolkit produces high performance models for large-scale datasets, and can easily surpass well-established hand-crafted feature sets such as openSMILE [43], which are better suited for small-size datasets. In addition, our toolkit can better model the characteristic of the speech signal, rather than other types of signals such as heart beats (see Sec. 5.3.3).

For future work our toolkit can be extended by introducing models able to handle text information. In addition, pre-trained models of large scale datasets is important to be included in the toolkit. As we saw transfer learning is a vital method when limited data are provided.

Chapter 6

Conclusion

Contents

5.1 Motivation 94

5.2 End2You Toolkit 98

5.3 Experiments 100

5.4 Conclusions 115

Affect recognition is a vital aspect in human behaviour, which is subconsciously performed by humans. Nowadays, more and more intelligent systems are using emotion recognition models to improve their interaction with humans by adapting their responses and behavioural patterns according to the human emotions and try to make the interaction more *natural*. Applications can be found in automatic tutoring system, and call centres.

Substantial efforts have been performed by the affective computing community towards emotional intelligence systems. However, there are several challenges that are still posed towards that goal. Most of the current systems operate on constrained (laboratory) conditions which do not allow them to generalise to naturalistic settings. In addition, due to the temporal variations that arise in the signals the fusion

of different modalities in a single system is not a straightforward process

In this thesis we aim at investigating deep learning approaches for the emotional recognition task in order to alleviate from these challenges. In particular, we present the following four contributions: *(i)* show that integrating multiple modalities in an end-to-end DNN model outperforms unimodal ones for both posed and naturalistic conditions, *(ii)* study the performance of deep learning models by analysing spontaneous and naturalistic behaviour, *(iii)* investigate the performance gains of attention methods to fuse multiple modalities to both deal with the discrepancy and temporal variations that exists between the modalities, and *(iv)* the End2You toolkit that provides end-to-end profiling capabilities.

The rest of the chapter presents a summary of the methods introduced, and concludes this thesis (Sec. 6.1). Finally, future research directions are presented (Sec. 6.2).

6.1 Contributions

We proposed deep learning approaches and models that were evaluated on both controlled and naturalistic environmental conditions using publicly available emotion databases. Our approaches were shown beneficial properties for automatic emotional behaviour analysis.

In Chapter 3, we introduced two speech networks: (i) a light network based on previous published studies that uses a smaller amount of parameters whilst also outperforming previous studies, and (ii) a network architecture that considers a small kernel size for the convolution layers and a large one for the max-pooling layers. We show that using an multimodal system outperforms a system that uses a single modality. In addition, we presented with an audiovisual emotion recognition system, that effectively operates in controlled conditions when trained with the concordance correlation coefficient (CCC) measure. Our experimental results indicate that our model can more efficiently capture complex dynamics in the signal compared to

more traditional methods that utilised hand-engineered features.

In Chapter 4, we extended our previous study by presenting deep learning models operating under naturalistic conditions. We included in our model a text architecture based on self-attention method that can efficiently capture the semantics of a sentence with the use to recurrent neural networks and attention. To combine this modality with the others (audio and visual), we use attention methods. Our experimental results indicate the benefits of our approach when modelling the text signal compared to more traditional approaches. In addition, we showed that using attention methods can improve the performance of multimodal systems when combining the unimodal information compared to more conventional approaches such as a simple concatenation. Finally, we performed a cross-dataset experiment and show that models trained in naturalistic conditions and fine-tuned on controlled ones, can outperform models trained on controlled conditions.

Finally, in Chapter 5 we introduced the End2You toolkit that provides end-to-end multimodal profiling capabilities. The input signal can be of one-, two-, or three-dimensions with multiple modalities. We extensively experimented with our toolkit in a number of tasks where we compare its performance with other well-established toolkits, and showed its beneficial capabilities when modelling speech, visual, and audiovisual modalities and when large-scale databases are utilised.

In summary, this thesis presented end-to-end deep learning-based approaches and showed the powerful capabilities of these networks to build better suited and more efficient multimodal emotional behaviour analysis systems that can operate in spontaneous and naturalistic (*in-the-wild*) conditions.

6.2 Future Work

Overall we have shown that emotional intelligence of systems can be achieved with deep learning architectures, models, and methodologies. The work presented in this

thesis can be extended in a number of ways.

Most of the studies in the literature fail to reflect on the cultural and linguistic diversities that exists among humans. Considering these factors can increase the generalisation capabilities of the emotion recognition systems, and hence improve their performance. In addition, few papers exist in the literature that take advantage of promising techniques such as zero-/one-shot learning [47], reinforcement learning [102], and federated learning [175]. Finally, an emerging technique where the affective computing literature lacks of methodologies is the Self-Supervised Learning (SSL) [134, 74, 120]. SSL allows deep learning networks to learn representations in an unsupervised manner by withholding some of the input data, and task the network in predicting them. Such methods are vital in cases where databases are not available, and it is expensive to collect data.

