Continuous Emotion Recognition in Dyadic Interactions

by

Syeda Narjis Fatima

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Department of Electrical Engineering

KOÇ ÜNİVERSİTESİ

June, 2020

Continuous Emotion Recognition in Dyadic Interactions

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Syeda Narjis Fatima

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Engin Erzin

_____

Prof. Yücel Yemez

_____

Prof. A. Murat Tekalp

_____

Prof. Çiğdem Eroğlu Erdem

_____

Prof. Levent Arslan

Date: _____

To my family!

ABSTRACT

Continuous Emotion Recognition in Dyadic Interactions
Syeda Narjis Fatima
Doctor of Philosophy in Department of Electrical Engineering
June, 2020

Understanding emotional dynamics of dyadic interactions is crucial for developing more natural human-computer interaction (HCI) systems. Emotional dependencies and affect context play important roles in dyadic interactions. The emotional state of a participant is modulated in many communication channels such as speech, head and body motion, vocal activity patterns as well as non-verbal vocalizations of speech during dyadic interactions. Recent studies have shown that affect recognition tasks can benefit by the incorporation of a particular interaction's context, however, particularly the investigation of the role and contribution of affect context and its incorporation into dyadic neural architectures remains a complex and open problem. Our work takes motivation from this perspective and therefore, in this thesis, a series of related studies targeting emotional dependencies during dyadic interactions are conducted to improve continuous emotion recognition (CER). Firstly, we define a convolutional neural network (CNN) architecture for single-subject CER based on speech and body motion data. We then introduce dyadic CER as a two-stage regression framework and explore ways in which cross-subject affect can be used to improve CER performance for a target subject. Specifically, we propose two dyadic CNN architectures where cross-subject contribution to the CER task is achieved by fusion of cross-subject affect and feature maps. As a conclusive work, we define dyadic affect context (DAC) and propose a new Convolutional LSTM (ConvLSTM) model that exploits it for dyadic CER. Our ConvLSTM model captures local spectro-temporal correlations in speech and body motion as well as the long-term affect inter-dependencies between subjects. Our multimodal analysis demonstrates that modelling and incorporation of the DAC in the proposed CER models provide significant performance improvements on the USC CreativeIT database and the achieved results compare favorably to the state-of-the-art.

ÖZETÇE

İkili Etkileşimde Sürekli Duygu Tanıma
Syeda Narjis Fatima
Elektrik ve Elektronik Mühendisliği, Doktora
Haziran, 2020

İkili etkileşimlerin duygusal dinamiklerini anlamak, daha doğal insan-bilgisayar etkileşimi (İBE) sistemleri geliştirmek için çok önemlidir. Duygusal devinimler ve duygu durum bağlamı ikili etkileşimlerde önemli rol oynamaktadır. İkili etkileşimde bir katılımcının duygusal durumu, konuşma, kafa ve vücut hareketi, vokal aktiviteler ve konuşmanın sözel olmayan etkenleri gibi birçok iletişim kanalında modüle edilir. Yakın zamanlarda yapılan çalışmalar, duygu durum tanıma için etkileşim bağlamının kullanılmasının faydalı olabileceğini göstermiştir, ancak özellikle duygu bağlamının rolü ve katkısının araştırılması ve derin sinir ağları mimarilerine dahil edilmesi karmaşık ve açık bir problem olmaya devam etmektedir. Tez çalışmamız motivasyonu bu perspektiften alarak sürekli duygu tanıma (SDT) problemine bir çözüm için ikili etkileşimler sırasında duygusal etkileşimlerin kullanılmasını hedefleyen bir dizi çalışma içermektedir. İlk olarak, tek kişili SDT için konuşma ve beden hareket verilerine dayalı evrişimsel bir sinir ağı (CNN) mimarisi tanımladık. Daha sonra iki aşamalı bir kestirici ile ikili etkileşim için bir SDT sistemi tanımladık ve etkileşimdeki bir kişinin (çapraz-kişi) duygu durumunun diğer kişinin SDT başarımını artırmak için nasıl kullanılabileceğini araştırdık. Özellikle, SDT için çapraz-kişinin katkısını duygu durum ve öznitelik haritalarının kaynaşmasıyla sağladığımız iki CNN mimarisi önerdik. Bu çalışmaların devamında, ikili-etkileşim duygu bağlamını (İDB) tanımladık ve ikili-etkileşimde SDT için yeni bir Evrişimsel Uzun-Kısa Süreli Bellek (ConvLSTM) modeli önerdik. ConvLSTM modelimiz, konuşma ve beden hareketlerinde izgesel-zamansal ilintileri ve kişiler arası uzun vadeli duygu durum bağımlılıklarını yakalar. Çok kipli analiz modellerimizde İDB bilgisinin önerilen SDT yaklaşımlarına dahil edilmesinin USC CreativeIT veritabanında önemli başarım iyileştirmeleri sağladığını ve elde edilen sonuçların literatürdeki en son başarımlar uyumlu olduğunu göstermektedir.

ACKNOWLEDGMENTS

I would like to express my gratitudes and sincere thanks to my advisor, Prof. Engin Erzin, who led me into this exciting area of continuous emotion recognition. I am indebted to Prof. Engin for his continued support, guidance, patience and understanding. I would like to express my sincere thanks to Prof. Yücel Yemez and Prof. Çiğdem Eroğlu Erdem for accepting to supervise this thesis and to drive it further with their valuable comments throughout the process.

TABLE OF CONTENTS

List of Figures **x**

List of Tables **xii**

Chapter 1: Introduction **1**

1.1 Motivation 1

1.2 Research Problem 2

1.3 Thesis Contributions 3

1.4 List of Publications 4

1.5 Report Outline 4

Chapter 2: Literature on Continuous Emotion Recognition (CER) **6**

2.1 CER from Speech 8

2.2 Other Modalities 8

2.3 Multimodal CER 9

2.4 Generative Models for CER 9

2.5 Neural Networks for CER 10

2.6 Dyadic CER Studies 12

Chapter 3: CER Resources **15**

3.1 USC CreativeIT Database 15

3.2 IEMOCAP Database 18

Chapter 4: Acoustic Feature Selection for Natural Dialog Identification **20**

4.1 Introduction 20

4.2 Methodology . . . 21
4.2.1 Segment Construction . . . 22
4.2.2 Feature Extraction and Processing . . . 23
4.2.3 Similarity based Feature Selection . . . 25
4.3 Experimental Evaluations . . . 26
4.3.1 Feature Selection Evaluations . . . 28

Chapter 5: Dyadic Cross-Subject CER 31
5.1 Introduction . . . 31
5.2 Methodology . . . 31
5.2.1 Feature extraction and window-level summarization . . . 32
5.2.2 Cross-subject continuous emotion recognition (CSCER) . . . 34
5.3 Experimental Evaluations . . . 35

Chapter 6: Non-Verbal Vocalizations for CER 41
6.1 Introduction . . . 41
6.2 Methodology . . . 42
6.2.1 Feature Extraction . . . 43
6.2.2 Feature Summarization . . . 44
6.2.3 NVV Histogram Features . . . 44
6.2.4 Continuous emotion recognition (CER) . . . 45
6.3 Experimental evaluations . . . 46

Chapter 7: Dyadic Affect Context for CER 51
7.1 Introduction . . . 51
7.2 Methodology . . . 52
7.2.1 Overview . . . 52
7.2.2 Single-Subject CNN Architecture . . . 53
7.2.3 Dyadic CNN Architectures . . . 55
7.2.4 Dyadic ConvLSTM Architecture . . . 58

7.3 Experimental Results 62
7.3.1 Feature Representations 63
7.3.2 Experimental Setup 63
7.3.3 Single-Subject CNN Architecture Results 64
7.3.4 Dyadic CNN Architecture Results 66
7.3.5 Dyadic ConvLSTM Architecture Results 66

Chapter 8: Conclusions and Future Work **70**

Bibliography **73**

LIST OF FIGURES

4.1 Identification of natural dialogs based on acoustic feature selection. . 21
4.2 Natural segment construction. 23
4.3 Feature set selection based on similarity and distance measures 25
4.4 Normalized feature-wise percentage retention breakdown of z-normalized features at session-level as a function of overall percentage of features retained under four categories: MFCCs [0-14], Loudness, Fundamental frequency, Probability of voicing. 29
4.5 Retention percentage of each z-normalized statistical functional at session-level as a function of the overall percentage of features retained. Note: functional indexes (x-axis) are in the order of their description in Section 4.2.2. 30

5.1 Cross-subject continuous emotion recognition (CSCER) framework. . 32
5.2 Sample comparison of (i) ground truth for Activation with estimation using (ii) modality data only, (c) modality data and true side information, and (d) modality data and estimated side information. . 39

6.1 Framework for continuous emotion recognition using histograms of non-verbal vocalizations. 42
6.2 CNN architecture for the proposed CER system. 46

7.1 Block diagram of a general dyadic two-stage CER framework. 52
7.2 CER-CNN - Single-subject continuous emotion recognition CNN architecture. 54

7.3 Block diagram of the dyadic CNN architecture for the fusion of cross-subject affect (FoA). 56
7.4 Block diagram of the dyadic CNN architecture for the fusion of cross-subject feature maps (FoM). 57
7.5 The proposed CER-ConvLSTM framework based on the FoA model for the prediction of activation attribute in dyadic interactions. 60
7.6 Organization of input to CER-ConvLSTM. 61
7.7 ConvLSTM architecture at *Stage 2*. 61
7.8 PCC Performance of multimodal AVD prediction as a function of context duration ($R \times (K-1) \times \Delta/60$) secs. 69

LIST OF TABLES

2.1 Summary of related work in CER. Some of the abbreviations include: CER Task column refers to affect dimensions (A)rousal/Activation, (V)alence, (P)ower, (E)xpectation, and (D)ominance; Modality column refers to Speech (S), speech-video (SV), body motion (BM); Evaluation Criteria column refers to weighted accuracy (WA), Pearson's correlation coefficient (PCC), Concordance correlation coefficient (CCC), Spearman's correlation coefficient (SCC) 7

3.1 Session-wise distribution of USC CreativeIT recordings within the working dataset. 17

4.1 Session level distribution of natural and artificial segments within the IEMOCAP database. 27

4.2 Average accuracy rates (%) on global and session-level scopes of z-normalized and MM-normalized features without feature selection (baseline exhaustive set). 27

4.3 Average percentage classification accuracy with KLD, Pearson correlation and Hellinger distance based feature selection schemes as a function of percentage feature retention (FR) of session-level z-normalized features 28

5.1 Cross-subject inter-attribute Pearson's correlation based on ground truth activation(A), valence(V) and dominance(D) attributes. 34

5.2 Cross-subject continuous emotion recognition (CSCER) at the window-level of AVD with ground truths as true side emotional information based on speech, motion and multimodal cues. We present mean absolute correlation values between estimated emotional curve and the ground truth. 37
5.3 CSCER at the window-level of activation, valence, dominance with estimated ground truths (Stage1) as side emotional information based on speech and motion cues. We present mean absolute correlation values between estimated emotional curve and the ground truth. . . . 38

6.1 Distribution of non-verbal vocalizations across five sessions of the working USC CreativeIT database. 43
6.2 The mean absolute correlation values between estimated emotional attributes and the ground truth for the GMR-based CER regression framework with MFCC and *eGemaps* features for speech, head motion features and NVV histogram features. 48
6.3 The mean absolute correlation values between estimated emotional attributes and the ground truth for the proposed CER system using CNN-based regression with *MFCC* features for speech, head motion features and NVV histogram features. 49

7.1 Cross-subject correlations (PCC) of the AVD in dyadic settings. . . . 62
7.2 PCC performance of the unimodal and multimodal AVD prediction with the Single-Subject CNN architecture (CER-CNN) for varying temporal window size N. 64
7.3 PCC performance of the multimodal activation prediction with the Single-Subject CNN architecture (CER-CNN) for varying batch sizes (temporal window size is fixed to $N = 40$). 65

7.4 PCC performance of the multimodal activation attribute prediction with the Single-Subject CNN architecture (CER-CNN) for various activation functions. 65
7.5 PCC performance of the unimodal and multimodal CER with the dyadic FoA model for the AVD attributes. 66
7.6 PCC performance of the unimodal and multimodal CER with the dyadic FoM model for the AVD attributes for varying max-pooling factor P_F. 67
7.7 PCC performance of the multimodal AVD prediciton with the Dyadic CER-ConvLSTM architecture for varying values of the number of feature images K and the context stride Δ. 68

Nomenclature

A Activation

CCA Canonical correlation analysis

CER Continuous emotion recognition

CER-CNN Single-subject CNN architecture

CNN Convolutional neural network

ConvLSTM Convolutional long short-term memory

CSCER Cross-subject continuous emotion recognition

D Dominance

FoA Fusion of cross-subject affect

FoM Fusion of cross-subject feature maps

FR Feature retention

GMR Gaussian mixture regression

HCI Human-computer interfaces

LOSO Leave-one-session-out cross-validation

LSTM Long-short term memory neural network

MFCC mel frequency cepstral coefficients

MM Min-max normalization

MSE Mean square error

NVV Non-verbal vocalizations

PCA Principal Component Analysis

SVM Support vector machine

V Valence

Chapter 1

INTRODUCTION

1.1 Motivation

Understanding and modeling of dyadic affective interactions are crucial in realizing intelligent human-computer interfaces (HCIs). There is a deep interest in benefiting from socially and emotionally intelligent HCIs in applications such as human-robot communication, call center dialog systems, computer games, and conversational agents [1].

Affective human interactions are well-defined psychological paradigms that reflect a multitude of human behaviors [2, 3]. Emotions are an inextricable constituent of human interactions. A common continuous representation of emotion uses the three principle dimensions of affect known as *Activation* (active - passive), *Valence* (pleasant - unpleasant) and *Dominance* (dominant - submissive) [4, 5]. Collectively, the three attributes summarize the global emotional state of a subject. Understanding emotion is crucial in development of emotionally intelligent machines. Research community around the world is increasingly targeting the realization of systems that are capable of automatically interpreting, understanding and responding to emotions and moods displayed by humans, which have various applications in very diverse areas such as HCI, security, healthcare and education. One way at advancing affect analysis requires attention to investigate how individuals in dyadic settings or a group influence the affective states of each other [6, 7].

Specifically, with regard to dyadic interactions, i.e. interactions between two participants, behaviours of individual participants are strongly influenced by their conversational partner and thus still require analysis of both participants to measure

and study crucial inter-dependency patterns within different data spaces. Moreover, since there is a known dependency between the participants of a conversation, devising dyadic architectures that are able to capture dyadic exchanges between interlocutors can be particularly breakthroughs. Our work thus takes motivation from these perspectives and we focus on the problem of continuous emotion recognition (CER) so that it can be used by HCI while being online with an emphasis on dyadic partners.

1.2 Research Problem

Based on the motivations presented, the major research questions this thesis aims to answers are:

- What acoustic features contribute towards the naturalness of the speech? We believe that since these features represent the naturalness of speech, they will be important in establishing the underlying characteristics of human conversations such as synchrony (interdependence of dyadic partners).

- How can we utilize cross-subject and cross-emotion relationships across emotional attributes in a dyadic interaction to improve CER? The existence of these relationships will validate the presence of dyadic synchrony as a natural inherent phenomenon in dyadic human interactions and this synchrony can then be used to improve CER.

- Do inter-emotional dependencies also reflect themselves in different modalities? If the answer to this question is in the positive, we may also incorporate the predicted emotion estimates learnt from different modalities into data fusion setups.

- Whether the hypothesis that participants in dyadic interactions exhibit higher emotion correlation around occurrences of non-verbal vocalizations as compared to verbal exchange during interactions is true? These occurences can then be considered as salient points where high synchrony is exhibited.

- Given that synchrony exists in dyadic conversations, how can the context of a dyadic conversation be quantified and how can this context be used to improve the CER estimation? Can neural network architectures be used to improve the emotion estimation? How do various architectures compare in tackling the CER problem in dyadic settings?

1.3 Thesis Contributions

In answering the research questions in the previous section, we made the following key contributions:

- We provide a definition for conversational context and quantify it for use in improvement of emotion recognition [8]. We introduce CNN-based affect and feature-map fusion models in dyadic settings for the proposed two-stage framework and demonstrate that these fusion models hold potential to further improve CER performance. As a final conclusive study, we propose a new convolutional long short-term memory (ConvLSTM) architecture in dyadic settings and show that exploiting dyadic affect context across subjects holds significant potential to further improve continuous emotion recognition. We propose first-ever dyadic CNN structures capturing dyadic emotional and context dependencies of interlocutors in the ConvLSTM frameworks.

- We conceptualize two-stage continuous emotion recognition frameworks where the emotion estimation from first stage is improved in a second-stage re-estimation [7]. We applied this approach on a number of different modalities with Gaussian mixture regression (GMR) and convolutional neural network (CNN) based estimation methods. Our work shows that cross-speaker dependencies can be parameterized to achieve realistic cross-speaker emotion regression in a dyadic setting based on acoustic and body motion modalities. In a related study we also observe that apart from cross-speaker dependencies, we also show that cross-modality relationships can also be used to improve the performance of CER.

- We identified a set of acoustic features that help in the discriminating natural and artificial conversations [9].

- We show that non-verbal vocalizations are areas of speech which have higher synchrony in comparison to other regions of speech and they can be used to improve CER [10].

1.4 List of Publications

- S. N. Fatima, E. Erzin, "Acoustic Feature Selection for Natural Dialog Identification," in 3rd International Workshop on Affective Social Multimedia Computing (ASMMC), 2017.

- S. N. Fatima, E. Erzin, "Cross-Subject Continuous Emotion Recognition Using Speech and Body Motion in Dyadic Interactions," in INTERSPEECH, pp. 1731-1735, 2017.

- S. N. Fatima, E. Erzin, "Use of non-verbal vocalizations for continuous emotion recognition from speech and head motion," in 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 433-437, 2019.

- S. N. Fatima, E. Erzin, "Use of Affect Context for Continuous Emotion Recognition in Dyadic Interactions," submitted to Speech Communication, 2019.

1.5 Report Outline

The rest of this thesis report is organized as follows. In chapter 2, we provide a detailed literature review of emotion recognition. The major datasets used for our various studies are described in chapter 3. Our study on identification of acoustic feature selection for classification of natural and artificial speech is described in chapter 4. Cross-subject interdependencies between participants of a dyadic conversation are exploited by means of 2-stage GMR-based regression networks in chapters 5. Chapter 6 describes the use of non-verbal vocalizations for improvement of CER

on CNN-based networks. In Chapter 7 we quantify affect context and use various data fusion methods to improve emotion recognition while also exploring the use of ConvLSTM network architecture. The thesis is finally concluded in chapter 8.

Chapter 2

LITERATURE ON CONTINUOUS EMOTION RECOGNITION (CER)

Affective human interactions are reflective of a multitude of human behaviors and interaction cues [2, 3]. Owing to emotion being an inherent state of human response, understanding emotion is crucial in development of emotionally intelligent machines, better emotion prediction in speech synthesis, and tracking of emotional responses in social scenarios such as dyadic interactions and counseling.

Continuous emotion recognition (CER) is a sub-field of affective computing wherein the emotional state of the target subject is estimated based on many data modalities typically including speech, video and body motion cues. CER increasingly attracted interest over the past few years [11, 12]. Speech and facial information are extensively used for CER [13, 14, 1] based on a wide variety of techniques such as k-means clustering, support vector regression, neural networks and Gaussian mixture model (GMM) regression [15, 16, 17, 18, 19]. Since CER is the most relevant to our work, we present a summary of some key research efforts most related to our work in Table 2.1. We discuss some of the major dimensions of this topic in the subsections below.

Table 2.1: Summary of related work in CER. Some of the abbreviations include: CER Task column refers to affect dimensions (A)rousal/Activation, (V)alence, (P)ower, (E)xpectation, and (D)ominance; Modality column refers to Speech (S), speech-video (SV), body motion (BM); Evaluation Criteria column refers to weighted accuracy (WA), Pearson's correlation coefficient (PCC), Concordance correlation coefficient (CCC), Spearman's correlation coefficient (SCC)

Ref.	**Year**	**CER Task**	**Modality**	**Model**	**Database**	**Database Size (hrs/speakers)**	**Evaluation Criteria**	**CER Performance**
[20]	2012	AVPE	S	SVR	SEMAINE	3.7 / 24	PCC	0.014 / 0.040 / 0.016 / 0.038
[16]	2013	AVD	S-BM	GMR	CreativeIT	∼5 / 16	PCC	0.5979 / 0.2247 / 0.3696
[21]	2013	AV	S	SVR	AViD	240 / 292	PCC	0.09 / 0.09
[22]	2015	AV	S	SVR	RECOLA	2.3 / 27	CCC	0.23 / 0.07
[23]	2016	AV	S	SVR	RECOLA	2.3 / 27	CCC	0.65 / 0.38
[23]	2016	AV	S	CRNN	RECOLA	2.3 / 27	CCC	0.69 / 0.26
[24]	2017	AV	S	SVR	SEWA	3 / 64	CCC	0.23 / 0.24
[25]	2017	AVD	S-BM	SVR	JestKOD	4.31 / 10	PCC	0.35 / 0.06 / 0.53
[26]	2017	AV	SV	SVR	CreativeIT	∼5 / 16	SCC	0.51 / 0.60
[7]	2017	AVD	S-BM	GMR	CreativeIT	∼5 / 16	PCC	0.62 / 0.37 / 0.35
[27]	2017	AV	S	Dilated CNN	RECOLA	∼3.5 / 27	CCC	0.68 / 0.49
[27]	2017	AV	S	Up-down Sampling	RECOLA	∼3.5 / 27	CCC	0.68 / 0.50
[28]	2017	AV	SV	CNN-LSTM ResNet-LSTM	RECOLA	∼9.5 / 46	CCC	0.71 / 0.61
[29]	2018	AVD	S	Dense vs. ResNet	MSP-Podcast	∼35 / 256	CCC	0.7 / 0.17 / 0.62
[30]	2018	AV	SV	CNN-LSTM	RECOLA	∼9.5 / 46	CCC	0.79 / 0.44
[31]	2018	AVD	S	Ladder Net	MSP-Podcast	∼39 / 256	CCC	0.77 / 0.29 / 0.69
[32]	2018	AV	SV	3D CNN-ConvLSTM	SEWA	∼3 / 64	CCC	0.583 / 0.654
[33]	2018	V	SV	3D CNN-ConvLSTM	RECOLA, SEWA	2.3 / 27, 3 / 64	CCC	0.546 / 0.612

2.1 CER from Speech

For the CER problem, various feature representations from a range of modalities are exploited to predict the evolution of continuous affect dimensions. These emotional states are expressed and communicated through various modalities such as speech, video, and body motion [34]. Among them, speech modality is widely used for emotion recognition [26, 7, 27, 28, 29, 35, 31]. Recently, a comprehensive summary of CER benchmark results is presented by Schuller et al. [36]. The top ranked performances over a series of challenges endorse that speech outperforms other modalities for the task of emotion recognition. Speech based end-to-end systems that are learning features from raw data demonstrate tremendous potential [35, 28, 36].

In [37], authors formulate the continuous regression of AV attributes as a linear combination of past AV attributes as well as current and past input speech features. Also as the primary modality, speech holds immense potential when augmented with other modalities such as body motion. Body expressions are also a tool of affective communication channel and a detailed review of literature on role of the body expressions in affect recognition is presented in [38].

2.2 Other Modalities

Apart from speech, modalities such as body motion data, have been extensively exploited [38, 11, 15]. Head and body gestures are correlated with prosody in speech [39] and participants are also able to recognize emotional intent with high accuracy from only head movements of the speaker [40, 41, 42], reflecting the significant role of head motion in communicating affect between dyadic participants. Non-verbal vocalizations (NVV) are also an integral component of spontaneous conversations and carry important emotional content [43]. Studies comparing interpretation accuracy between verbal and full-channel (verbal-plus-nonverbal) cues reveal that nonverbal cues can also be crucial to social interpretation [44, 45].

2.3 Multimodal CER

Multimodal analyses such as involving both speech and motion, has also been extensively used for CER [1, 13, 14, 35]. For the multimodal CER systems, fusion of modality representations plays a crucial role. While Sargin et al. in [46] propose methods for multimodal synchronization and fusion of speech and lip modalities in the context of open-set speaker identification, their proposed canonical correlation analysis (CCA) based fusion strategy suggests ways of fusing correlated and uncorrelated representations. Ngiam et al. in [47] explore multimodal deep learning wherein they apply deep networks to learn shared representations between multiple modalities. They highlight several multimodal learning settings for audio and video data covering classic deep learning, multimodal fusion, cross modality learning and shared representation learning. Their multimodal results reinforce that while audio information alone performs reasonably well for emotion recognition, fusing audio with another modality, for instance video, provides complementary information to the audio modality. They suggest that one approach to learning a shared representation is to find transformations for the modalities that maximize correlations. Some other challenges faced in multimodal machine learning such as representation, translation, alignment, fusion, and co-learning are reclassified and described in detail in [48].

2.4 Generative Models for CER

Earlier works in CER are based on a wide variety of generative techniques that include k-means clustering, support vector regression and Gaussian mixture model (GMM) regression [15, 16, 17, 19, 49]. These techniques have now been superseded by deep neural networks and since earliest works using deep learning for speech emotion recognition about a decade ago to present, a multitude of deep neural architectures have been proposed for continuous emotion recognition yielding significant performances [50, 51, 52], however, there are still some open challenges.

2.5 Neural Networks for CER

CNNs have outperformed conventional neural networks for many emotion recognition tasks [53, 54]. This is attributed to their translational invariance and weight sharing properties that ensures CNNs as excellent feature learners. With the advent of long-term dependencies where larger temporal windows are studied, the context-capturing ability of LSTMs looses significance as CNNs in essence achieve temporal modeling over larger windows. In fact, in a recent study, it is demonstrated that fully convolutional network achieves superior performances as compared to LSTM architecture for continuous emotion recognition task [55]. CNN based architectures have advantage of less trainable parameters and, therefore, a faster training process. Their results suggest that delay compensation does not have a considerable influence on the results for the CNN model, suggesting CNNs as suitable to model any long-term context patterns. Similarly in [27], the authors demonstrate that contrary to LSTMs, CNNs are also capable of capturing long-term temporal dependencies in audio data that can facilitate CER. Their results reveal that dilated convolutions can provide means of incorporating long-term temporal information by using filters with varying dilation factors. Another means of incorporating long-term dependencies with CNN architectures is with the use of downsampling/upsampling networks wherein a series of convolutions and max-poolings get a global view of the features at reduced resolution thus also facilitating computational resources. Performance scores from this study provide us reference AV baselines however over a different dataset. Furthermore, CNNs have also been employed for automatic feature learning in speech recognition tasks [56].

LSTM based frameworks are also well known to benefit CER owing to their ability to capture long-term contextual dependencies. Recently, attention based LSTM models have been proposed where attention mechanism maximizes the contribution of the relevant encoding context vectors and minimizes the influence of the irrelevant ones for the construction of the decoding context. In this direction, authors in [57] address the problem of emotional relevant feature extraction from speech by lever-

aging Attention-based Bidirectional Long Short-Term Memory Recurrent Neural Networks (BLSTM) with fully convolutional networks (FCN) in order to automatically learn the best spatio-temporal representations of speech signals. Some other recent works on LSTM based prediction models include cross-corpus and cross-task acoustic emotion recognition presented in [58] and multimodal interaction strategies to imitate the real interaction patterns in [59]. In dyadic interaction scenarios, LTMs have also been used in conjugation with several multimodal interaction feature strategies to capture interlocutor's multimodal information[60].

LSTM networks are primarily exploited to model memory and context in a large range of applications. LSTMs are units of recurrent neural networks (RNNs) that are often used to relate signals over arbitrary time lags. Although LSTMs can capture temporal relationships, they cannot capture spatial relationships in data. CNNs, on the other hand, owing to their convolutional structures can capture patterns in the data better. In order for LSTMs to consider convolution structures, a multi-dimensional architecture is required that forms the basis of ConvLSTM architecture.

CNNs and LSTMs, when implemented together, provide strong complementarity. Karita et al. apply Convolutional LSTMs for acoustic modeling, as well as ConvLSTM based extended architectures using forward-backward contexts [61]. Some recent works have applied ConvLSTM specifically to CER problem such as in [32], where the authors present an end-to-end continuous emotion recognition framework which exploits 3D convolutional networks with ConvLSTM and applies it to AVEC 2017 database (subset of SEWA database) using the video modality in a non-dyadic fashion. A similar approach is presented in [33] where ConvLSTM modules are used to learn spatio-temporal attention to selectively focus on emotional salient parts within facial videos of RECOLA and AVEC 2017 databases for valence prediction. Using ConvLSTM and a new loss function, the authors of [62] present an emotion recognition problem as a spectral-temporal sequence classification problem of bipolar EEG signals.

2.6 Dyadic CER Studies

When people communicate, they often adapt their interaction styles [63] and this synchrony in behavior can be exploited to further improve CER. The dependence between discrete emotional states of the dialog partners is investigated in [64]. In this study, authors formulate four aspects in dyadic interactions namely (a) dependence between the emotional states of the dialog partners, (b) dependence between the emotion of one subject and the expressive behaviors of the other, (c) dependence between heterogeneous behaviors from the dialog partners, and (d) effect of cross-subject multimodal information for emotion discrimination. Based on their co-occurrence analysis between the discrete emotional states displayed by speakers and listeners in the turns during spontaneous dialogs, the authors reveal that 72 percent of the time both conversation partners share the same emotion. This result supports the emotional adaptation hypothesis.

Further, in [65], the relationship of activation and valence with gesture and speech based synchrony is explored. These studies suggest that there exits a relationship between the emotional attributes of the participants in a dyadic interaction. Similarly, we hypothesized that while the three dimensions are independent, there exists a synchrony between emotions at a cross-subject level. This dependency should reflect itself in the prediction of emotion in a dyadic setting. Tzirakis et al. in [28] utilize audio and video features of the RECOLA dataset for CER. As feature extractors, CNNs and ResNets are employed. Their work provides us corresponding baseline benchmarks, however across a different dataset. They provide a comprehensive comparison among different speech and visual features and highlight highest unimodal and multimodal concordance correlation scores for the AV attributes. There are also other lines of work that have studied CER during dyadic interactions on the course of either addressing same problem using other approaches or an entirely different emotional database. For dyadic CER studies, the RECOLA database has been extensively used [66], and while in the database the continuous levels of arousal and valence are annotated by 7 raters at frame-level, emotion labeling for both speakers

is not available simultaneously. Same is the case with SEWA database that has earlier been used in affect sub-challenge tasks [67].

In [26], Chang et al. use audio and video features of the CreativeIT dataset for cross-corpus emotion recognition. They integrate the joint emotionally-relevant information across both databases by deriving a weighted fused gram matrix for SVR. This work validates that emotionally relevant information improves emotion recognition.

Recently, dense neural network architectures have been applied to address the CER problem. In [32], ConvLSTMs are applied to the video modality for end-to-end CER. In [29], authors use audio features from the MSP-PODCAST corpus to predict AVD attributes using dense network architectures. This study also provides us some score trends for emotion attributes across a different emotional speech database. They explore performance of regression models for AVD as a function of several network parameters. They suggest that the most effective approach to improve performance is to increase the size of the training set.

On the DER end, a recent study on a Gated Recurrent Unit (GRU) based conversational memory network is proposed to model emotional context and identify utterance-level emotions in dyadic conversational videos [68]. The framework takes a multimodal approach comprising audio, visual and textual features to model past utterances of each speaker into memories. These memories are then merged using an attention mechanism to form a merged utterance representation that captures inter-speaker dependencies and this new representation is used for learning discrete emotion classes of the IEMOCAP database [69]. Similarly, another recent study models speaker and interlocutor's context using self-attention mechanism and fuses the adaptive context with the present behavior to predict the current emotional states [70]. Context attention layers are used to model interaction contexts in the framework. Similarly, a comparison of the attentive convolutional neural network (ACNN) with CNNs in predicting the discrete emotional classes has been investigated over the IEMOCAP database using various feature types [71]. These recent studies indicate that DER in dyadic interactions has been explored using a multitude

of deep neural architectures ranging from CNNs, LSTMs, 3D convolutional recurrent neural networks (3D-CRNN), attentive CNNs and residual CNNs. However, specific to models where interlocutor's interdependencies are modeled as context within deep neural networks, works in fields of DER and CER are still few. Thus, context modeling within neural network architectures may significantly improve the CER performance.

In conclusion, the prevalent challenges in CER necessitate new approaches by combining wider range of cues to attain improved and generalized solutions. However this is a challenging task since the estimated performance is context-dependent and varies across databases. This shortcoming may be mitigated if the overall relationship between the participants of a conversation is taken in account during the continuous emotion recognition.

Chapter 3

CER RESOURCES

Having appropriate datasets that can truly explore a classification or regression task is just as important as the methodology one proposes to solve these problems. Since the problems explored in the current thesis were multifaceted, we used a number of datasets to explore the different problems. We discuss these in the rest of this section.

3.1 USC CreativeIT Database

Some of our works use the USC CreativeIT database, which is a multimodal database of theatrical improvisations [72, 73]. The database is collected using cameras, microphones, motion capture and contains detailed audiovisual information of the actors' body language, speech cues as well as emotional annotations. The database serves two purposes. Firstly, it provides insights into the creative and cognitive processes of actors during theatrical improvisation. Second, the database offers a well-designed and well-controlled opportunity to study expressive behaviors and natural human interactions.

The database design is based on well-established theatrical improvisation technique of Active Analysis and results from a close collaboration of theater experts, actors and engineers. Motion Capture technology is deployed to obtain detailed body language information of the actors, in addition to microphones, video and carefully designed post-performance interviews of the participants. The database is annotated for continuous emotional descriptors as well as discrete emotional labels. The frame-level continuous emotional annotations of both subjects in a dyadic improvisation facilitate many research directions including dyadic affect and behavior inter-dependencies, continuous emotion tracking and multimodal modeling. Active

analysis is the primary acting technique utilized in this database. A key feature of active analysis is the use of verbs in driving the the actors' responses. This technique facilitates the interaction and behavior of the actors to be more expressive and closer to natural, which is crucial in the context of emotion recognition.

The database comprises of two different theatrical techniques including the two-sentence exercise and the paraphrase, both of which originate from the Active Analysis methodology. The database contains the recording of nine full sessions, each of which contains approximately one hour of audiovisual data. In total there are 40 two-sentence exercises and 19 paraphrases with 19 actors. Each interaction on average has a length of 3.5 minutes and contains two recordings, one for each actor in a pair.

On the equipment and technical end, Vicon motion capture system is used with 12 motion capture cameras to record 45 marker's (x, y, z) position for each actor. Two Full HD cameras are placed at corners of the recording room to capture the performance of the actors. Also, each actor has a close-up microphone to record actors' speech at 48KHz with 24 bits. Some of the post-processing steps include mapping of each of the markers captured into a subjects defined body model and emotional annotations.