Finally, our End2You toolkit (Chapter 5) can be extended by incorporating also the text modality in the pipeline, and provide current state-of-the-art word embedding deep architecture, along with recently proposed attention-based fusion strategies for efficient multimodal fusion. In addition, pre-train models on other fields than emotion recognition can also be provided.

List of Tables

3.1 The replication of each bottleneck architecture of the ResNet-50 along with the size of the features maps of the convolutions. 50

3.2 Network architecture for Emo-16, and Emo-18 speech networks. . . . 53

3.3 The partitioning of the RECOLA database. 56

3.4 Results (in terms of CCC) on arousal and valence after 60 epochs when varying the sequence length for the speech and visual networks. 59

3.5 Results (in terms of CCC) on arousal and valence for the speech and visual networks when no post-processing is introduced and the same sequence length as in training is used (**NO**), when post-processing is applied (**PP**), and when the whole sequence is evaluated at once without any intermediate processing methods (**WH**). 60

3.6 RECOLA database results (in terms of CCC) for prediction of arousal and valence. In parenthesis are the performances obtained on the development set. In *a)* we optimised the models w. r. t. MSE whereas in *b)* w. r. t. CCC. 61

3.7 RECOLA database results (in terms of CCC) for the prediction of arousal and valence using the speech networks. In parenthesis are the performances obtained on the development set. A dash is inserted if the results were not reported in the original papers. 62

3.8 Performance comparison (w. r. t. CCC) between the proposed method and other state-of-the-art methods. In parenthesis are the performance obtained on the validation set. 63

3.9 RECOLA database results (in terms of CCC) for prediction of arousal and valence using the Visual Network. In parenthesis are the performances obtained on the development set. A dash is inserted if the results were not reported in the original papers. 64

3.10 RECOLA database results (in terms of CCC) for the prediction of arousal and valence using the proposed multimodal emotion recognition network. In parenthesis are the performance obtained on the development set. 67

4.1 Summary of the SEWA database. Number of cultures, conversations, subjects and their corresponding duration. 83

4.2 Range of values for hyper-parameter tuning. 84

4.3 SEWA database development results (in terms of CCC) for the visual, text, and audio modalities, for the prediction of arousal and valence. . 86

4.4 SEWA database test results (in terms of CCC) for the text modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set. 87

4.5 SEWA database test results (in terms of CCC) for the audio modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set. 88

4.6 SEWA database test results (in terms of CCC) for the visual modality, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set. A dash is inserted if the results could not be obtained. 89

4.7 Results (in terms of CCC) on the development set of the SEWA database for the prediction of arousal and valence using five different fusion strategies, namely, concatenation, self-attention, hierarchical attention, residual self-attention, and cross-modal self-attention. . . . 90

4.8 SEWA database results (in terms of CCC) for our multimodal network, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set. 91

4.9 Cross-dataset results (in terms of CCC) for the visual, audio, and audiovisual modalities, and for the prediction of arousal and valence. In parenthesis are the performances obtained on the development set. 92

4.10 SEWA database results (in terms of CCC) for the multimodal model on the development set for different hyper-parameters of the model. . 93

5.1 Results on the test set of the RECOLA database (w. r. t. CCC) for the unimodal (audio, video, and physio) and multimodal models. In parenthesis are the performances obtained on the development set. . . 102

5.2 Performance on the RECOLA dataset (w. r. t. CCC) between the proposed model and the rest. In parenthesis are the performance obtained on the validation set. A dash is inserted if the results could not be obtained. 103

5.3 ComParE 2017 databases with the number of instances per split. . . . 105

5.4 ComParE 2018 databases with the number of instances per split. . . . 106

5.5 ComParE 2021 databases with the number of instances per split. . . . 107

5.6 ComParE 2017 baseline results (w. r. t. UAR) for the Addressee, Cold, and Snoring sub-challenges. 107

5.7 ComParE 2018 baseline results (w. r. t. UAR) for the Atypical, Self-Ass, Crying, and Heart sub-challenges. 108

5.8 ComParE 2021 baseline results (w. r. t. UAR) for the Covid-19 Cough and Speech, Escalation, and Primates sub-challenges. 109

5.9 Partitioning of the *MuSe-Car* dataset, applied for each of the tasks. Reported are the number of unique videos, and the duration for each task hh:mm ss. 110

5.10 Showing the results for the *MuSe-Wild* and *MuSe-Trust* tasks (w. r. t. CCC). As feature sets FastText (FT), eGeMAPS (eG), DeepSpectrum (DS), VGGface (VG), and Xception (X) and all visual (aV) are fed into the models. Furthermore, the raw audio signal (RA) is used by End2You, and LLD are utilised for *MuSe-Trust*. 111

5.11 *MuSe-Topic* Results using the Unweighted Average Recall (UAR), F1, and Combined ($0.66 \cdot F1 + 0.34 \cdot UAR$). As feature sets FastText (FT), FastText (FT), eGeMAPS (eG), DeepSpectrum (DS), Xception (X), VGGface (VG), and all visual feature set (aV) are used. 112

5.12 Number of instances per class in the train/test split used for the EAT challenge. 113

5.13 Baseline results for audio, visual, and audiovisual modalities. Results for FOOD TYPE (7-classes) and *Likability* (2-classes) are in terms of UAR, and results for *Difficulty* (regression) are in terms of CCC. . . . 114

List of Figures

2.1 Circumplex of Affect. The horizontal axis represents the valence and the vertical the arousal. [119] 30

2.2 The three main fusion strategies: (a) early fusion, (b) model-level fusion, and (c) late fusion. 40

3.1 Depicting the (a) Emo-16, and (b) Emo-18 architectures. Emo-16 is comprised of two blocks of convolution and max-pooling with the second's block max-pooling to be applied in the channels. On the other hand, Emo-18 is comprised of three blocks of convolution and max-pooling, all of which operate in the time domain. 52

3.2 The network comprises of two parts: the multimodal feature extraction part and the RNN part. The multimodal part extracts features from raw speech and visual signals. The extracted features are concatenated and used to feed a 2-layer LSTM network. These are used to capture the temporal information in the data. 55

4.1 Text network architecture. The input is a word-embedding matrix that is passed to a N-gram feature extractor (padding is not shown), before being concatenated with the N-gram features. Then two linear projections are applied such that different contexts are extracted. Each context-matrix is processed by a 2-layer Bi-LSTM to get the semantics in the sentence before the selector that weights the outputs. Finally, the outputs of each recurrent network is concatenated and passed by a highway network to get the semantic-aware feature vector. (Best viewed in colour) 75

4.2 Attentional pooling layer. First, an attentional feature map is applied before taking a weighted average of these maps to compute the final feature representation. 77

4.3 Hierarchical attention fusion methodology. Each modality features are combined using a weighted average of them, and the corresponding outputs are again combined in the same manner. 77

4.4 Self-attention fusion methodology. All modality-specific features are concatenated and passed through a self-attention layer to combine them in the final fusion vector. 79

4.5 Residual self-attention fusion methodology. All modality-specific features are concatenated and passed through a self-attention layer, and a residual connection from the input to the output is added to combine them in the final fusion vector. 80

4.6 Showing the (a) Cross-modal attention (CMA) method, and (b) Cross-modal self-attention fusion architecture. 81

4.7 The multimodal network is comprised of the features extracted from the text ($\mathbf{x}^t$), audio ($\mathbf{x}^a$), and visual ($\mathbf{x}^v$) modalities, the attentional layer that fuses the extracted features ($\mathbf{x}^{fusion}$) using a weighted vector, and 1-layer LSTM that capture the temporal dynamics in the data before the final prediction. (Best viewed in colour) 82

5.1 The End2You workflow is comprised of the data generator that transforms the raw file format to hdf5, the data provider that feeds the either static or dynamic data to the models, and finally, the prediction. 100

Bibliography

[1] T. A., C. J. S., and A. Zisserman. The conversation: Deep audio-visual speech enhancement. In *Proc. Interspeech*, pages 3244–3248, 2018.

[2] S. M. Alarcao and M. J. Fonseca. Emotions recognition using eeg signals: A survey. *IEEE Transactions on Affective Computing*, 10(3):374–393, 2017.

[3] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 579–586, 2005.

[4] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller. Snore sound classification using image-based deep spectrum features. In *Interspeech 2017*, pages 3512–3516, 2017.

[5] C. N. Anagnostopoulos, T. Iliou, and I. Giannoukos. Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011. *Artificial Intelligence Review*, 43:155–177, 2015.

[6] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proc. International Conference on Spoken Language Processing*, 2002.

[7] S. Angelidis and M. Lapata. Multiple Instance Learning Networks for Fine-Grained Sentiment Analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31, 2018.

[8] B. Atmaja and M. Akagi. Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4482–4486, 2020.

[9] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379, 2010.