During improvised dyadic sessions there is a continuous flow of body language and dialog and a diverse expression of emotions and intentions. In order to preserve this flow, the sessions are annotated using continuous labels instead of truncated sentences or other arbitrary chunks. Furthermore, since commonly used categorical emotional attributes (angry, happy etc) may not be applicable or sufficient for our data, a more comprehensive set of attributes are chosen. These contain dimensional emotional descriptors (valence, activation, dominance) as well as theater performance ratings (interest, naturalness), which may facilitate future theatrical performance analysis.

All continuous annotations are performed by watching the session videos and using the Feeltrace software. Feeltrace is a publicly available emotional annotation tool, described in [74], that enables the user to continuously move the mouse along

the computer screen so as to indicate the attribute value, ranging from -1 to 1. For the discrete annotations, annotators are asked to provide a label ranging from 1 to 5. The annotated attributes are, to a large extent, subjective.

Statistical analysis is performed to test the annotation agreement between multiple annotators. In order to examine linear relationships between the annotations, the Pearson correlation coefficients are computed between all pairs of annotations, which are all found significant at the 0.01 level (2-tailed). Furthermore, the Pearson correlation coefficients between all pairs of annotations for all continuous emotional descriptors are also computed. They were all found statistically significant at the 0.01 level (2-tailed) and positive, which indicates consistency between different annotators despite their possibly different styles.

In this study, the working database comprises of five sessions after aligning the three emotion attributes over a common time range with the audio and motion data, and removing noisy recording and bad annotations. Since our work is formulated for a cross-subject analysis, we require emotional attributes from both participants to evaluate the proposed system. Thus, after eliminating some recordings, which do not have emotional attributes for both participants, our final experimental dataset includes 65 time-synchronized and AVD aligned recordings. We use the mean inter-annotator ratings of three human annotators as the ground truth attributes. Also note that each session is performed by a unique pair of actors and therefore, session independence implies subject independence. Table 3.1 presents the session-wise distribution of recordings within the working dataset.

Table 3.1: Session-wise distribution of USC CreativeIT recordings within the working dataset.

Sessions	Apr23	Apr30	Mar5	May12	May13	Total
Recordings	4x2	4x2	6x2	2x2	5x2	42

3.2 IEMOCAP Database

Some of our work is based on Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, which is extensively used to study human emotional behaviors [69]. We used this database in our study on natural dialog identification as described in Chapter 4. The database contains five sessions where ten actors (two per session) were trained to perform two types of sessions namely scripted dialogs and improvised dyadic interactions that evoke various emotions.

The most important consideration in the design of this database was to have a large emotional corpus with many subjects, who were able to express genuine emotions. In order to achieve this goal, the material and actor selection was carefully curated. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration and neutral state). The average duration of the dialogs is approximately 5 minutes. The actors wear markers on the face, head, and hands, which provide detailed information about their facial expressions and hand movements during scripted and spontaneous spoken communication scenarios. The corpus contains approximately twelve hours of data, which was manually segmented and transcribed at the turn level. Note that each session is performed by a unique pair of actors and therefore, session independence implies subject independence.

Each of the sessions was recorded with 53 markers (diameter 4 mm) attached to the face of one of the subjects in the dyad to capture detailed facial expression information, while keeping the markers far from each other to increase the accuracy in the trajectory reconstruction step. Once the scenarios were recorded for one actor, the markers were placed on the other subject. This approach was adopted to avoid interference between two separate setups. The audio was simultaneously recorded using two high quality shotgun microphones (Schoeps CMIT 5U) directed at each of the participants. The sample rate was set to 48KHz. In addition, two high-resolution digital cameras (Sony DCRTRV340) were used to record a semi frontal view of the participants. The recordings were synchronized by using a clapboard

with reflective markers attached to its ends.

As part of post-processing, the dialogs were manually segmented at the dialog turn level (speaker turn), defined as continuous segments in which one of the actors was actively speaking. For this database, two of the most popular assessment schemes were used: discrete categorical based annotations and continuous attribute based annotations. On the categorical emotional descriptors end, six human evaluators were asked to assess the emotional content of the database in terms of emotional categories. The final emotional categories selected for annotation were anger, sadness, happiness, disgust, fear and surprise as the basic emotions plus frustration, excited and neutral states. Statistical analysis was done to analyze the inter-evaluator agreement. On the continuous emotional descriptors end, the emotional content of an utterance is described by the use of primitive attributes such as valence, activation (or arousal), and dominance.

In general, the detailed motion capture information, the interactive setting to elicit authentic emotions, and the size of the database make the IEMOCAP database a valuable choice for our natural dialog study.

Chapter 4

ACOUSTIC FEATURE SELECTION FOR NATURAL DIALOG IDENTIFICATION

4.1 Introduction

In behavioral signal processing, extensive work is being carried out on improving voice quality and naturalness of text-to-speech synthesis (TTS) systems [75, 76, 77] and the synthesis of human-like communication in man-machine interfaces [75, 78, 79]. Often naturalness and intelligibility are evaluation criteria. Judgment of naturalness of synthesized and vocoded speech so far has mainly been based on perceptual evaluations such as MOS tests performed by listeners [80]. However, identification of naturalness capturing parametric acoustic cues and their incorporation into TTS systems still has plenty of room for improvement. Also, contrary to its frequent use, naturalness remains a less defined concept in the field of behavior signal processing. While extensive psychology literature details natural elements of speech [81], their parametric representations for recognition are often not well-defined or require optimization [82, 83]. New work is needed to explore the manipulation of prosodic phenomena (pitch, loudness, voice quality, inter-lexical pause duration) in conjunction with other elements such as fillers and emotions with the goal to further develop synthetic voices that convey appropriate naturalness and personality traits. While prosodic feature sets that convey personality traits have been explored in literature recently [84, 85], feature sets that are rich in temporal and spectral characteristics of natural speech are not yet explored.

Apart from applications in TTS systems, identification of prosodic and spectral 'naturalness' in features under natural and artificial dialog classification setting can also have important contributions to interesting applications such as in forensic audio

analysis to differentiate between natural and doctored speech samples. Therefore, we investigate retention of several prosodic and spectral features based on similarity metrics under a novel two-class natural and artificial interaction classification approach to identify vocal features that help to distinguish characteristics of natural dyadic speech turns from artificial turns.

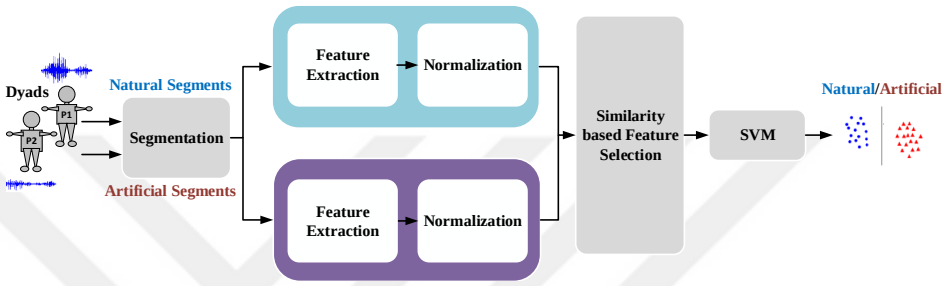


Figure 4.1: Identification of natural dialogs based on acoustic feature selection.

4.2 Methodology

Figure 4.1 summarizes the procedure for identification of natural dialogs based on acoustic cues. We conduct feature-level analysis to study the *naturalness* of speech. We adopt a 'natural' and 'artificial' dyadic segment classification approach where high classification rate indicates confidence towards features that capture most naturalness. We validate the predictive ability of our reduced feature sets by training support vector machine (SVM) based classifier model on the data and testing it to distinguish between tuples of contiguous natural utterances in dyadic interactions and similar tuples where randomization is introduced in feature domain to create artificial speech. For performance comparison, we conduct our own baseline experiment with an exhaustive set of acoustic features. We perform data processing on two new scopes of data that we refer to as 'global' and 'session-level' to improve classification accuracy.

Please note that the results in this chapter are based on Interactive Emotional Dyadic Motion Capture (IEMOCAP) database, which is extensively used to study

human emotional behaviors [69].

4.2.1 Segment Construction

Due to their adaptive and interactive progression, dyadic interactions attract more interest from researchers [64, 69, 86]. As a first step, dialogs are segmented into appropriate length turns to ensure availability of reasonable amount of prosodic information per instance in the dataset. We refer to four appended turns as a 'segment. Turn length being too short or too long is informatively insufficient or computationally exhaustive respectively. Segments are constructed in two settings; (a) natural conversation order and (b) artificial/arbitrary order, with their respective labels as *natural* and *artificial*.

In dyadic interactions, two participants 1 and 2 interact for which turn-level speech recordings are available. We denote the k-th natural segment as

$$D_k^{nat} = [T_{2k-1}^{S1}, T_{2k-1}^{S2}, T_{2k}^{S1}, T_{2k}^{S2}],$$

where T_k^{Sn} is the k-th turn of speaker Sn. We denote the k-th artificial segment as

$$D_k^{art} = [T_{n1}^{S1}, T_{n2}^{S2}, T_{n3}^{S1}, T_{n4}^{S2}],$$

where $S1$ and $S2$ are two speakers who are paired in the dataset sessions, and the turn indexes $n1$, $n2$, $n3$ and $n4$ are randomly selected such that consecutive turns do not follow the natural order of speech.

Figure 4.2 illustrates the construction of a natural segment from four *speaking* turns in the natural order of speaking thus preserving discourse coherence through the conversation. Note that due to the dataset setup, the speakers $S1$ and $S2$ are male and female speakers in random order. Artificial speech segments are appended features extracted over the duration of each turn constituting a segment independently and randomized within the scope of a session. This feature space randomization generates artificial feature contours demonstrating pseudo patterns that are unlikely to occur in real-life natural conversations. Prior experimentation showed that time-domain random concatenation of speaking turns leads to false inclusion of jitter as top-ranked discriminating feature. To avoid this, we randomly

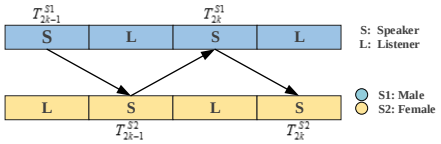


Figure 4.2: Natural segment construction.

concatenate feature rather than temporal representations. The artificial segment construction procedure is designed to ensure the following key points:

1. Appended turns alternate over both dyads for gender progression consistency ensuring that the only variability in natural and artificial segmentation is the randomness in turn taking of dyads.

2. Randomizations are done only within each session since different sessions have different compositions.

3. Conversation content variability is guaranteed by ensuring a turn difference of at least ten between consecutive male/male and female/female turns in each artificial segment.

4.2.2 Feature Extraction and Processing

In existing literature, acoustic features extracted for synthesis and recognition tasks vary significantly in dimensionality ranging from 10 basic features in [87] to 276 and 1134 features [88, 89]. The wide variation in feature set size is due to lack of an optimal feature set, dependency of data on end classification tasks [11]. Pitch, Loudness, voice quality, speaking rate and the effect of pausing are all factors that have been shown to influence perception of synthesized speech [90].

We extract frame-level feature descriptors over the duration of each natural and artificial segment independently and normalize feature data to reduce variabilities. We denote each feature vector as $f_k(i)$ for i-th frame of segment k. Each feature vector is 18 dimensional including fundamental frequency and probability of voicing

as prosodic, mel frequency cepstral coefficients (MFCCS $0-14$) as spectral and loudness as energy related low-level descriptors (LLDs). Pitch and energy related features are computed over 60 ms windows while MFCCs are extracted over 25 ms windows, both with 10 ms frame shifts. Note that speaking rate is not included in the feature sets because it is known that slower speech rates lead to slower reaction times in synthetic *and* in natural speech thus demonstrating slow perception patterns in both natural and synthetic speech [91].

Statistical functionals are computed over the temporal duration of each segment across each of the four LLDs and comprise 13 statistical representations including arithmetic mean and standard deviation of contour values (1, 2), skewness (3), kurtosis (4), 25^{th}, 50^{th} and 75^{th} percentiles with 1% percentile (5, 6, 7), 99% percentile, and percentile range $1\% - 99\%$ for outliers robust max, min and signal range (8, 9, 10, 11, 12, 13) respectively. With four turns constituting a single dialog segment, each natural and artificial segment is thus summarized in $234 \times 4 = 936$ functionals. To make data comparable, we exploit z-normalization and min-max normalization to minimize inter-speaker variability. Our z-normalized (Z) on a feature vector is defined as

$$f_n^z = \frac{f_n - \mu_n}{\sigma_n}, \tag{4.1}$$

where f_n is the n-th feature with mean μ_n and standard deviation σ_n. Min-max (MM) normalization is defined as

$$f_n^{mm} = \frac{f_n - f_n^{max}}{f_n^{max} - f_n^{min}}, \tag{4.2}$$

where f_n^{max} and f_n^{min} are the maximum and minimum feature value of the feature f_n.

We define two scopes of normalizations namely *global* and *session*. In global scope, normalization across all data irrespective of session breakdown is performed whereas in session, data is normalized within a session. In the latter case, since we constrain μ and σ within a session, we expect better performance because the effects of speaker vocal characteristics and inter-speaker differences are significantly reduced.

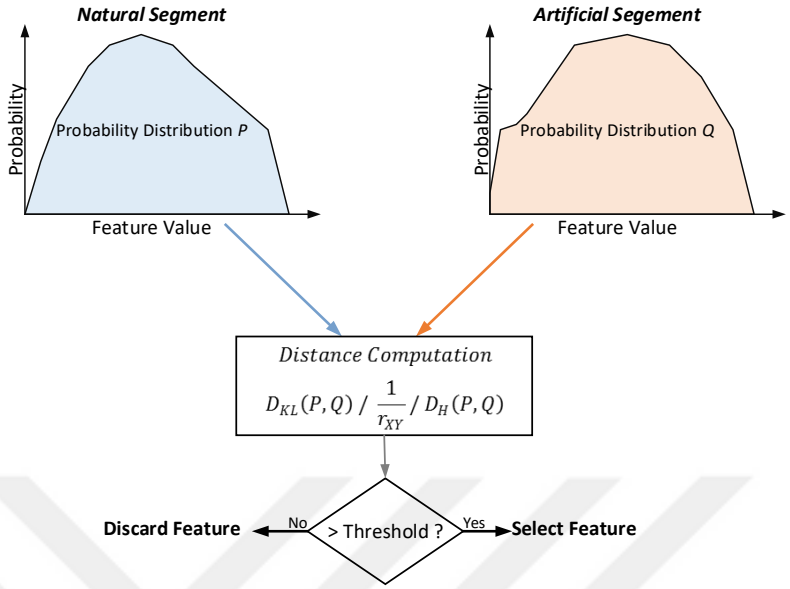


Figure 4.3: Feature set selection based on similarity and distance measures

4.2.3 Similarity based Feature Selection

Since the effectiveness of our setup depends on the discrimination capacity between natural and artificial classes, features with class dependent statistical characterizations can be more discriminating. To see this experimentally, we deploy feature set reduction using similarity measures such as KLD distance, Pearson's correlation coefficient and Hellinger distance. We reduce the feature set by computing these metrics for each pair of corresponding feature between the natural and artificial class. In the case KLD and Hellinger distance based selection, we retain features for which the dissimilarity is above a threshold whereas for Pearson correlation, we retain features for which correlation is below a threshold. In effect, we remove informatively redundant features at the cost of some controlled loss of data.

Figure 4.3 illustrates our feature selection approach. Considering the natural subset of a feature as $X_N = \{x_1, x_2, ..., x_n\}$ and $Y_A = \{y_1, y_2, ..., y_n\}$ as the artificial subset with probability distributions $P = \{p_1, p_2, ..., p_k\}$ and $Q = \{q_1, q_2, ..., q_k\}$,

respectively, the KLD metric is given as

$$D_{KL}(P,Q) = \sum_{i=1}^{k} p(i) \log \frac{p(i)}{q(i)}. \tag{4.3}$$

High KLD distance suggests better feature-wise statistical separation for the classes, therefore, we select features above a threshold to reduce features. The Pearson correlation coefficient between the two samples is given as

$$r_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{4.4}$$

where $\overline{x}$ and $\overline{y}$ denote the sample means of X_N and Y_A. r_{XY} ranges between -1 and $+1$, where values around $+1$, -1, 0 indicate high positive, high negative and no correlation respectively. We reduce the feature set by thresholding close to 0 since features in this region show the least correlation between the classes and thus are most discriminating. Finally, the Hellinger distance is given as

$$D_H(P,Q) = \frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}. \tag{4.5}$$

Hellinger distance is bounded over the range of $[0, 1]$ where 0 indicates complete similarity and 1 indicates complete dissimilarity. Thresholding similar to KLD case is applied here.

4.3 Experimental Evaluations

There is evidence that speech produced spontaneously in a conversation is perceived more natural in comparison to read prompts [92]. For this reason, we construct speech segments from improvisation data of IEMOCAP corpus. The working dataset comprises of improvisations in sessions 2, 3, 4 and 5 of IEMOCAP corpus. To ensure speaker independent testing, session 2 is fixed for testing while the training set is built on sessions 3, 4 and 5. A quantitative breakdown of the constructed segments is tabulated in Table 4.1. A two-class SVM model is trained to differentiate between the natural and artificial segments with radial basis kernel [93]. Classification rates are used as a confidence of discrimination between natural and artificial

Table 4.1: Session level distribution of natural and artificial segments within the IEMOCAP database.

	Sessions			
	2	**3**	**4**	**5**
Natural	364	346	611	738
Artificial	364	346	611	738
Testing	728			
Training		3390		

Table 4.2: Average accuracy rates (%) on global and session-level scopes of z-normalized and MM-normalized features without feature selection (baseline exhaustive set).