[10] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proc. Advances in neural information processing systems (NeurIPS)*, 2016.

[11] K. Bahreini, R. Nadolski, and W. Westera. Data fusion for real-time multimodal emotion recognition through webcams and microphones in e-learning. *International Journal of Human–Computer Interaction*, 32:415–430, 2016.

[12] T. Baltrušaitis, C. Ahuja, and L. P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[13] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang. Emotion recognition in the wild from videos using images. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, page 433–436, 2016.

[14] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 568–573, 2005.

[15] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 223–230, 2006.

[16] S. Bilakhia, S. Petridis, A. Nijholt, and M. Pantic. The MAHNOB Mimicry Database: A database of naturalistic human interactions. *Pattern Recognition Letters*, 66:52–61, 2015.

[17] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[18] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 97–104, 2016.

[19] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211, 2004.

[20] R. A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods and their applications. *Transactions on Affective Computing*, 1:18–37, 2010.

[21] E. Cambria, N. Howard, J. Hsu, and A. Hussain. Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics. In *Proc. IEEE symposium on computational intelligence for human-like intelligence (CIHLI)*, pages 108–117, 2013.

[22] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[23] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters*, 125:264–270, 2019.

[24] S. Chen, Q. Jin, J. Zhao, and S. Wang. Multimodal multi-task learning for dimensional and continuous emotion recognition. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 19–26, 2017.

[25] K. Cho, B. Van M., C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[26] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017.

[27] J. F. Cohn. Foundations of human computing: facial expression and emotion. In *Proc. International Conference on Multimodal Interfaces (ICMI)*, pages 233–238, 2006.

[28] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps. Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 27–35, 2017.

[29] J. Deng, G. Trigeorgis, Y. Zhou, and S. Zafeiriou. Joint multi-view face alignment in the wild. *IEEE Transactions on Image Processing*, 28:3636–3648, 2019.

[30] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[31] A. Dhall, R. Goecke, J. Joshi, J. Hoey, and T. Gedeon. Emotiw 2016: Video and group-level emotion recognition challenges. In *Proc. ACM International Conference on Multimodal Interaction (ICMI)*, pages 427–432, 2016.

[32] A. Dhall, R. Goecke, J. Joshi, M. Wagner, and T. Gedeon. Emotion recognition in the wild challenge (emotiw) challenge and workshop summary. In *Proc. International Conference on Multimodal Interaction (ICMI)*, pages 371–372, 2013.

[33] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon. Collecting Large, Richly Annotated Facial-Expression Databases from Movies. *MultiMedia*, 19:34–41, 2012.

[34] S. K. D'mello and J. Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.

[35] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J. C. Martin., L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis. The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Proc. International Conference on Affective Computing and Intelligent Interaction*, pages 488–500, 2007.

[36] K. Ebrahimi, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proc. IEEE International Conference on Multimodal Interfaces (ICMI)*, pages 467–474, 2015.

[37] J. Edwards, H. Jackson, and P. Pattison. Emotion recognition via facial expression and affective prosody in schizophrenia: a methodological review. *Clinical psychology review*, 22(6):789–832, 2002.

[38] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

[39] P. Ekman. Emotional and conversational nonverbal signals. In *Proc. Language, Knowledge and Representation*, pages 39–50. 2004.

[40] P. Ekman and W. V. Friesen. *Unmasking the face: A guide to recognizing emotions from facial clues*, volume 10. Ishk, 2003.

[41] M. El Ayadi, M. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[42] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *Transactions on Affective Computing*, pages 190–202, 2016.

[43] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. ACM International Conference on Multimedia*, pages 1459–1462, 2010.

[44] M. Faisal and S. Manzoor. Deep learning for lip reading using audio-visual information for urdu language. *arXiv preprint arXiv:1802.05521*, 2018.

[45] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth. The world of emotions is not two-dimensional. *Psychological Science*, 18:1050–1057, 2007.

[46] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller. Audeep: Unsupervised learning of representations from audio with deep recurrent neural networks. *J. Mach. Learn. Res.*, 18(1):6340–6344, 2017.

[47] Y. Fu, T. Xiang, Y. G. Jiang, X. Xue, L. Sigal, and S. Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35:112–125, 2018.

[48] R. Girdhar and D. Ramanan. Attentional pooling for action recognition. In *Proc. Advances in neural information processing systems (NIPS)*, pages 34–45, 2017.

[49] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

[50] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer. Technique for automatic emotion recognition by body gesture analysis. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–6. IEEE, 2008.

[51] D. Grandjean, D. Sander, and K. Scherer. Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization. *Consciousness and Cognition*, 17:484–495, 2008.

[52] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *Proc. International Conference on Machine Learning (ICML)*, volume 14, pages 1764–1772, 2014.

[53] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic. Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proc. ACM Association for Computational Linguistics (ACL)*, pages 1–11, 2018.

[54] P. Gupta and N. Rajput. Two-stream emotion recognition for call center monitoring. In *Proc. Conference of the International Speech Communication Association*, 2007.

[55] J. Han, Z. Zhang, N. Cummins, F. Ringeval, and B. Schuller. Strength modelling for real-world automatic continuous affect recognition from audiovisual signals. *Image and Vision Computing*, 65:76–86, 2017.

[56] J. Han, Z. Zhang, F. Ringeval, and B. Schuller. Prediction-based learning for continuous emotion recognition in speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5005–5009, 2017.

[57] J. Han, Z. Zhang, F. Ringeval, and B. Schuller. Reconstruction-error-based learning for continuous emotion recognition in speech. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2367–2371, 2017.

[58] S. Hantke, M. Schmitt, P. Tzirakis, and B. Schuller. Eat–the icmi 2018 eating analysis and tracking challenge. 2018.

[59] S. Hantke, F. Weninger, R. Kurle, F. Ringeval, A. Batliner, A. El-Desoky Mousa, and B. Schuller. I Hear You Eat and Speak: Automatic Recognition of Eating Condition and Food Types, Use-Cases, and Impact on ASR Performance. *PLoS ONE*, 11:1–24, 2016.

[60] D. Hazarika, S. Gorantla, S. Poria, and R. Zimmermann. Self-attentive feature-level fusion for multimodal emotion detection. In *Proc. IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 196–201, 2018.

[61] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann. Icon: interactive conversational memory network for multimodal emotion detection.

In *Proc. ACM Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2594–2604, 2018.

[62] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[63] G. Hinton, L. Deng, Y. Dong, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. *Signal Processing Magazine*, 29(6):82–97, 2012.

[64] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[65] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 328–339, 2018.

[66] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017.

[67] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[68] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, and J. Yi. Continuous multimodal emotion prediction based on long short term memory recurrent neural network. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 11–18, 2017.

[69] Y. Huang and H. Lu. Deep learning driven hypergraph representation for image-based emotion recognition. In *Proc. International Conference on Multimodal Interaction*, pages 243–247, 2016.

[70] Z. Huang, B. Stasak, T. Dang, K. W. Gamage, P. Le, V. Sethu, and J. Epps. Staircase regression in oa rvm, data selection and gender dependency in avec 2016. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 19–26, 2016.

[71] G. L. Huttar. Relations between prosodic variables and emotions in normal American English utterances. *Journal of Speech, Language and Hearing Research*, 11:481–487, 1968.

[72] Q. Ji, Z. Zhu, and P. Lan. Real-time nonintrusive monitoring and prediction of driver fatigue. *IEEE Transactions on Vehicular Technology*, 53(4):1052–1068, 2004.

[73] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai. Emotional audio-visual speech synthesis based on PAD. *Transactions on Audio, Speech and Language Processing*, 19:570–582, 2011.

[74] L. Jing and Y. Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020.

[75] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels. Ensemble methods for continuous affect recognition: Multimodality, temporality and challenges. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 9–16, 2015.

[76] S. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, et al. Emon-

ets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.

[77] S. Khorram, M. McInnis, and E. Mower Provost. Jointly Aligning and Predicting Continuous Emotion Annotations. *IEEE Transactions on Affective Computing*, pages 1–1, 2019.

[78] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *IEEE transactions on pattern analysis and machine intelligence*, 30(12):2067–2083, 2008.

[79] Y. Kim, Y. Jernite, D. Sontag, and A. Rush. Character-aware neural language models. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, pages 2741–2749, 2016.

[80] Y. Kim, H. Lee, and E. Provost. Deep learning for robust feature generation in audiovisual emotion recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3687–3691, 2013.

[81] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[82] R. Kiros, R. Salakhutdinov, and R. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[83] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) workshops*, pages 1972–1979, 2017.

[84] D. Kollias, P. Tzirakis, M. A. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou. Deep affect prediction in-the-wild: Aff-wild

database and challenge, deep architectures and beyond. *International Journal of Computer Vision*, 127:907–929, 2019.

[85] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[86] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. Albert: A lite bert for self-supervised learning of language representations. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

[87] I. Lawrence and K. Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45:255–268, 1989.

[88] Y. Lecun. Generalization and network design strategies. In *Connectionism in perspective*. 1989.

[89] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, pages 6765–6816, 2017.

[90] S. Li and W. Deng. Deep facial expression recognition: A survey. *Transactions on Affective Computing*, 2020.