Normalization	**Global**	**Session**
Z	85.71	87.81
MM	68.27	68.96
Chance	50	

conversations. We conduct a leave-one-session- out cross-validation on testing data. Table 4.2 presents average classification accuracies on global and session-level scope of normalization using Z and MM approaches. Z outperforms MM with a relative 18.85% improvement at session-level, and yields an absolute 37.81% improvement over our chance baseline (without normalization). On global level, the absolute improvement over baseline is 35.71%. The improved performance of Z over MM can be attributed to its ability to spread outliers over the entire distribution.

Table 4.3: Average percentage classification accuracy with KLD, Pearson correlation and Hellinger distance based feature selection schemes as a function of percentage feature retention (FR) of session-level z-normalized features

FR (%)	$D_{KL}(P,Q)$	$1/r_{XY}$	$D_H(P,Q)$
90	**90.8**	86.54	84.89
80	**88.46**	85.03	84.07
70	**86.26**	84.34	83.65
60	83.38	**83.93**	82.97
50	82.69	**82.83**	81.04

4.3.1 Feature Selection Evaluations

We first evaluate the performance of our selected feature reduction techniques. Table 4.3 illustrates the obtained percentage accuracy with the three deployed feature set reduction criteria. For a feature retention between 70% to 90% corresponding to a $D_{KL}(P,Q)$ threshold between 0.134 and 0.0166 respectively, $D_{KL}(P,Q)$ based feature selection yields highest accuracies peaking at 90.80%. Note that thresholds under each of the three selections are a function of percentage feature retention thus vary differently under each of the three techniques. At lower retention rates(last two rows), r_{XY} based feature selection outperforms $D_{KL}(P,Q)$ and $D_H(P,Q)$ based selections. As observed, with reduction in number of retained features, there is an expected fall in performance due to loss of information, however,even when only 50% of features are retained (last row), we still achieve a remarkable accuracy of over 81% for all cases.

In [81], the features of *naturalness* in a conversation are discussed from a psychology point of view. Naturalness defined as the discourse coherence, and is examined under four interdependent sub-categories: alignment, feedback, reflectiveness and intonation. For a conversation to be coherent and progress well such that the par-

ticipants pursue their goals co-operatively rather than individually, alignment and feedback are essential requirements for discourse coherence. Furthermore, the discourse intonation system consists of a set of choices available to speakers namely prominence, tone, key and termination. These choices are not formulated with reference to grammar and do not have fixed attitudinal meanings. A natural conversation is reflective of these contributing factors in the ongoing negotiation of meaning between discourse participants. While co-operativeness, feedback and language ability are different to capture prosodically, these findings particularly suggest that the rise and fall of the speaking voice of engaged partners as is captured in *pitch* is a governing quality of natural dialogs.

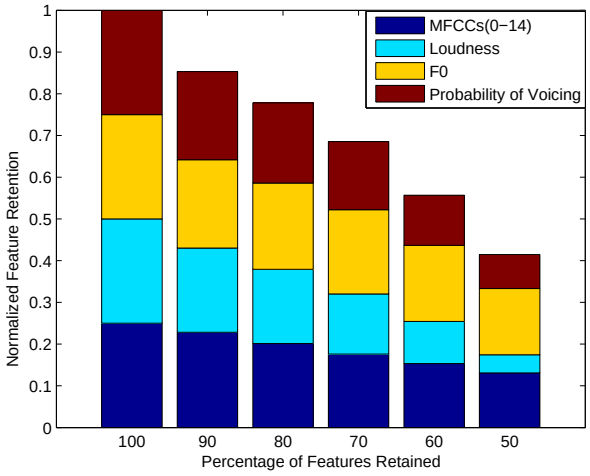


Figure 4.4: Normalized feature-wise percentage retention breakdown of z-normalized features at session-level as a function of overall percentage of features retained under four categories: MFCCs [0-14], Loudness, Fundamental frequency, Probability of voicing.

We validate this psychological evidence by observing percentage retention patterns to identify feature descriptors and statistical functionals that hold the most information using only $D_{KL}(P, Q)$ as the selection criteria. We observe in Figure 4.4 that as the number of overall retained features reduces, fundamental frequency (in legend: yellow) related features are the top-ranked retained features suggesting their

highest discriminatory power from amongst all feature descriptors. These findings reinforce the crucial role of fundamental frequency in describing the intonation and speaking melody in speech synthesis models.

Additionally, the functional analysis shown in Figure 4.5 highlights 50^{th} percentile and skewness as the top most retained functionals across features indicating that measures of variability, distribution asymmetries and location parameters are crucial for natural dialog identification.

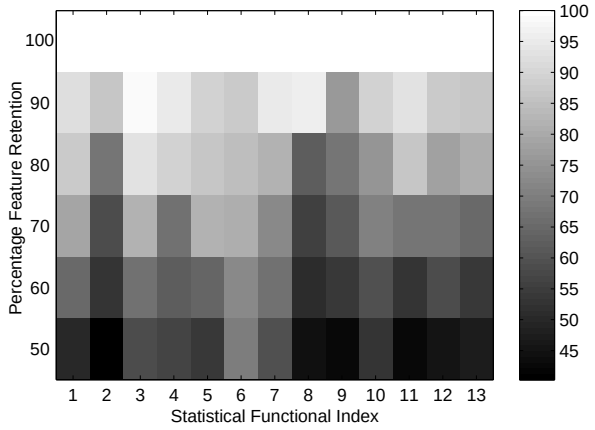


Figure 4.5: Retention percentage of each z-normalized statistical functional at session-level as a function of the overall percentage of features retained. Note: functional indexes (x-axis) are in the order of their description in Section 4.2.2.

Chapter 5

DYADIC CROSS-SUBJECT CER

5.1 Introduction

In underlying hypothesis for this study is that there exists a synchrony between dimensions of emotion at a cross-subject level and this dependency should reflect itself in the prediction of emotion in a dyadic setting. To achieve this, we propose a novel two-stage framework for emotion estimation utilizing the cross-subject dependence at the emotion-level. In the first stage, we perform conventional CER for participants in a dyadic interaction using their speech and body motion data to estimate activation, valence and dominance (AVD) attributes independently. We use Gaussian Mixture Model (GMM) regression for attribute estimation similar to [15, 16, 49]. In the second stage, we improve the first stage estimates by performing CER of the selected speaker using her/his speech and body motion modalities as well as using the estimated affective attribute(s) of the other speaker as side emotional information (SEI). Our experimental evaluations indicate that the two stage, cross-subject continuous emotion recognition (CSCER), provides complementary information to recognize the affective state, and delivers promising improvements for the continuous emotion estimation problem.

5.2 Methodology

The block diagram of the proposed two stage continuous emotion recognition framework is given in Figure 5.1. In dyadic interaction setup, two participants, 1 and 2, interact, and we have speech and body motion recordings from these two participants from the USC CreativeIT database. We extract frame level speech and body motion features and represent them as $f_i^S(k)$ and $f_i^M(k)$ at frame k for i-th par-

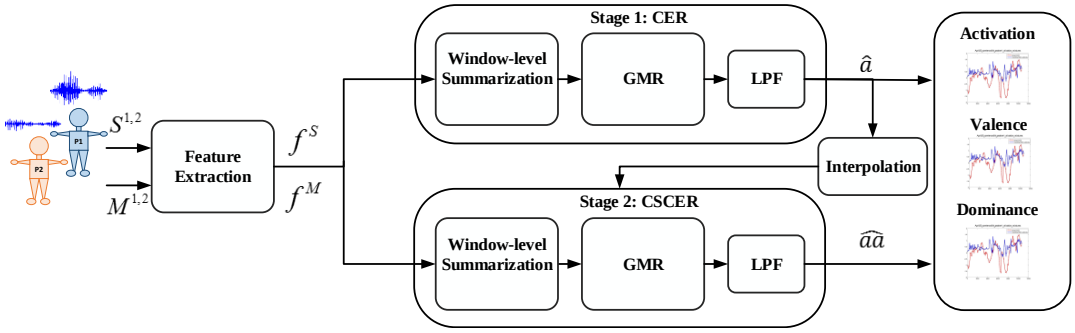


Figure 5.1: Cross-subject continuous emotion recognition (CSCER) framework.

ticipant. Correspondingly, $a_i^A(k)$, $a_i^V(k)$ and $a_i^D(k)$ are defined as the underlying activation, valence and dominance attributes for the i-th participant, respectively.

At the first stage, we obtain a window-level summarization of speech, motion and multimodal features for both participants independently to estimate their corresponding emotion attributes in a 3-class recognition setup. Subsequently, we exploit Gaussian Mixture Regression (GMR) to construct a statistical mapping between the underlying observed summarized speech, motion and multimodal features and the hidden window-level emotional attributes, a. After GMR, the estimated attribute, $\hat{a}$, is low-pass filtered for smoothing as in [16, 19].

In the second stage, we aim to improve the first stage estimates by performing CER of the i-th participant using her/his speech and body motion modalities as well as using, $\hat{a}_j$, the estimated affective attribute(s) of the j-th participant as a feature. This is expected to improve CER since cross-subject emotions are related to each other. GMR is used at the second stage again for the regression analysis. In essence, we validate the predictive ability of the first stage cross-speaker AVD estimates by re-iterating them for CER to generate our second stage updated estimate, $\widehat{aa}$.

5.2.1 Feature extraction and window-level summarization

Frame level acoustic features are extracted as 39 dimensional feature vectors including energy, the first 12 MFCCs together with the first and second time derivatives. Frame level motion feature vector is 24 dimensional and includes the Euler rotation

angles in directions (x,y,z) of the arm and forearm joints together with their first derivatives over each frame. Acoustic features are computed every 16.67 ms with 8.33 ms overlap to match the motion capture rate.

In our window-level framework, we perform CER over windows of size 3 sec with an overlap of 1 sec between adjacent windows to generate the summarized speech, motion and multimodal feature vectors. We extract frame level feature vectors over the temporal duration of the window and construct matrices of features as $F_n^S = [f_1^S, ..., f_K^S]$ for the n-th window with dimensions $39 \times K$. Similarly, our motion feature matrix is constructed as $F_n^M = [f_1^M, ..., f_K^M]$ with dimensions $24 \times K$. Statistical functionals are computed over feature frames within the temporal duration of each window to provide a down-sampled statistical representation as in [16]. The resulting matrices of speech, motion and multimodal features are reduced by applying Principal Component Analysis (PCA) retaining 50 most informative principal components that corresponds to more than 90% of the total variance on average. We incorporate dynamic information by augmenting the PCA summarized features with their first order temporal derivatives. Speech, motion and multimodal data along with the emotional attributes are z-normalized using the global means and standard deviations of the dataset. Finally, we define our summarized speech, motion and multimodal feature vectors as h^S, h^M and h^{SM}, respectively. Note that in the proposed CSCER system, the cross-speaker emotion descriptor is appended to the feature set, which increases the feature dimensionality by one.

Since emotional attributes change slowly in comparison to audiovisual data, the resulting window-level covariance matrices may become singular causing the GMR to fail. One possible way to ensure increased signal variability while preserving signal spectral characteristics is to selectively add a small amount of noise to the slow-varying signal. Therefore, we add additive white Gaussian noise (AWGN) noise with a signal to noise ratio (SNR) of 200dB in the SEI to increase the data variance.

We perform a pre-analysis to access the correlation between the emotional attributes of two participants engaged in a dyadic interaction. We analyze the correlation between mean inter-annotator agreements across sessions as ground truths

Table 5.1: Cross-subject inter-attribute Pearson's correlation based on ground truth activation(A), valence(V) and dominance(D) attributes.

		Participant 1		
		a_1^A	a_1^V	a_1^D
Participant 2	a_2^A	0.5783	-0.2887	0.1283
	a_2^V	-	0.2248	-0.1974
	a_2^D	-	-	-0.2617

for the three emotional attributes of USC CreativeIT database. As presented in Table 5.1, there is a significant correlation between the activation of the two speakers. The cross-attribute correlation generally seems insignificant for valence and dominance of both speakers since they are slightly correlated. Based on this analysis, we do not expect significant improvement in estimation of valence and dominance attributes under our CSCER framework. Nonetheless, the inter-dependencies may still render slight improvements. On the other hand, we expect activation prediction to reflect most significant advantage of utilizing cross-subject emotional dependencies as activation of both speakers seems promisingly correlated. In this analysis, it should be kept in mind that the annotations are context dependent and subjective to dataset size thus it is hard to generalize inter-attribute correlations across datasets.

5.2.2 Cross-subject continuous emotion recognition (CSCER)

Continuous emotion recognition, which performs estimation of the affective AVD attributes, is formulated using the GMR framework as

$$\hat{a}(h) = \arg\max_{a} P(a \mid h, \lambda(h, a)), \tag{5.1}$$

where $\lambda(h, a)$ is the joint Gaussian mixture density of the observation and affect variables, h and a.

In the first stage, we adopt the GMR framework to estimate the affective AVD attributes using speech ($\hat{a}_i(h^S)$), motion ($\hat{a}_i(h^M)$) and multimodal ($\hat{a}_i(h^{SM})$) observations for the i-th participant.

In our second stage CSCER, we estimate attribute of the i-th participant using her/his speech/motion/multimodal data as well as using the estimated cross-subject affective attribute of the j-th participant. Note that the observation space dimension increases by one, and the second stage GMR estimation becomes,

$$\widehat{aa}_i(h_i, \hat{a}_j(h_j)) = \arg\max_a P(a \mid h_i, \hat{a}_j(h_j), \lambda(h, a)). \tag{5.2}$$

In an independent experimental setup, we also deploy ground truth from human agreements on AVD attributes as the true side information to set an upper performance bound for the second stage estimates. As the true side information is not available in a realistic scenario, it sets theoretically an upper performance benchmark. We can define the second stage CER estimate with true side information as $\widehat{aa}_i(h_i, a_j)$.

We perform a comparative analysis of CSCER estimates with the first stage (baseline) results and our ideal true side information based upper bound reference. For each of the estimated attributes, we deploy individual i.e. (a) activation, (b) valence and (c) dominance descriptions. In general, our experimental setup tests the predictive ability of intensity, pleasure and control of each participant on the prediction of emotional experience of other with and without supplementary cross-subject emotional descriptions.

5.3 Experimental Evaluations

We devise three different experimental setups based on speech, motion and a multimodal data. The absolute relationship of features to emotional attributes is ensured thereby we report the absolute weighted mean correlation across sessions. It is empirically deduced that this approach is not sensitive to increased number of mixtures. We conduct five-fold-leave-one-session out cross-validation for testing.

Table 5.2 presents CSCER results with true side information. As compared to

the first stage, wherein AVD estimation is based on speech, motion and multimodal data only (bottom-most row), the true side information delivers considerable improvement for all the three estimated attributes. The advantage of cross-speaker emotional dependencies is best captured in activation as expected by our pre-analysis. While no significant dependency is observed between valence-dominance and dominance-activation, we observe that the slight correlation between activation-valence and vice versa shows advantage in activation and valence estimation over the first stage estimates. We obtain the highest improvements with like-attribute side AVD information for diagonals of Table 5.2, and this observation is consistent over all three modalities. We expected a slightly low performance in the multimodal case as reported due to overfitting because of low sample to feature ratio.

Table 5.2: Cross-subject continuous emotion recognition (CSCER) at the window-level of AVD with ground truths as true side emotional information based on speech, motion and multimodal cues. We present mean absolute correlation values between estimated emotional curve and the ground truth.

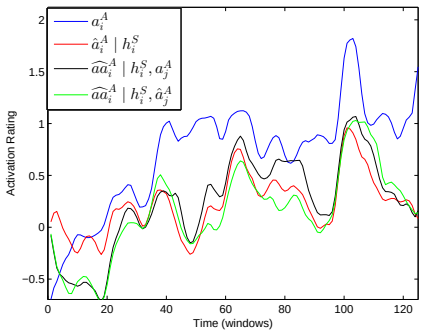
		Estimated Attribute								
		h^S			h^M			h^{SM}		
		$\widehat{aa}_i^A$	$\widehat{aa}_i^V$	$\widehat{aa}_i^D$	$\widehat{aa}_i^A$	$\widehat{aa}_i^V$	$\widehat{aa}_i^D$	$\widehat{aa}_i^A$	$\widehat{aa}_i^V$	$\widehat{aa}_i^D$
Cross-subject side information	a_j^A	**0.6321**	0.3290	0.2526	**0.5113**	0.3135	0.2441	**0.5017**	0.2963	0.2504
	a_j^V	0.5164	0.3220	0.3530	0.3932	**0.3559**	0.3267	0.4734	0.3104	0.2609
	a_j^D	0.5106	0.3173	**0.3675**	0.3366	0.3446	0.3466	0.4681	0.2976	0.2870
	Stage1	0.5265	0.2896	0.3088	0.3686	0.2891	0.2350	0.5275	0.2913	0.2615

Table 5.3: CSCER at the window-level of activation, valence, dominance with estimated ground truths (Stage1) as side emotional information based on speech and motion cues. We present mean absolute correlation values between estimated emotional curve and the ground truth.