[91] Y. Li, T. Zhao, and T. Kawahara. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. *Proc. Interspeech 2019*, pages 2803–2807, 2019.

[92] W. Liu, W.-L. Zheng, and B.-L. Lu. Emotion recognition using multimodal deep learning. In *Proc. International Conference on Neural Information Processing*, pages 521–529, 2016.

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

[94] Y. Liu, C. Sun, L. Lin, and X. Wang. Learning natural language inference using bidirectional LSTM model and inner-attention. *arXiv preprint arXiv:1605.09090*, 2016.

[95] L.-S. Low, N. Maddage, M. Lech, L. Sheeber, and N. Allen. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering*, 58(3):574–586, 2010.

[96] M. T. Luong, H. Pham, and C. Manning. Effective approaches to attention-based neural machine translation. In *Proc. Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

[97] S. Mai, H. Hu, and S. Xing. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Proc. ACM Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 481–492, 2019.

[98] S. Mariooryad and C. Busso. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 6(2):97–108, 2015.

[99] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations Between a Person and a Limited Agent. *Transactions on Affective Computing*, 3:5–17, 2012.

[100] S. Minaee and A. Abdolrashidi. Deep-emotion: Facial expression recognition using attentional convolutional network. *arXiv preprint arXiv:1902.01019*, 2019.

[101] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao. Deep learning based text classification: A comprehensive review. *arXiv preprint arXiv:2004.03705*, 2020.

[102] T. M. Moerland, J. Broekens, and C. M. Jonker. Emotion in reinforcement learning agents and robots: a survey. *Machine Learning*, 107:443–480, 2018.

[103] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4293–4302, 2016.

[104] A. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE Access*, 7:19143–19165, 2019.

[105] M. Neumann and N. Vu. Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length and acted speech. *arXiv preprint arXiv:1706.00612*, 2017.

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

[107] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *Proc. British Machine Vision Conference (BMVC)*, page 6, 2015.

[108] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Advances in neural information processing systems (NIPS)*, pages 8026–8037, 2019.

[109] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition. In *Proc. IEEE International Con-*

ference on Acoustics, Speech and Signal Processing (ICASSP), pages 6548–6552, 2018.

[110] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, and A. Hussain. Multimodal sentiment analysis: Addressing key issues and setting up the baselines. *IEEE Intelligent Systems*, 33(6):17–25, 2018.

[111] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel. Multimodal emotion recognition for avec 2016 challenge. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 75–82, 2016.

[112] D. Priyasad, T. Fernando, S. Denman, S. Sridharan, and C. Fookes. Attention driven fusion for multi-modal emotion recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3227–3231, 2020.

[113] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller. Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, 66:22–30, 2015.

[114] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic. Avec 2015: The 5th international audio/visual emotion challenge and workshop. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 1335–1336, 2015.

[115] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmi, and M. Pantic. Avec 2017–real-life depression and affect recognition workshop and challenge. In *Proc. International Workshop on Audio/Visual Emotion Challenge, ACM Multimedia*, pages 3–9, 2017.

[116] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic. AV+EC 2015 – The First Affect Recognition Challenge Bridging Across Audio, Video and Physiological Data. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 3–8, 2015.

[117] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne. Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proc. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2013.

[118] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252, 2015.

[119] J. Russell. A Circular Model of Affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[120] P. Sarkar and A. Etemad. Self-supervised learning for ecg-based emotion recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3217–3221, 2020.

[121] K. Sarma, Y. Liang, and A. Sethares. Domain adapted word embeddings for improved sentiment classification. In *Proc. ACM Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 37–42, 2018.

[122] K. R. Scherer. On the nature and function of emotion: A component process approach. *Approaches to emotion*, 2293:317, 1984.

[123] M. Schmitt, F. Ringeval, and B. Schuller. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In *Proc.*

Annual Conference of the International Speech Communication Association, pages 495–499, 2016.

[124] M. Schmitt and B. Schuller. Openxbow: Introducing the passau open-source crossmodal bag-of-words toolkit. *J. Mach. Learn. Res.*, 18(1):3370–3374, 2017.

[125] B. Schuller and A. Batliner. *Computational paralinguistics: emotion, affect and personality in speech and language processing*. John Wiley & Sons, 2013.

[126] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages II–1, 2003.

[127] B. Schuller, S. Steidl, A. Batliner, E. Bergelson, J. Krajewski, C. Janott, A. Amatuni, M. Casillas, A. Seidl, M. Soderstrom, et al. The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In *Proc. Interspeech, Computational Paralinguistics Challenge (ComParE)*, pages 3442–3446, 2017.