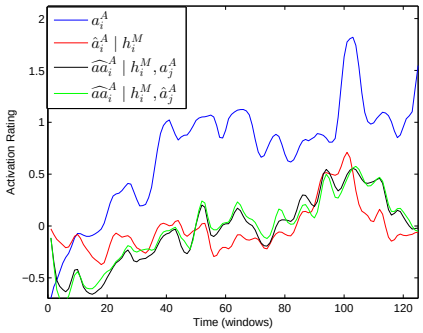
	Estimated Emotion		
	$\widehat{aa}_i^A \mid \hat{a}_j^A$	$\widehat{aa}_i^V \mid \hat{a}_j^V$	$\widehat{aa}_i^D \mid \hat{a}_j^D$
h_i^S	**0.6189**	0.3441	0.3330
h_i^M	0.5079	0.3645	0.3484
h_i^{SM}	0.5361	0.3374	0.3240

Table 5.3 presents CSCER results with estimated side information. We observe that our first stage estimate based CSCER renders performance that approaches the maximal bound validating a strong inter-subject emotional dependency that holds potential for improved CER. To validate our approach in a practical setting where true cross-subject emotions are not available on run-time, our goal was to obtain correlation close to that using the true side descriptions. We observe that our cross-subject estimated side information from the first stage approaches upper bounds set by true side information from Table 5.2. We obtain best results with speech followed by motion and multimodal data. Lower motion based performance may be due to arm and forearm features not reflecting sufficient movements in dyadic interactions. The correlation with estimated side information slightly exceeds the upper bound set by true side information. This is due to the unavoidable need of interpolation on the estimated data to align two partner sides and deliberate noise addition in the experimental process. Here again, use of cross-subject emotions is best captured in activation dimension.

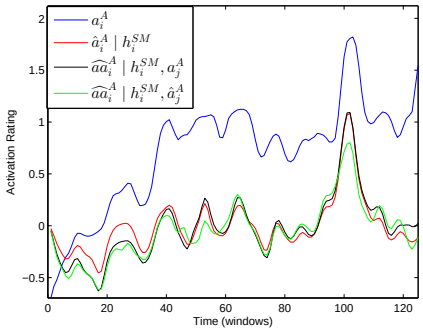
Figure 5.2 presents a sample performance comparison of activation recognition



(a) Speech



(b) Motion



(c) Multimodal

Figure 5.2: Sample comparison of (i) ground truth for Activation with estimation using (ii) modality data only, (c) modality data and true side information, and (d) modality data and estimated side information.

under CSCER using speech, motion and multimodal cues. We obtain highest performance with speech and speech couplings as shown in Figure 5.2a. Additionally, we observe that the results achieved by the first stage are improved when we incorporate the cross-subject side information.

Chapter 6

NON-VERBAL VOCALIZATIONS FOR CER

6.1 Introduction

Non-verbal vocalizations (NVV) are an integral component of spontaneous conversations and carry important emotional content within them but they are not extensively studied[43]. Studies comparing interpretation accuracy between verbal and full-channel (verbal-plus-nonverbal) cues reveal that nonverbal cues are crucial to social interpretation [44, 45]. Thus, the use of non-verbal vocalizations as feature descriptors can significantly improve continuous emotion recognition (CER) performance.

In this chapter, we hypothesize that participants in dyadic interactions exhibit higher emotion correlation around occurrences of non-verbal vocalizations as compared to verbal exchange during interactions. Particularly, some important NVV cues based on speech include laughters, sighs and breathing noises [94, 95, 96]. Voice activity features indicating silence and speech regions are also a crucial source of context information [97]. Most traditional CER systems represent features as low-level descriptors (LLDs) extracted over the entire duration of the utterance and summarize them in terms of statistical functionals. This time-series approach describes the global characteristics of the given utterance assuming that all regions of the utterance contain relevant emotional information thus diluting useful information contained in shorter salient regions.

With the motivation discussed above, we focus on capturing the advantage of emotionally rich NVVs as salient regions within speech through two CER frameworks.

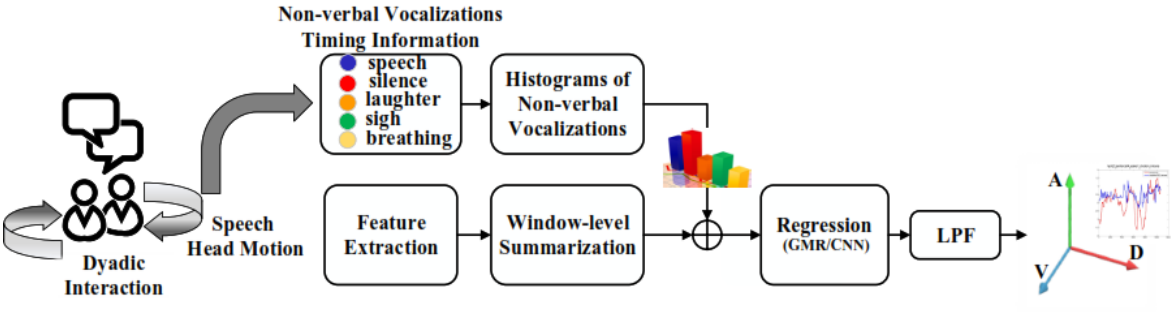


Figure 6.1: Framework for continuous emotion recognition using histograms of non-verbal vocalizations.

6.2 Methodology

The block diagram of our proposed framework is depicted in Figure 6.1. The aim of this chapter is to explore the predictive ability of NVV context, in conjugation with speech and head motion data on recognizing the underlying emotional attributes e that we define as a, v and d for activation, valence and dominance dimensions respectively. We exploit labeled data extracted from dyadic speech recordings if the USC CreativeIT database to approximate histograms of NVV context. In all five speech events are used to construct our set of non-verbal vocalizations namely laughter, sighs, breathing noises along with voice activity indicators for speech and silence.

Table 6.1 tabulates the session-wise distribution of NVV events and voice activity labels which are separated by a dotted line. Note that naturally occurring laughter, sighs and breathing noises are bound to occur sporadically thus, NVV labels, alone, are almost always insufficient to be encoded in recognition systems. One possible straightforward way of exploring their ability is to encode them with voice activity indicators which on their own are also quite useful features [97]. This way, the NVV context labels can be interpreted as key tags within the voice activity progression.

6.2.1 Feature Extraction

We define our frame-level speech and head motion feature vectors as $f^S(k)$ and $f^M(k)$ at a given frame k, respectively. Here we utilized two different acoustic feature representations, f_C^S and f_G^S, where both are computed every 16.67 ms over 25 ms windows to match the motion capture rate. The first acoustic feature f_C^S is defined as the 39 dimensional MFCC feature vector including the first 12 MFCC coefficients and energy together with the first and second time derivatives to capture the dynamic variations. The second acoustic feature f_G^S is extracted from the comprehensive *eGemaps* extended parameter set to capture frequency, energy, temporal and dynamic variation characteristics [98]. We remove static features of the *eGemaps*, and define the f_G^S feature as a reduced 81 dimensional vector. Similarly, the frame-level head motion feature set comprises of 7 head and neck features with their Euler rotation angles in (x,y,z) directions. Specifically, we use left and right front hand, left and right back hand, chest, sternum and top spine out of the full motion capture feature set. This results in a 21 dimensional head motion feature vector that we refer to as f^M.

Table 6.1: Distribution of non-verbal vocalizations across five sessions of the working USC CreativeIT database.

	session-wise count					
NVV class	**1**	**2**	**3**	**4**	**5**	**Total**
Laughter	22	15	20	18	15	90
Sigh	22	15	11	1	25	71
Breathing	36	22	29	10	29	126
Speech	380	465	494	434	489	2262
Silence	461	499	533	439	519	2451

6.2.2 Feature Summarization

In our GMR-based CER framework, we use speech, motion and multimodal data over windows of size 3 sec with an overlap of 1 sec between adjacent windows to generate summarized feature vectors taking a lead from [7]. In order to obtain statistical summarization of the speech and head motion features, we use 11 statistical aggregation functions as defined in [16] including mean, standard deviation, skewness, etc. These functions are applied to each of the frame-level speech and head motion normalized feature vectors over the duration of each window to generate a long summarized feature vector per window. We use principal component analysis (PCA) for dimensionality reduction by adjusting the number of principal components to retain more than 90% of the total variance on average. We then incorporate dynamic information by augmenting the PCA summarized features with their first order temporal derivatives. Hence, the input of the window-level feature summarization is a sequence of the frame-level feature vectors, f, and we denote the output summarized feature representation as h.

For our CNN-based CER framework, the optimization of key CNN parameters such as window size and overlap was done independently and the values of the best parameters were selected for the current chapter. Therefore, we use 2 sec windows with 0.33 sec overlap. For each window, we extract a sequence of frame-level feature vectors of speech or motion. Then, we apply z-normalization to the feature vector sequence using the global mean and standard deviation vectors of the whole dataset. By appending the histogram of NVV classes as features, we increase the feature dimensionality within each window by five.

6.2.3 NVV Histogram Features

We define the NVV context labels as five distinct classes in Table 6.1. Associated with the frame-level speech and head motion features, we also have an indicator NVV context label for the frame of interest. We compute histogram of the 5-dimensional discrete NVV context label variable over the 3 sec analysis windows and denote it as h^V. Note that the NVV histogram feature vector, h^V, is expected

to carry a summarized NVV context information for each analysis window that can contribute to estimation of the emotional attribute by defining a vocalization context. Furthermore, in our CNN-based CER framework, we append the frame-level NVV context vector to speech and head motion features thus appending the dimensionality of each by one. Note that convolutional neural networks typically self-extract and learn features thus instead of functionals and histogram summarizations, it is more appropriate to append NVV feature variations.

With the use of NVV histograms, we ensure that the non-verbal activity within the speech utterance gets better utilized in the regression. Instead of descriptions derived from the entire speech utterance, this approach, in effect, magnifies the predictive power of some salient regions within speech that are more correlated and relevant to emotional information. We now provide the details for each of these subtopics.

6.2.4 Continuous emotion recognition (CER)

GMR-based CER

In our first experiment, continuous emotion recognition framework is formulated as Gaussian mixture regression (GMR),

$$\hat{e}(h) = \arg\max_{e} P(e \mid h, \lambda(h, e)), \tag{6.1}$$

where $\lambda(h, e)$ is the joint Gaussian mixture density of the observation and emotional attributes, respectively h and e. Additional details about this framework can be looked up in [7].

CNN-based CER

Speech as well as body motion data can be interpreted as a 2D-image or a spectral pattern over temporal and spectral axis. Such a visualization of data enables analyzing patterns of spectral and euler features by applying CNNs for unimodal and multimodal speech and body motion emotion regression.

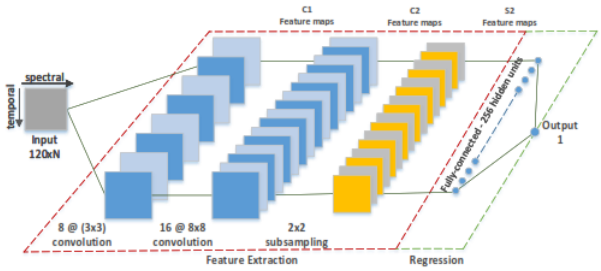


Figure 6.2: CNN architecture for the proposed CER system.

Thus, in our second experiment, we deploy CNNs for CER and demonstrate that NVV context can also bring about improvements in deep neural network based CER systems. After experimentation with a number of CNN architectures, the discussion for which we skip for brevity, the structure of CNN we chose for our experiments is shown in Figure 6.2 is summarized below. Here, we deploy 2 convolutional layers, 1 maxpooling layer, and 1 fully connected layer with a soft-max layer. The input of network is a 120 x N image generated from speech and head motion data where N is the dimensionality of each modality. The initial convolutional layer has 8 (3 x 3) kernels with a stride setting of 1 pixels. It is followed by the second convolutional layer with 16 (8 x 8) kernels. Next applied is a max pooling layer of size 2 x 2 with default stride 2. The activation at two convolutional layers is also performed by ReLU. Following maxpooling, we perform batch normalization making the training less sensitive to initialization and other optimization parameters. Then a fully-connected layer with 256 neurons is applied, followed by a soft-max layer with 1 neuron. The output corresponds to A,V and D attributes independently. Moreover, to avoid over-fitting, the first two convolutional layers and the fully-connected layer are followed by dropout layers having a ratio of 25% and 50% respectively.

6.3 Experimental evaluations

In both our continuous emotion recognition frameworks, we conduct 5-fold leave-one-session-out cross validation for experimental evaluations. For the GMR-based CER framework, training of the GMR is performed for each emotional attribute

activation, valence and dominance using 4 mixtures. We restrict the number of mixtures to 4, since it is empirically deduced that this approach is not sensitive to increased the number of mixtures used. Furthermore, each window segment is assigned the average emotional attribute of all frames within the window in the training phase.

We implement our CNN-based CER framework models in python using Keras with tensorflow backend. The training is performed on a single NVidia GeForce Titan XP GPU with 12 GB onboard memory as well as clusters on Koç University Advanced Computing Center (KUACC). We deploy Adam optimizer [99] and train a separate model for each emotional dimension. We train the models to maximize the Pearson correlation coefficient (PCC) between the ground truth continuous emotional label of the temporal window (x) and its estimate (y). Furthermore, the training process is run for 30 epochs with a batch size set to 200. Initial learning rate was set to 0.01 with a decay of 0.1 after every 20 epochs.

For both our frameworks, we further split our experiments into three. Experiment (a) includes unimodal regression evaluation, whereas experiment (b) investigates contribution of the NVV context information to the unimodal regression and experiment (c) investigates multimodal regression together with the NVV context information. Since the absolute relationship of features to emotional attributes is ensured in our experiment, we therefore report the mean absolute correlation of ground truth and estimated emotional attributes across sessions.

Table 6.2 presents correlation results for the three experiment sets of the GMR-based CER framework. Our baseline results from experiment (a) reveal that MFCC features outperform the *eGemaps* and head motion features with a relative improvement of 35.84% and 28.10% for activation attribute, respectively. Valence and dominance attributes are best predicted using the *eGemaps* and head features, respectively. Experiment (b) demonstrates that the advantage of NVV context information is best captured in speech rather than head motion. Particularly, in presence of NVV histograms, MFCC features yield the strongest correlation of 0.6021 in activation dimension, followed by head features in dominance and *eGemaps* in valence

Table 6.2: The mean absolute correlation values between estimated emotional attributes and the ground truth for the GMR-based CER regression framework with MFCC and *eGemaps* features for speech, head motion features and NVV histogram features.

Exp	**Features**	**A**	**V**	**D**
a	h_C^S	**0.5265**	0.2896	0.3088
	h_G^S	0.3876	**0.3476**	0.2886
	h^M	0.4110	0.3414	**0.3521**
b	$h_C^S + h^V$	**0.6021**	0.2812	0.3391
	$h_G^S + h^V$	0.4911	**0.2988**	0.3235
	$h^M + h^V$	0.4752	0.2751	**0.3560**
c	$h_C^S + h^M$	0.5312	0.2736	0.3281
	$h_G^S + h^M$	0.3957	**0.3595**	0.3308
	$h_C^S + h^M + h^V$	**0.5914**	0.2709	0.3723
	$h_G^S + h^M + h^V$	0.5007	0.3023	**0.3836**

dimension. The improvement in speech modality was expected since NVV context information is extracted from it. Finally, in experiment (c) the best multimodal results for activation are achieved by a combination of MFCC, head motion, NVV descriptors. Replacing MFCC with *eGemaps*, we get the best results for dominance. The best result for valence is achieved by the use of *eGemaps* along with head motion features.

Furthermore, Table 6.3 presents correlation results for the three experiment sets of our CNN-based CER framework. In these set of experiments, we exclusively exploit the MFCC features as our speech feature set since they outperform *eGemaps* in our GMR-based CER experimental analysis. Also, detailed statistical descriptions

Table 6.3: The mean absolute correlation values between estimated emotional attributes and the ground truth for the proposed CER system using CNN-based regression with *MFCC* features for speech, head motion features and NVV histogram features.

Exp	**Features**	**A**	**V**	**D**
a	h_C^S	**0.2396**	0.1091	0.1219
	h^M	0.1486	0.0884	0.1337
b	$h_C^S + h^V$	**0.3071**	0.1387	0.1858
	$h^M + h^V$	0.2382	0.1101	0.1532
c	$h_C^S + h^M$	0.5401	0.2339	0.2244
	$h_C^S + h^M + h^V$	**0.6069**	0.2886	0.2695

as captured in the *eGemaps* set are not suited for CNN architectures since CNNs are designed for minimal preprocessing.

Our baseline results from experiment (a) reveal that CNN learns MFCC speech features better than head motion features. This is expected since variations over MFCC spectrum are more prominent in comparison to head motion feature movements. As results from experiment (b) illustrate, we observe a lower relative improvement of 6.75% with speech in presence of NVV context. Based on the multimodal results from experiment (c), baseline multimodal performance of our CNN-CER framework is slightly lower than that achieved by the GMR-CER, the advantage of NVV context in predicting all three AVD attributes, however, is enhanced with multimodal data learned by the CNN.

In general, our experimental results indicate strong improvement of CER performance in activation predictions with incorporation of the NVV context information, while improvements in valence and dominance are primarily attributed to modality feature descriptions. The advantage of NVV context is better captured by the

GMR framework using only speech modality. In our unimodal experiments with CNN, valence and dominance attributes are weakly predicted using both the MFCC and head features. The is due to the fact that a CNN is sensitive to dimensions of the input image (feature matrix) which is also the reason for observing a higher multimodal performance as compared to that observed with the GMR learning over multimodal data. Head gestures are mostly correlated with prosody and verbal content thus justifying their relatively lower recognition performance relative to speech in presence of non-verbal context. Our findings reinforce that characterization of the non-verbal vocalizations reflects the best in activation attribute predictions.

Chapter 7

DYADIC AFFECT CONTEXT FOR CER

7.1 Introduction

To advance improvements in affect analysis, more attention is needed to investigate how individuals in dyadic settings or a group influence the affective states of each other [6, 7]. Moreover, in recent years, deep neural networks, particularly convolutional neural networks (CNNs) and long-short term memory neural networks (LSTMs) have been used successfully in emotion recognition from audio-visual signals. While CNNs and LSTMs independently as well as in conjugation have been applied to a diverse set of applications, formulating a generalizable mechanism for recognizing emotional states still remains a challenging task. Specifically, with regard to dyadic interactions, behaviors may be influenced by those of an interlocutor and thus still require further analysis of both participants, especially to measure and study crucial inter-dependency patterns within different data spaces. Moreover, since there is a known dependency between the participants of a conversation, devising dyadic architectures that are able to capture dyadic exchanges between interlocutors can be particularly breakthroughs. Our work thus takes motivation from this perspective. Although some recent works propose dyadic neural network architectures for discrete emotion recognition (DER) as we will discuss in Section 2, to the best of our knowledge, continuous emotion recognition (CER) under affective dyadic settings has not been explored using deep neural network architectures.

Therefore, we investigate the incorporating affect context in dyadic interactions into neural network architectures to improve continuous emotion recognition (CER) in this chapter.

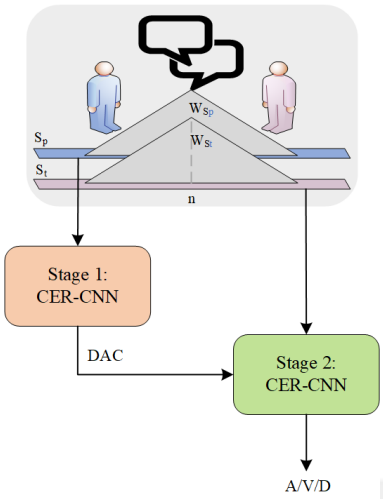


Figure 7.1: Block diagram of a general dyadic two-stage CER framework.

7.2 Methodology

This section presents a system overview, defines the affect context in dyadic settings, and introduces our proposed neural network architectures.

7.2.1 Overview

The motivation for the current investigation resides in the fact that the flow of emotions in dyadic conversation is a complex process based on a number of behavioral factors. A dyadic interaction system is expected to be causal in nature. This means that the emotional state of a subject at a particular time is a function of the emotional states of themselves and their partner up to that time. Furthermore, effects of a particular emotional instant manifest generally after some time owing to the slow-varying nature of the underlying emotions.

A block diagram outlining the general framework of our dyadic two-stage CER system is given in Figure 7.1. We define the dyadic interaction between a partner and a target subject. The fundamental goal of this study is to exploit cross-subject emotional context that for the rest of the paper we refer to as *dyadic affect context* (DAC), in realistic scenarios and integrate it within neural network frameworks to

improve emotion recognition in dyadic interaction settings. As we established in our recent analysis [7], DAC facilitates emotion recognition since emotion attributes of the partner and target subjects are significantly correlated.

Our dyadic framework constitutes of two stages, where *Stage 1* is associated with the partner subject and *Stage 2* is associated with the target subject. *Stage 1* deploys a CNN to estimate the AVD affect attributes, generally denoted as $\hat{a}_n^p$ at time frame n for the partner subject. Note that *Stage 1* performs feature-to-emotion mapping for the partner independently and outputs either affect estimates or feature maps that then defines the DAC for *Stage 2*. Outputs of the *Stage 1* are then exploited at *Stage 2* as features in either affect and feature map data fusion configurations to estimate the AVD affect attributes, generally denoted as $\tilde{a}_n^t$ for the target subject. The estimated AVD attributes are finally evaluated for performance against the corresponding mean annotated ground truths of the target subject. In the following, we introduce each of the three proposed neural network structures that are used in the rest of this manuscript. We begin with our single-subject CNN architecture in Section 7.2.2 that forms the building block of the other proposed models. We then propose dyadic two-stage CNN architectures that results in the two fusion configurations described in Section 7.2.3. Finally, based on the dyadic CNN architectures, we formalize the concept of affect context in dyadic settings and demonstrate how it can be incorporated into our novel ConvLSTM frameworks that are introduced in Section 7.2.4.

In the following, we first present the single-subject CNN architecture, then introduce the dyadic two-stage CNN architectures with two fusion configurations to merge the *Stage 1* DAC with the *Stage 2* CER network, and finally define the proposed dyadic two-stage ConvLSTM architecture.

7.2.2 Single-Subject CNN Architecture

Figure 7.2 illustrates our CNN architecture for single-subject continuous emotion recognition (CER-CNN). We consider this architecture as a baseline and as well use it as the DAC estimator in *Stage 1* for the two-stage dyadic models.

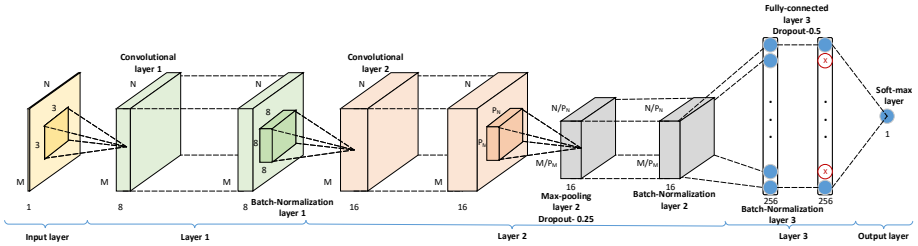


Figure 7.2: CER-CNN - Single-subject continuous emotion recognition CNN architecture.

The CER-CNN system is driven by speech spectral and/or body motion features, $\mathbf{f}_n \in \mathbb{R}^M$, which are M dimensional feature vectors that are extracted at time frame n. The input to the CER-CNN is defined as an image of speech spectral and/or body motion features over a temporal window centered at frame Rn as

$$\mathbf{F}_{Rn} = [\mathbf{f}_{Rn-l+1}...\mathbf{f}_{Rn-1}, \mathbf{f}_{Rn}, \mathbf{f}_{Rn+1}...\mathbf{f}_{Rn+l}] \in \mathbb{R}^{M \times N}, \tag{7.1}$$

where $N = 2l$ is the length of the temporal window and $\mathbf{F}_{Rn}$ is computed with R-frames skip. The CER-CNN has four layers, where the first two layers are CNN, the third layer is a fully-connected dense layer, and the last layer is a soft-max output layer. The first layer CNN has 8 number of (3×3) kernels with padding adjusted to retain spatial dimensions, and includes a batch normalization at the output. Whereas, the second layer CNN has 16 number of (8×8) kernels again with padding adjusted to retain spatial dimensions. The second layer CNN follows a max-pooling, which is performed with the pool size $(P_M \times P_N)$ and the stride factors default to pool size. The output of the second layer CNN has 25% dropout to avoid over-fitting and a batch normalization. The third layer of the CER-CNN is a fully-connected dense layer with 256 neurons, which follows a 50% dropout. The final output soft-max layer has a single output neuron, which corresponds to one of the estimated AVD affect attributes.

Note that we empirically tested a number of different CNN network configurations used in the literature before selecting the one presented here. The presented

CNN has a simple architecture and delivers competitive CER performance. The selection of a simple design is essential, especially to attain computationally feasible deep ConvLSTM architectures in the two-stage models.

7.2.3 Dyadic CNN Architectures

We consider the continuous emotion recognition in dyadic settings as a two-stage regression framework, which is outlined in Figure 7.1. In this framework let us refer to s_p as the partner subject and s_t as the target subject. Then, *Stage 1* performs CER for the partner subject to extract DAC as a form of affect estimates or feature maps. Later, *Stage 2* exploits DAC together with audio-visual inputs to perform CER for the target subject. The fusion of DAC and audio-visual inputs in *Stage 2* can be executed as data fusion (also referred as early integration) or decision fusion (late integration). Data fusion is effective when the modalities are correlated, on the other hand decision fusion is advantageous for uncorrelated modalities [46]. Furthermore, our earlier study on dyadic interaction reveals a significant correlation across subjects in the activation dimension, while the other two dimensions are also weakly correlated [7]. Based on these facts we consider data fusion to exploit DAC in two different settings, either cross-subject affect or feature maps, for the CNN based models.

Fusion of cross-subject affect (FoA)

CER-CNN performs regression of affect attribute at a rate R times slower than the frame rate. Hence in *Stage 1*, the CER-CNN performs regression of affect attribute to output $\hat{a}^p_{Rn}$ for the partner subject that corresponds to the input feature image of the partner $\mathbf{F}^p_{Rn}$. The cross-subject affect estimate, $\hat{a}^p_{Rn}$, is used as the DAC for the target subject in *Stage 2*. Figure 7.3 presents the block diagram of the dyadic CER system with the data fusion of cross-subject affect (FoA). The CER-CNN network in *Stage 2* receives augmented image of features where the cross-subject affect estimates and the audio-visual features are fused in $\mathbf{G}^t_{Rn} \in \mathbb{R}^{(M+1)\mathrm{x}N}$ at frame

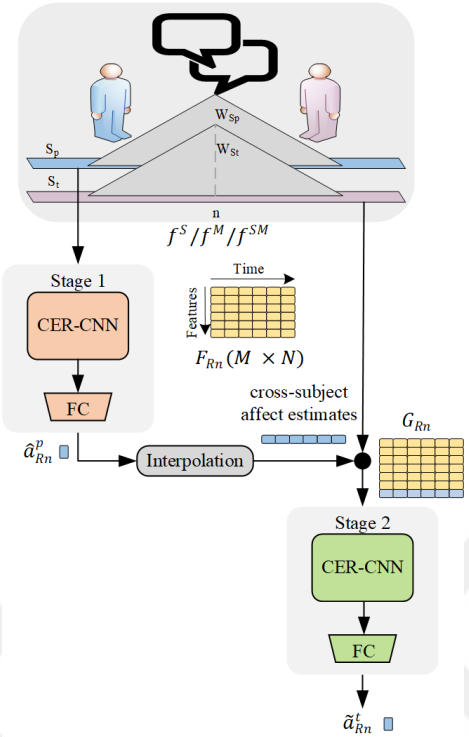


Figure 7.3: Block diagram of the dyadic CNN architecture for the fusion of cross-subject affect (FoA).

index Rn as

$$\mathbf{G}_{Rn}^{t} = \begin{bmatrix} \mathbf{f}_{Rn-l+1}^{t}, \ldots, \mathbf{f}_{Rn-1}^{t}, \mathbf{f}_{Rn}^{t}, \mathbf{f}_{Rn+1}^{t}, \ldots, \mathbf{f}_{Rn+l}^{t} \\ \hat{a}_{Rn-l+1}^{p}, \ldots, \hat{a}_{Rn-1}^{p}, \hat{a}_{Rn}^{p}, \hat{a}_{Rn+1}^{p}, \ldots, \hat{a}_{Rn+l}^{p} \end{bmatrix}. \tag{7.2}$$

Note that the sequence of affect attribute estimates, $\{\hat{a}_{Rn}^{p}\}_n$, has to be interpolated to the frame rate in order to construct $\mathbf{G}_{Rn}^{t}$. Then, the CER-CNN in *Stage 2* receives $\mathbf{G}_{Rn}^{t}$ as input and outputs affect estimate $\tilde{a}_{Rn}^{t}$ for the target subject.

Fusion of cross-subject feature maps (FoM)

In this subsection, we look at a mechanism for integration of cross-subject feature maps in an early fusion configuration. The fusion of feature maps (FoM) is configured similar to the FoA model, where the second stage closely follows CER-CNN, however, the first stage network is adapted for the FoM. Figure 7.4 presents the FoM model

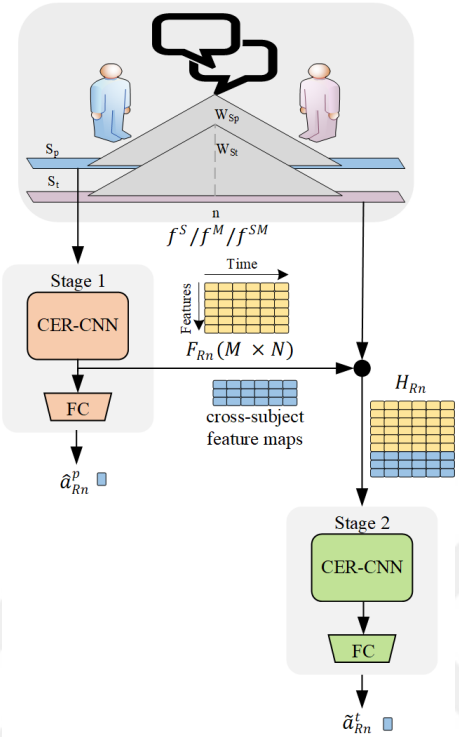


Figure 7.4: Block diagram of the dyadic CNN architecture for the fusion of cross-subject feature maps (FoM).

in dyadic CER framework.

While augmenting cross-subject data from *Stage 1* into *Stage 2*, it needs to be ensured that the cross-subject feature maps of the partner subject shall be a compact representation of the DAC and has a lower dimensional representation than the feature image of the target subject. Furthermore, temporal alignment between the cross-subject feature maps of the partner and feature image of the target subjects need to be controlled such that the temporal interpolation is not required. Keeping these two points in mind, we define a new network (FoM-CNN) for *Stage 1* by modifying the second convolutional layer of the CER-CNN to include a single (8×8) kernel, which then follows a max-pooling with size $(P_F \times 1)$. With these two modifications, when the network receives input feature image $\mathbf{F}^p_{Rn}$, the feature map image at the output of the second layer can be represented as $[\mathbf{h}^p_{Rn-l+1}...\mathbf{h}^p_{Rn-1}, \mathbf{h}^p_{Rn}, \mathbf{h}^p_{Rn+1}...\mathbf{h}^p_{Rn+l}] \in \mathbb{R}^{\lfloor M/P_F \rfloor \times N}$ for the partner subject at frame

index Rn. Then, the CER-CNN network in *Stage 2* receives augmented image of features where the cross-subject feature maps of partner and the audio-visual features of target subjects are fused in $\mathbf{H}^t_{Rn} \in \mathbb{R}^{M+\lfloor M/P_F \rfloor \times N}$ at frame index Rn as

$$\mathbf{H}^t_{Rn} = \begin{bmatrix} \mathbf{f}^t_{Rn-l+1}, \ldots, \mathbf{f}^t_{Rn-1}, \mathbf{f}^t_{Rn}, \mathbf{f}^t_{Rn+1}, \ldots, \mathbf{f}^t_{Rn+l} \\ \mathbf{h}^p_{Rn-l+1}, \ldots, \mathbf{h}^p_{Rn-1}, \mathbf{h}^p_{Rn}, \mathbf{h}^p_{Rn+1}, \ldots, \mathbf{h}^p_{Rn+l} \end{bmatrix}. \tag{7.3}$$

7.2.4 Dyadic ConvLSTM Architecture

LSTM networks are primarily exploited to model memory and context in a large range of applications. LSTMs are units of recurrent neural networks (RNNs) that are often used to relate signals over arbitrary time lags. Although LSTMs can capture temporal relationships, they cannot capture spatial relationships in data. CNNs, on the other hand, owing to their convolutional structures can capture patterns in the data better. In order for LSTMs to consider convolution structures, a multi-dimensional architecture is required that forms the basis of ConvLSTM architecture.

In traditional LSTMs, the input-to-state and state-to-state transitions are fully connected which means that dense layers are applied to data. Thus, temporal dependencies of data are retained, however, spatial dependencies are not taken care of. In ConvLSTM structure, the input and recurrent transformations are both convolutional. The issues discussed above are addressed by the ConvLSTMs since they retain shape of data along with the temporal flow. LSTM captures temporal dependencies while on the spatial axis, which is spectral and motion dimensions in our case, CNN learns features and captures feature patterns.

Recent literature suggests that context modeling may provide useful insights into emotion triggering concepts leading to improved predictions of individual's emotional state [100]. In other recent studies, context is modeled as a linear combination of past emotion states as well as past and current input features [37]. While these studies are increasingly investigating the contribution of context modeling, particularly for CER, the amount and type of context required and an optimal strategy to model it, is still an open question.

Extending the concept of context, we introduce the concept of *dyadic context* as a construct to simplify the inclusion of causal cross-dyadic relationships in the dyadic CNN. We deploy ConvLSTM framework to model dyadic affect context. We define dyadic context as a method where the ConvLSTM inputs will include partner affect information from past feature windows. The partner affect data from the same as well as previous time then constitutes the context to a conversation for a particular frame index Rn. In the current chapter, we formulate context as the extent of past image of features a ConvLSTM model looks at for the affect prediction.

Before moving on to discuss our design, we note that in prior literature reporting the use of ConvLSTMs for various applications, significant data variation is ensured from one frame to the next. Since emotions are slow-varying, consecutive feature windows may not always have significant additional information. Moreover, input size of the ConvLSTM network is reciprocal to its memory requirements.

Recent studies have also established that the activation dimension significantly benefits from cross-subject emotional dependencies [7]. Moreover, valence dimension is observed to be more sensitive to past context in the temporal space [70]. We extend these ideas and further hypothesize that using the CNN structure, we can capture and represent feature correlations over a larger spatio-temporal feature space in the form of intermediate feature maps learned by the CNN. Further exploiting these dual-space representations in an early fusion setting can benefit valence and dominance prediction performance. Based on this motivaton and the dyadic CNN models in Section 7.2.3, we propose two ConvLSTM networks. The FoA-based ConvLSTM network is used for Activation prediction while the FoM-based ConvLSTM network is used for Valence and Dominance prediction. We begin by describing FoA-based ConvLSTM network illustrated in Figure 7.5.