[128] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH*, pages 2794–2797, 2010.

[129] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan. Paralinguistics in speech and language—state-of-the-art and the challenge. *Computer Speech & Language*, 27(1):4–39, 2013.

[130] B. Schuller, S. Steidl, A. Batliner, S. Hantke, F. Hönig, J. R. Orozco-Arroyave, E. Nöth, Y. Zhang, and F. Weninger. The INTERSPEECH 2015 computational paralinguistics challenge: Nativeness, parkinson's & eating condition. In *Proc. of INTERSPEECH*, pages 478–482, Dresden, Germany, 2015.

[131] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E. Rathner, K. Bartl-Pokorny, et al. The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats. In *Proc. Interspeech, Computational Paralinguistics Challenge (ComParE)*, pages 122–126, 2018.

[132] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011–the first international audio/visual emotion challenge. *Affective Computing and Intelligent Interaction*, 6975:415–424, 2011.

[133] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen, S. Ottl, M. Gerczuk, P. Tzirakis, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, J. L. Rothkrantz, J. Zwerts, J. Treep, and C. Kaandorp. The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates. *arXiv preprint arXiv:2102.13468*, 2021.

[134] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine. Time-contrastive networks: Self-supervised learning from video. In *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, pages 1134–1141, 2018.

[135] N. Singh, N. Singh, and A. Dhall. Continuous Multimodal Emotion Recognition Approach for AVEC 2017. *arXiv preprint arXiv:1709.05861*, 2017.

[136] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty. The belfast induced natural emotion database. *Transactions on Affective Computing*, 3:32–41, 2012.

[137] R. Socher, J. Pennington, E. Huang, A. Ng, and C. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proc. con-*

ference on empirical methods in natural language processing, pages 151–161, 2011.

[138] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proc. ACM Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, 2013.

[139] K. Somandepalli, E. Gupta, M. Nasir, B. M. Booth, S. Lee, and S. Narayanan. Online affect tracking with multimodal kalman filters. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 59–66, 2016.

[140] K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Proc. Advances in neural information processing systems (NeurIPS)*, pages 2377–2385, 2015.

[141] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[142] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris. MuSe 2020 Challenge and Workshop: Multimodal Sentiment Analysis, Emotion-Target Engagement and Trustworthiness Detection in Real-Life Media: Emotional Car Reviews in-the-Wild. In *Proc. ACM Multimedia, International on Multimodal Sentiment Analysis in Real-Life Media Challenge and Workshop*, page 35–44. Association for Computing Machinery, 2020.

[143] L. Stappen, A. Baird, L. Schumann, and B. Schuller. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *arXiv preprint arXiv:2101.06053*, 2021.

[144] L. Stappen, X. Du, V. Karas, S. Müller, and B. Schuller. Go-card - generic, optical car part recognition and detection: Collection, insights, and applications. *ArXiv*, abs/2006.08521, 2020.

[145] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. Deep neural networks for acoustic emotion recognition: raising the benchmarks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5688–5691, 2011.

[146] B. Sun, S. Cao, L. Li, J. He, and L. Yu. Exploring multimodal visual features for continuous affect recognition. In *Proc. International Workshop on Audio/Visual Emotion Challenge*, pages 83–88, October 2016.

[147] F. Tao and C. Busso. Aligning audiovisual features for audiovisual speech recognition. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2018.

[148] V. A. Tatsis, C. Tjortjis, and P. Tzirakis. Evaluating data mining algorithms using molecular dynamics trajectories. *International journal of data mining and bioinformatics*, 8:169–187, 2013.

[149] H. N. Tran and E. Cambria. Ensemble application of ELM and GPU for real-time multimodal sentiment analysis. *Memetic Computing*, 10(1):3–13, 2018.

[150] A. Triantafyllopoulos, G. Keren, J. Wagner, I. Steiner, and B. Schuller. Towards robust speech emotion recognition using deep residual networks for speech enhancement. In *Proc. Interspeech*, pages 1691–1695, 2019.

[151] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, 2016.

[152] Y. H. Tsai, S. Bai, P. Liang, Z. Kolter, L. P. Morency, and R. Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proc. ACM Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6558–6569, 2019.

[153] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, July 2019.

[154] P. Tzirakis, D. Boros, E. Hajiyev, and B. W. Schuller. Facial emotion recognition using deep residual networks in real-world environments. *arXiv preprint arXiv:2111.02717*, 2021.

[155] P. Tzirakis, A. Kumar, and J. Donley. Multi-channel speech enhancement using graph neural networks. *arXiv preprint arXiv:2102.06934*, 2021.

[156] P. Tzirakis, A. Papaioannou, A. Lattas, M. Tarasiou, B. Schuller, and S. Zafeiriou. Synthesising 3d facial motion from" in-the-wild" speech. In *Proc. IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 627–634.

[157] P. Tzirakis, A. Shiarella, R. Ewers, and B. Schuller. Computer Audition for Continuous Rainforest Occupancy Monitoring: The Case of Bornean Gibbons' Call Detection. *Proc. Interspeech 2020*, pages 1211–1215, 2020.

[158] P. Tzirakis and C. Tjortjis. T3C: improving a decision tree classification algorithm's interval splits on continuous attributes. *Advances in Data Analysis and Classification*, 11:353–370, 2017.

[159] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. End-to-End Multimodal Emotion Recognition using Deep Neural Networks. *Journal of Selected Topics in Signal Processing*, 11:1301–1309, 2017.

[160] P. Tzirakis, S. Zafeiriou, and B. W. Schuller. End2You–The Imperial Toolkit for Multimodal Profiling by End-to-End Learning. *arXiv preprint arXiv:1802.01115*, 2018.

[161] P. Tzirakis, J. Zhang, and B. Schuller. End-to-end speech emotion recognition using deep neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5200–5204, 2018.

[162] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016: Depression, mood and emotion recognition workshop and challenge. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 3–10, 2016.

[163] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. Advances in neural information processing systems (NeurIPS)*, pages 5998–6008, 2017.

[164] J. Wagner, J. Kim, and E. André. From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification. In *2005 IEEE international conference on multimedia and expo*, pages 940–943. IEEE, 2005.

[165] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[166] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7794–7803, 2018.

[167] Y. Wang, C. Weth, Y. Zhang, K. H. Low, V. K. Singh, and M. Kankanhalli. Proceedings of the concept based hybrid fusion of multimodal event signals. In *Proc. International Symposium on Multimedia*, pages 14–19, 2016.

[168] N. Watanabe and M. Haruno. Effects of subconscious and conscious emotions on human cue–reward association learning. *Scientific reports*, 5(1):1–6, 2015.

[169] R. Weber, V. Barrielle, C. Soladié, and R. Séguier. High-level geometry-based features of video modality for emotion prediction. In *Proc. ACM Multimedia, International Workshop on Audio/Visual Emotion Challenge (AVEC)*, pages 51–58, 2016.

[170] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie. Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies. In *Proc. Annual Conference of the International Speech Communication Association*, pages 597–600, 2008.

[171] C. H. Wu, J. C. Lin, and W. L. Wei. Survey on audiovisual emotion recognition: databases, features and data fusion strategies. *Transactions on Signal and Information Processing*, 3:572–587, 2014.

[172] M. Xu, F. Zhang, and S. U. Khan. Improve accuracy of speech emotion recognition with attention head fusion. In *Proc. IEEE Annual Computing and Communication Workshop and Conference (CCWC)*, pages 1058–1064, 2020.

[173] Y. Xu, H. Xu, and J. Zou. Hgfm: A hierarchical grained and feature model for acoustic emotion recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6499–6503, 2020.

[174] H. Yang, U. Ciftci, and L. Yin. Facial expression recognition by de-expression residue learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2168–2177, 2018.

[175] Q. Yang, Y. Liu, T. Chen, and Y. Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10:1–19, 2019.

[176] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. Advances in neural information processing systems (NeurIPS)*, pages 5754–5764, 2019.

[177] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proc. conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[178] S. Yoon, S. Byun, and K. Jung. Multimodal speech emotion recognition using audio and text. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118, 2018.

[179] W. J. Yoon, Y. H. Cho, and K. S. Park. A study of speech emotion recognition and its application to mobile services. In *Proc. IEEE International Conference on Ubiquitous Intelligence and Computing*, pages 758–766, 2007.

[180] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[181] S. Zhang, S. Zhang, T. Huang, and W. Gao. Multimodal deep convolutional neural network for audio-visual emotion recognition. In *Proc. International Conference on Multimedia Retrieval (ICMR)*, pages 281–284, 2016.

[182] Z. Zhang, E. Coutinho, J. Deng, and B. Schuller. Cooperative learning and its application to emotion recognition from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):115–126, 2014.

[183] W. Zheng, J. Yu, and Y. Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *Proc. IEEE Affective Computing and Intelligent Interaction (ACII)*, pages 827–831, 2015.

[184] X. Zhou, X. Wan, and J. Xiao. Attention-based lstm network for cross-lingual sentiment classification. In *Proc. conference on empirical methods in natural language processing*, pages 247–256, 2016.