At *Stage 1*, the CER-CNN performs regression of affect attribute to output $\hat{a}^{p}_{Rn}$ for the partner subject with the corresponding input feature image $\mathbf{F}^{p}_{Rn}$. At *Stage 1*, we deploy the same model as described in Section 7.2.3. At *Stage 2*, the input to the CER-ConvLSTM is a temporal sequence of K feature images in the form of a

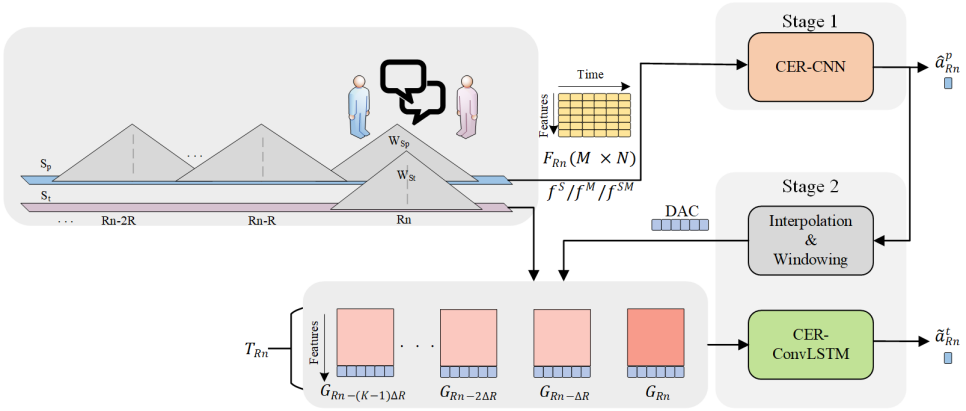


Figure 7.5: The proposed CER-ConvLSTM framework based on the FoA model for the prediction of activation attribute in dyadic interactions.

3D tensor defined as

$$\mathbf{T}_{Rn}^{FoA} = \{\mathbf{G}_{R(n-k\Delta)}^{t}\}_{k=0}^{K-1}, \tag{7.4}$$

where Δ is the context stride or the shift between selected feature images along the context axis. The input feature image $\mathbf{G}_{R(n-k\Delta)}^{t}$ is constructed in the same way as described in Section 7.2.3 and the 3D tensor, $T_{Rn}^{FoA} \in \mathbb{R}^{(M+1)\times N\times K}$, includes a sequence of K feature images ending at the current frame index Rn. Note that K and Δ are our context parameters and different settings of these parameters can conveniently deliver varying duration of affect context.

Recall that for a FoM-based model, we perform an early fusion of the cross-subject feature maps of the partner speaker obtained using the FoM-CNN in *Stage 1*. Similar to the FoA-based ConvLSTM discussed above, in the FoM-based ConvLSTM framework, we use a temporal sequence of H_{Rn} windows defining T_{Rn}^{FoM} as

$$\mathbf{T}_{Rn}^{FoM} = \{\mathbf{H}_{R(n-k\Delta)}^{t}\}_{k=0}^{K-1}, \tag{7.5}$$

where $T_{Rn}^{FoM} \in \mathbb{R}^{(M+\lfloor M/P_F \rfloor)\times N\times K}$.

The input organization to the *Stage 2* CER-ConvLSTM network is shown in Figure 7.6. We tested several variants of the ConvLSTM models and empirically opt for the model illustrated in Figure 7.7. The *Stage 2* CER-ConvLSTM has three

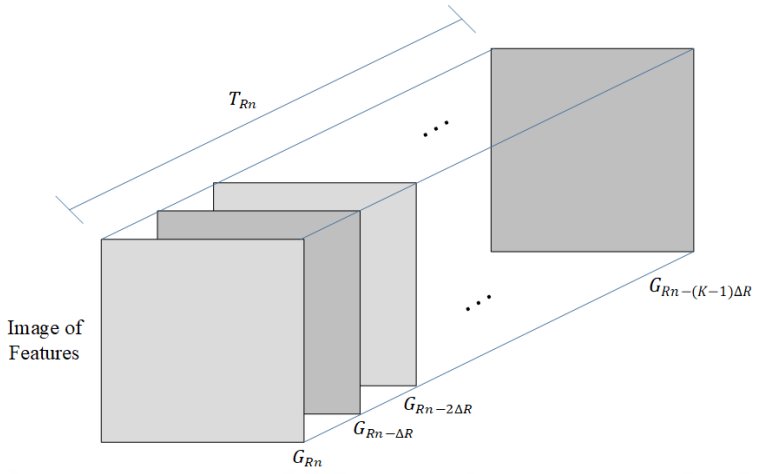


Figure 7.6: Organization of input to CER-ConvLSTM.

layers, where the first layer is a Convolutional LSTM layer, the second layer is a fully-connected dense layer, and the last layer is a soft-max output layer. The convolutional LSTM layer has 8 number of (5×5) kernels with padding adjusted to retain spatial dimensions, and includes a batch normalization at the output. The activation at this layer is performed by tanh (selected instead of ReLU, see the Results section). The ConvLSTM layer is followed by max-pooling with pool size (4×1) and the stride factors default to pool size that in effect sub-samples the feature map by 4 along the spectral dimension. The output of the first layer has 25% dropout to avoid over-fitting. The second layer of the CER-ConvLSTM is a

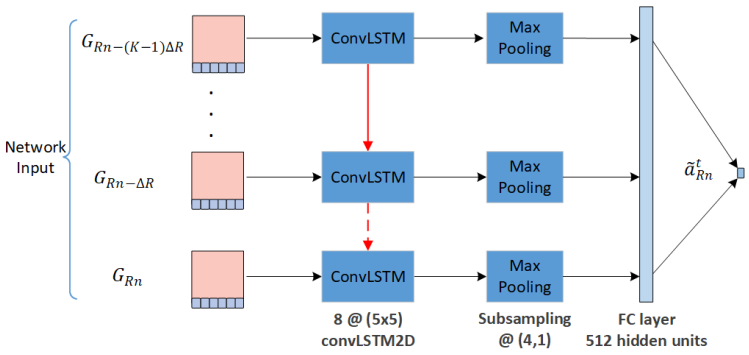


Figure 7.7: ConvLSTM architecture at *Stage 2*.

fully-connected layer with 512 neurons, which is followed by a 50% dropout. The final output soft-max layer has a single output neuron, which corresponds to one of the estimated AVD affect attributes, $\tilde{a}^t_{Rn}$, for the target subject.

Note that at *Stage 2* of the ConvLSTM networks, if regression rate is equal to data rate, no interpolation is required, however, this comes at a very high computational expense. Moreover, since emotions are slow-varying, contiguous selection of context windows causes redundancy and leads to wastage of resources. At the same time, estimates from *Stage 1* govern the re-estimation performance at *Stage 2*, thereby a well-weighted trade-off needs to be set.

Thus, for CER, the ConvLSTM model can be adapted to model the local spectro-temporal correlations in speech and body motion data as well as the long term dependencies in the feature evolution. While in speech recognition tasks, the ConvLSTM method demonstrates potential with a 4-6% relative improvement in word error rate (WER) over an LSTM [101], and to the best of our knowledge this thesis is the first attempt at using ConvLSTM models for CER using speech and body motion cues.

7.3 Experimental Results

In this section, we describe the feature representations of the speech and motion modalities, and the experimental setup followed by our experimental results and discussion. Please note that we use the USC CreativeIT database in the current work [72, 73]. A pre-analysis of the database is performed to investigate the correlation

Table 7.1: Cross-subject correlations (PCC) of the AVD in dyadic settings.

		Subject 1		
		a_1^A	a_1^V	a_1^D
Subject 2	a_2^A	0.5783	-0.2887	0.1283
	a_2^V	-	0.2248	-0.1974
	a_2^D	-	-	-0.2617

across the emotional attributes of the partner and the target subjects engaged in a dyadic interaction. Ground truth annotations are extracted as the mean inter-annotator ratings for the three emotional attributes of the USC CreativeIT dataset. The Pearson correlation coefficient (PCC) has been used as the evaluation metric both for the pre-analysis and later for the CER to assess goodness of the predicted affect attributes.

Table 7.1 presents cross-subject correlations of the affect attributes, where the correlation across activation attribute is significant. The cross-subject and cross-attribute correlations are generally lower for valence and dominance.

7.3.1 Feature Representations

Frame level acoustic features are extracted as 39 dimensional MFCC feature vectors including energy, the first 12 MFCCs together with the first and second time derivatives. Frame level motion feature vector is 24 dimensional and includes the Euler rotation angles in directions (x, y, z) of the arm and forearm joints together with their first derivatives over each frame. An acoustic frame is computed every 16.67 ms over a 25 ms analysis window where consecutive windows have an 8.33 ms overlap. Thus, we attain the same frame rate as motion capture, which is 60 fps.

7.3.2 Experimental Setup

In the experimental evaluations, we utilized subject independent train/test protocols. For subject independent evaluations, we perform leave-one-session-out (LOSO) cross-validation. All the models described in this work are implemented in Python using Keras with a tensorflow backend. All the experiments are run on a single NVidia GeForce Titan XP GPU with 12 GB onboard memory for 20 epochs with an initial learning rate of 0.01 and a decay of 0.1 after each 20 epochs. Networks are optimized by using the Adam optimizer [99]. Loss function for the CNNs is used as the mean square error (MSE).

Note that in our dyadic FoA model, interpolation may be required on the predicted affect estimates at output of *Stage 1* so that they can be augmented with

Table 7.2: PCC performance of the unimodal and multimodal AVD prediction with the Single-Subject CNN architecture (CER-CNN) for varying temporal window size N.

	Speech			**Motion**			**Multimodal**		
N	**A**	**V**	**D**	**A**	**V**	**D**	**A**	**V**	**D**
60	0.2042	0.0791	0.0892	0.1157	0.0449	0.0704	0.4299	0.2373	0.2229
50	0.2101	0.0650	0.0793	0.0797	0.0923	0.0585	0.4108	0.2301	0.2127
40	0.2189	0.0746	0.1087	0.0910	0.0799	0.0748	**0.4252**	**0.2822**	**0.2494**
30	0.2082	0.0980	0.0964	0.1038	0.0514	0.0596	0.4065	0.2568	0.2356
20	0.1911	0.1002	0.0771	0.0993	0.0779	0.0743	0.4012	0.2695	0.2245
10	0.1963	0.0942	0.0648	0.0882	0.0831	0.0628	0.3861	0.2278	0.2024

feature data of the target subject to form *Stage 2* input. Feature data is available at 60 Hz which is the data rate. At *Stage 1*, we perform windowing every $R = 10$ frames such that our selected windowing rate is 6 Hz which is also the regression rate since we obtain one regression per window. At the output of *Stage 1*, we have emotion estimates available at 6 Hz. We perform spline interpolation to up-sample the rate from 6 Hz to 60 Hz while passing the dyadic affect context from *Stage 1* to *Stage 2*.

7.3.3 Single-Subject CNN Architecture Results

We begin by presenting our Single-Subject CNN architecture results. Table 7.2 summarizes our CER-CNN model baseline results in PCC for varying temporal window size, N, on the unimodal and multimodal prediction of AVD. Results are obtained with batch size 100, number of epochs 20, and the frame skip of $R = 10$ as described in Section 7.2.2. We observe that increasing N from 10 to 60 generally improves the CER performance since the network is enabled to iterate training over larger data. We obtain best multimodal performance at $N = 40$ where prediction of activation outperforms prediction of valence and dominance. After

fixing the temporal window size to $N = 40$, we investigate the effect of batch size and activation function type to the prediction of activation dimension in the multimodal setup. Table 7.3 presents the PCC performance of the activation dimension for varying batch sizes from 100 to 1000, where the performance is decreasing as batch size is increasing. In general, batch size selection is constraint by the available memory resources. Larger batches are computationally faster but they require more memory. We choose to set batch size as 100 since it delivers the best performance and eases on memory resources. Further on, Table 7.4 presents the effect of activation function to the prediction performance. Conventionally, CNNs are used with ReLU activation and while ReLU function and its derivative is monotonic and gives good performance for data in [0, 1], the downside is that all the negative values become zero immediately which decreases the ability of the model to properly fit or train from the data outside the [0, 1] range. Our results reveal that tanh activation fits better to our data as the features also reside in the negative space. Based on this finding, we deployed tanh activation for the following experiments.

Table 7.3: PCC performance of the multimodal activation prediction with the Single-Subject CNN architecture (CER-CNN) for varying batch sizes (temporal window size is fixed to $N = 40$).

Batch Size	100	250	500	750	1000
Multimodal	**0.4252**	0.4086	0.3937	0.4002	0.3619

Table 7.4: PCC performance of the multimodal activation attribute prediction with the Single-Subject CNN architecture (CER-CNN) for various activation functions.

Activation Function	ReLU	elu	selu	sigmoid	tanh
Multimodal	0.4252	0.4200	0.4283	0.4132	**0.4519**

7.3.4 Dyadic CNN Architecture Results

We investigate two data fusion schemes with the dyadic CNN architectures using either cross-subject affect (the FoA model) or feature maps (the FoM model). Table 7.5 presents the PCC performance of the FoA model for speech, motion and multimodal cues. In general multimodal cue performs better than the unimodal cues and the prediction of the activation attribute from multimodal cue is significantly better than the others by reaching a higher PCC (0.5363) than the one (0.4519) achieved with the CER-CNN model.

Table 7.5: PCC performance of the unimodal and multimodal CER with the dyadic FoA model for the AVD attributes.

Features	**A**	**V**	**D**
Speech	0.2411	0.0931	0.0822
Motion	0.1143	0.0792	0.0762
Multimodal	**0.5363**	**0.1607**	**0.2444**

Table 7.6 presents the PCC performance of the FoM model for speech, motion and multimodal cues for varying values of the max-pooling factor, P_F. With the fusion of feature maps the CER performance is better especially for the valence and dominance attributes. The highest VD prediction performances are obtained with the max-pooling factor $P_F = 3$ for the multimodal cues achieving correlations of 0.5061 and 0.5588, respectively. The multimodal FoM model at $P_F = 4$ also provides good CER performance for the activation attribute but not as high as the FoA model.

7.3.5 Dyadic ConvLSTM Architecture Results

Relying on the findings of the dyadic CNN architecture, we construct our CER-ConvLSTM architecture for the prediction of activation attribute based on the FoA model while that for the prediction of valence and dominance attributes based on

Table 7.6: PCC performance of the unimodal and multimodal CER with the dyadic FoM model for the AVD attributes for varying max-pooling factor P_F.

P_F	**Speech**			**Motion**			**Multimodal**		
	A	**V**	**D**	**A**	**V**	**D**	**A**	**V**	**D**
1	0.0816	0.0739	0.0551	0.0703	0.0403	0.0560	0.4401	0.4562	0.3880
2	0.1467	0.0734	0.0715	0.0652	0.0327	0.0542	0.4479	0.3186	0.3760
3	0.2123	0.0820	0.0961	0.0997	0.0352	0.0717	0.4993	**0.5061**	**0.5588**
4	0.2149	0.0792	0.0733	0.0414	0.0319	0.0468	**0.5130**	0.3494	0.3720

the FoM model. Performance of the CER-ConvLSTM is investigated with the multimodal input for varying values of the context parameters, which are the number of feature images K and the context stride Δ. Table 7.7 presents the PCC performance of the CER-ConvLSTM for prediction of valence and dominance attributes with the FoM-based ConvLSTM and activation attribute with the FoA-based ConvLSTM. The proposed ConvLSTM architectures achieves significant improvements over baseline results, reported in Table 7.2, for all three emotion dimensions. Using the FoA and FoM-based ConvLSTM models, we observe maximum improvement of 22.76%, 28.97% and 33.99% for activation, valence and dominance, respectively. For activation, FoA-based ConvLSTM framework provides highest PCC performance at 0.6528 with the context duration $K = 12$ and context stride $\Delta = 2$. In particular, for valence and dominance, FoM-based ConvLSTM framework provides highest relative improvement with PCC score of 0.5779 and 0.5893 at $\{K = 18, \Delta = 2\}$ and $\{K = 9, \Delta = 3\}$, respectively. To compare against CER results of other studies on the same database, we see from the results summary reported in Table 1 that the best reported results of PCC from various studies are 0.62/0.37/0.37 for A/V/D, respectively. This translates to PCC improvements of 5.29% for Activation, 56.19% for Valence and 59.27% for Dominance recognition in comparison to existing studies for the same database.

Figure 7.8 presents the PCC performance of multimodal AVD prediction as a

Table 7.7: PCC performance of the multimodal AVD prediciton with the Dyadic CER-ConvLSTM architecture for varying values of the number of feature images K and the context stride Δ.

	Activation (*FoA*-based ConvLSTM)			**Valence (*FoM*-based ConvLSTM)**			**Dominance (*FoM*-based ConvLSTM)**		
K	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$	$\Delta = 1$	$\Delta = 2$	$\Delta = 3$
3	0.5089	0.5296	0.5426	0.3917	0.4584	0.4890	0.5015	0.2831	0.5391
6	0.4997	0.5623	0.5876	0.4135	0.4659	0.3643	0.4525	0.3847	0.5138
9	0.5457	0.6178	0.6063	0.4991	0.4893	0.5028	0.4376	0.4823	**0.5893**
12	0.5548	**0.6528**	0.5963	0.4854	0.4286	0.3639	0.4620	0.4936	0.4618
15	0.5715	0.6437	0.6013	0.4865	0.4639	0.4313	0.4358	0.4864	0.3606
18	0.5755	0.6008	0.6115	0.4097	**0.5779**	0.4618	0.4368	0.4916	0.4336
21	0.5780	0.5812	0.5742	0.4823	0.5001	0.2840	0.4259	0.4392	0.3605

function of context duration. We observe that the best past context duration for Activation, Valence, Dominance prediction is 3.667, 5.667, 4 secs respectively. This suggests that amongst all three dimensions, Valence is most dependent on past context. Our observed context duration ranges between 0.33 to 10 secs. Additionally, for all the affect dimensions, we observe a performance improvement with a context stride greater than 1 that indicates the benefit of decimation and looking further back in time. Covering contexts as large as $K = 21$ windows yields improvements to performance as compared to results obtained from covering no or smaller contexts. Our context duration findings are analogous to kernel size results that suggest that increasing the receptive field of filters at convolutional layers can improve CER.

Our results demonstrate similar findings reported recently in [37] where valence prediction is studied as a function of past temporal dependencies. Their results show that valence is highly dependent on the past valence intensities, as well as the past speech characteristics. Another recent study in [70] models speaker and interlocutor's context using self attention mechanism. Their results also suggest

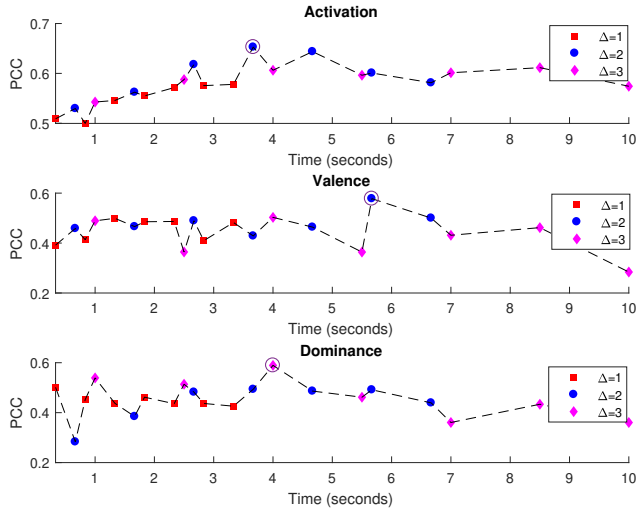


Figure 7.8: PCC Performance of multimodal AVD prediction as a function of context duration $(R \times (K-1) \times \Delta/60)$ secs.

that while the arousal dimension is less sensitive to context, the valence dimension captures significant improvements over the LSTM and no context DNN models. A possible reasoning to this improvement is attributed to valence dimension capturing the polarity of the emotion, i.e. whether emotion is positive or negative, and the assessment of this effect is reflected in relatively long-term variations explaining the benefit of both self and interlocutor's contexts for the valence recognition. Additionally, our dominance prediction results also reveal that the advantage of past context is significant however, unlike valence, relatively shorter context yields better performance. In general, our results validate that modeling and incorporating dyadic affect context of the partner subject for emotion prediction of the target subject, capturing long-term spectro-temporal contextual information and incorporating it within the ConvLSTM framework is beneficial for CER and provides significant advantage over no context baseline results.

Chapter 8

CONCLUSIONS AND FUTURE WORK

We conclude this thesis by discussing some take-home messages from the experimental results across the conducted studies.

Firstly, our results show that speech spectral features hold potential in recognizing emotional states. While body motion features do not fare well for emotion prediction on their own, they complement speech features rather nicely to bring significant improvement to CER for multimodal scenarios. Secondly, we see that amongst the three attributes capturing the global emotional meaning, activation attribute of subjects is the most correlated as it measures the strength of the emotion expressed in a conversation. This means along the coarse of interaction, subjects are influenced by each other's emotional intensities. As a next step, the nature and directionality of affective influences can be studied where they may be in synchrony, diverging or accommodating in nature. Effect of several other feature types can be explored. Also, the proposed framework can be implemented in an end-to-end fashion where CNNs perform automatic feature learning.

During dyadic interaction, subjects influence and adapt to affective state changes of their partners. This manifestation can be observed as cross-speaker, cross-modality and non-verbal vocalization analysis. Furthermore, any of these could be used to improve the performance of CER. The two-stage networks proposed by various experiments in this thesis show that re-estimation of emotional attributes using the properties of synchronies we established, we can improve the performance of CER. This holds despite the type of regression network used at each stage. Note that the underlying hypothesis for this study is that there exists an inter-dependence between dimensions of emotion at a cross-subject level and this dependency should reflect itself in the prediction of emotion in a dyadic setting. To this end, the data resources

relevant for this study require emotional annotations of both dyadic subjects. Thus, our proposed contribution is useful in dyadic scenarios where the emotional changes of both subjects are tracked. Dyadic scenarios where this work may not be beneficial include scenarios where multimodal data of one of the subjects of an interaction is not available or where the multimodal data of each subject is exploited to improve only it's own emotion prediction.

We also observe that CNNs are suitable for studying CER during dyadic interactions. We propose a two-stage CNN-based CER framework that incorporates dyadic interaction setup and enables us to realize the potential of cross-subject affect in realistic scenarios. The predicted affective states, as well as intermediate feature maps, are utilized to strengthen the predictive power of AVD. We have investigated methods to explore cross-subject dependencies in early fusion settings. Our experimental results reveal that early fusion is suitable since cross-subject emotions are strongly correlated. We incorporate cross-subject dependencies within the proposed CNN frameworks.

Specifically, in terms of network performance for the multimodal CER task, we observe that the dyadic FoA-based ConvLSTM is the best performing model for Activation prediction whereas dyadic FoM-based ConvLSTM is the best performing model for Valence and Dominance prediction. Comparisons with the existing literature indicate significant performance improvements in Valence and Dominance predictions that indicates strong potential of the proposed dyadic FoM.

Dyadic affect context is a relatively new concept that we aimed to formulate and model using ConvLSTM framework as the major contribution of this work. In fact we observe that spectro-temporal and spectro-emotion dependencies bring significant advantage in improving the CER performance. Experimental results of the proposed CER-ConvLSTM architecture for dyadic interaction show that exploiting DAC across subjects holds significant potential to further improve CER. In the future, affect modeling can also be done using attention mechanisms where ConvLSTMs can be used to identify emotional salient parts of the data modality.

It is important to note that real-time continuous annotations implicate subject

dependent delays that are inevitable due to perceptual processing, and to some extent persist even when annotation procedures are modified [73]. To mitigate this challenging limitation of annotations, ConvLSTMs provide an excellent framework where the effect of delays in emotion ratings are not only neutralized but affect context history enables studying the effects of long as well as short term emotional dependencies during the dyadic exchanges.

BIBLIOGRAPHY

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[2] S. Tomkins, *Affect imagery consciousness: Volume I: The positive affects*, vol. 1. Springer publishing company, 1962.

[3] J. A. Russell, "Core affect and the psychological construction of emotion.," *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[4] H. Schlosberg, "Three dimensions of emotion.," *Psychological review*, vol. 61, no. 2, p. 81, 1954.

[5] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[6] W. Mou, H. Gunes, and I. Patras, "Your fellows matter: Affect analysis across subjects in group videos," in *14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–5, IEEE, 2019.

[7] S. N. Fatima and E. Erzin, "Cross-subject continuous emotion recognition using speech and body motion in dyadic interactions," *Proceedings of Interspeech*, pp. 1731–1735, 2017.

[8] S. N. Fatima and E. Erzin, "Use of affect context in dyadic interactions for continuous emotion recognition," *submitted to Speech Communication*, 2020.

[9] S. N. Fatima and E. Erzin, "Acoustic feature selection for natural dialog identification," in *3rd International Workshop on Affective Social Multimedia Computing (ASMMC)*, IEEE, 2017.

[10] S. N. Fatima and E. Erzin, "Use of non-verbal vocalizations for continuous emotion recognition from speech and head motion," in *14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 433–437, IEEE, 2019.

[11] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10, pp. 787–800, 2007.

[12] F. Eyben, M. Wöllmer, A. Graves, B. Schuller, E. Douglas-Cowie, and R. Cowie, "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 7–19, 2010.

[13] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis, "Multimodal emotion recognition from expressive faces, body gestures and speech," in *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 375–388, Springer, 2007.

[14] G. Castellano, L. Kessous, and G. Caridakis, "Emotion recognition through multiple modalities: face, body gesture, speech," in *Affect and Emotion in Human-Computer Interaction*, pp. 92–103, Springer, 2008.

[15] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2288–2291, IEEE, 2011.

[16] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, 2013.

[17] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[18] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived pca-based quantification scheme with application to affect analysis in married couple interactions," *Computer Speech & Language*, vol. 28, no. 2, pp. 518–539, 2014.

[19] H. Khaki and E. Erzin, "Use of agreement/disagreement classification in dyadic interactions for continuous emotion recognition," *Proceedings of Interspeech, San Francisco, USA*, 2016.

[20] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "Avec 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, pp. 449–456, ACM, 2012.

[21] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: the continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, ACM, 2013.

[22] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, and M. Pantic, "Avec 2015: The 5th international audio/visual emotion challenge and workshop," in *Proceedings of the 23rd ACM International Conference on Multimedia*, pp. 1335–1336, ACM, 2015.

[23] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pp. 3–10, ACM, 2016.

[24] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pp. 3–9, ACM, 2017.

[25] H. Khaki and E. Erzin, "Use of affect based interaction classification for continuous emotion tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2881–2885, IEEE, 2017.

[26] C.-M. Chang, B.-H. Su, S.-C. Lin, J.-L. Li, and C.-C. Lee, "A bootstrapped multi-view weighted kernel fusion framework for cross-corpus integration of multimodal emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*, pp. 377–382, IEEE, 2017.

[27] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis, and E. M. Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," *arXiv preprint arXiv:1708.07050*, 2017.

[28] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[29] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada*, 2018.

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

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

[32] J. Huang, Y. Li, J. Tao, Z. Lian, and J. Yi, "End-to-end continuous emotion recognition from video using 3d convlstm networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6837–6841, IEEE, 2018.

[33] J. Lee, S. Kim, S. Kiim, and K. Sohn, "Spatiotemporal attention based deep neural networks for emotion recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1513–1517, IEEE, 2018.

[34] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1, pp. 5–32, 2003.

[35] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5200–5204, IEEE, 2016.

[36] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[37] V. S. E. A. Anda Ouyang, Ting Dang, "Speech based emotion prediction: Can a linear model work?," in *Proceedings of Interspeech*, pp. 2813–2817, 2019.

[38] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[39] H. P. Graf, E. Cosatto, V. Strom, and F. J. Huang, "Visual prosody: Facial movements accompanying speech," in *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on*, pp. 396–401, IEEE, 2002.

[40] H. Gunes and M. Pantic, "Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners," in *Intelligent Virtual Agents*, pp. 371–377, Springer, 2010.

[41] A. Samanta and T. Guha, "On the role of head motion in affective expression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2886–2890, IEEE, 2017.

[42] Z. Hammal, J. F. Cohn, C. Heike, and M. L. Speltz, "What can head and facial movements convey about positive and negative affect?," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pp. 281–287, IEEE, 2015.

[43] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion," *The Quarterly Journal of Experimental Psychology*, vol. 63, no. 11, pp. 2251–2272, 2010.

[44] D. Archer and R. M. Akert, "Words and everything else: Verbal and nonverbal cues in social interpretation.," *Journal of Personality and Social Psychology*, vol. 35, no. 6, p. 443, 1977.

[45] M. Argyle, V. Salter, H. Nicholson, M. Williams, and P. Burgess, "The communication of inferior and superior attitudes by verbal and non-verbal signals," *British Journal of Clinical Psychology*, vol. 9, no. 3, pp. 222–231, 1970.

[46] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.

[47] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 689–696, 2011.

[48] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[49] T. Toda, A. W. Black, and K. Tokuda, "Statistical mapping between articulatory movements and acoustic spectrum using a gaussian mixture model," *Speech Communication*, vol. 50, no. 3, pp. 215–227, 2008.

[50] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[51] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "Lstm-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, 2013.

[52] M. Wöllmer, F. Eyben, S. Reiter, B. W. Schuller, C. Cox, E. Douglas-Cowie, R. Cowie, *et al.*, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies.," in *Proceeedings of Interspeech*, vol. 2008, pp. 597–600, 2008.

[53] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[54] Z. Aldeneh and E. M. Provost, "Using regional saliency for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2741–2745, IEEE, 2017.

[55] M. Schmitt, N. Cummins, and B. Schuller, "Continuous emotion recognition in speech–do we need recurrence?," *Training*, vol. 34, no. 93, p. 12, 2019.

[56] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[57] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition," in *Proceedings of Interspeech*, pp. 272–276, 2018.

[58] H. Kaya, D. Fedotov, A. Yeşilkanat, O. Verkholyak, Y. Zhang, and A. Karpov, "Lstm based cross-corpus and cross-task acoustic emotion recognition," in *Proceedings of INTERSPEECH*, pp. 521–525, 2018.

[59] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, pp. 65–72, 2018.

[60] J. Zhao, R. Li, J. Liang, S. Chen, and Q. Jin, "Adversarial domain adaption for multi-cultural dimensional emotion recognition in dyadic interactions," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pp. 37–45, 2019.

[61] S. Karita, A. Ogawa, M. Delcroix, and T. Nakatani, "Forward-backward convolutional lstm for acoustic modeling," in *Proceedings of Interspeech*, pp. 1601–1605, 2017.

[62] B. H. Kim and S. Jo, "Deep physiological affect network for the recognition of human emotions," *IEEE Transactions on Affective Computing*, 2018.

[63] J. K. Burgoon, L. A. Stern, and L. Dillman, *Interpersonal adaptation: Dyadic Interaction Patterns*. Cambridge University Press, 2007.

[64] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, 2013.

[65] Z. Yang and S. Narayanan, "Analyzing temporal dynamics of dyadic synchrony in affective interactions," in *Proceedings of Interspeech*, pp. 42–46, 2016.

[66] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pp. 1–8, IEEE, 2013.

[67] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, B. Schuller, K. Star, *et al.*, "Sewa db: A rich database for audio-visual emotion and sentiment research in the wild," *arXiv preprint arXiv:1901.02839*, 2019.

[68] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2122–2132, 2018.

[69] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[70] J. L. Q. J. Jinming Zhao, Shizhe Chen, "Speech emotion recognition in dyadic dialogues with attentive interaction modeling," in *Proceedings of Interspeech*, pp. 1671–1675, 2019.

[71] M. Neumann and N. T. 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.

[72] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database: A multimodal database of theatrical improvisation," *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, p. 55, 2010.

[73] A. Metallinou, Z. Yang, C.-c. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The usc creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations," *Language resources and evaluation*, vol. 50, no. 3, pp. 497–521, 2016.

[74] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[75] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[76] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Picheny, and J. Pitrelli, "A corpus-based approach to expressive speech synthesis," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[77] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The hmm-based speech synthesis system (hts) version 2.0.," in *SSW*, pp. 294–299, Citeseer, 2007.

[78] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 373–376, IEEE, 1996.

[79] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter," *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 527–535, 2007.

[80] M. R. Barrick and M. K. Mount, "The big five personality dimensions and job performance: a meta-analysis," *Personnel psychology*, vol. 44, no. 1, pp. 1–26, 1991.

[81] M. Warren, *Features of naturalness in conversation*, vol. 152. John Benjamins Publishing, 2006.

[82] A. W. Black, "Unit selection and emotional speech," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[83] M. Schröder, "Emotional speech synthesis: a review.," in *Proceedings of Interspeech*, pp. 561–564, 2001.

[84] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–284, 2012.

[85] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features extended abstract," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pp. 484–490, IEEE, 2015.

[86] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, P. G. Georgiou, and S. Narayanan, "An analysis of pca-based vocal entrainment measures in married couples' affective spoken interactions.," in *Proceedings of Interspeech*, vol. 2011, p. 12th, 2011.

[87] C. M. Lee, S. Narayanan, and R. Pieraccini, "Recognition of negative emotions from the speech signal," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*, pp. 240–243, IEEE, 2001.

[88] B. Schuller, M. Lang, G. Rigoll, *et al.*, "Recognition of spontaneous emotions by speech within automotive environment," *Fortschritte der Akustik*, vol. 32, no. 1, p. 57, 2006.

[89] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge.," in *Proceedings of Interspeech*, pp. 3201–3204, 2011.

[90] M. Wester, M. P. Aylett, M. Tomalin, and R. Dall, "Artificial personality and disfluency.," in *Proceedings of Interspeech*, pp. 3365–3369, 2015.

[91] R. Dall, M. Wester, and M. Corley, "The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech.," in *Proceedings of Interspeech*, pp. 56–60, 2014.

[92] R. Dall, J. Yamagishi, and S. King, "Rating naturalness in speech synthesis: The effect of style and expectation," *Proceedings Speech Prosody, Dublin, Ireland*, 2014.

[93] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[94] R. Gupta, N. Nath, T. Agrawal, P. Georgiou, D. Atkins, and S. Narayanan, "Laughter valence prediction in motivational interviewing based on lexical and acoustic cues," *Proceedings of Interspeech*, pp. 505–509, 2016.

[95] K. H. Teigen, "Is a sigh "just a sigh"? sighs as emotional signals and responses to a difficult task," *Scandinavian journal of Psychology*, vol. 49, no. 1, pp. 49–57, 2008.

[96] R. Gupta, C.-C. Lee, and S. Narayanan, "Classification of emotional content of sighs in dyadic human interactions," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 2265–2268, IEEE, 2012.

[97] 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*, pp. 205–211, ACM, 2004.

[98] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

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

[100] E.-M. M. G. R. Aniruddha Tammewar, Alessandra Cervone, "Modeling user context for valence prediction from narratives," in *Proceedings of Interspeech*, pp. 3252–3256, 2019.

[101] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4580–4584, IEEE, 2015.