# USC-SIPI Report #450

# Behavior Understanding from Speech under Constrained Conditions: Exploring Sparse Networks, Transfer and Unsupervised Learning

By
Haoqi Li

**December 2020**

Signal and Image Processing Institute
**UNIVERSITY OF SOUTHERN CALIFORNIA**
USC Viterbi School of Engineering
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Suite 400
Los Angeles, CA 90089-2564 U.S.A.

# Behavior Understanding from Speech under Constrained Conditions: Exploring Sparse Networks, Transfer and Unsupervised Learning

by

Haoqi Li

---

A Dissertation Presented to the
FACULTY OF THE GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
DOCTOR OF PHILOSOPHY
(Electrical Engineering)

December 2020

This dissertation is dedicated to my parents Yurong Zhang and Runlin Li.

# Acknowledgements

First and foremost, I want to express my deep gratitude to my advisors Prof. Shrikanth (Shri) Narayanan and Prof. Panayiotis (Panos) Georgiou for their help and support during my PhD study. Prof. Shri's vision, insightful discussion and general openness have led me to freely explore my interested research areas and gain the courage of out-of-the-box explorations in research. Prof. Panos's responsibility, patience and selfless support help me to shape my research aptitude. Moreover, their enthusiasm and genuine personality always make me feel supported, and enable me to have self-confidence and optimism during my hard time.

I would like to thank Prof. Morteza Dehghani, Prof. C.-C. Jay Kuo and Prof. Jonathan Gratch for being my dissertation and qualifying committee members. I also would like to thank Prof. Brian Baucom for his help and suggestions from psychological science. All their constructive comments and insightful feedback contributed to the work in this thesis.

I also want to thank my labmates in SCUBA and SAIL, my TA partners and my friends at USC, particularly, Sandeep Nallan Chakravarthula, Shao-Yen Tseng, Prashanth Gurunath Shivakumar, Arindam Jati, Taejin Park, Raghuveer Peri, Md Nasir, Rimita Lahiri, Bin Wang and Ching-Han Lee, for an enjoyable collaboration experience, their friendship and support. Especially for those who worked together with me in EEB Room B16, I really cannot forget those days when we worked in the basement: fixing AC's thermostat, solving servers' noise issues, assembling furniture etc. All those experiences make us real engineers.

I also want to thank Amazon Inc. JD.com Inc and Sony PlayStation Inc. for providing my great internship opportunities. I thank Naveen Kumar, Yelin Kim, Ming Tu, Ruxin Chen, Jing

Huang, and Cheng-Hao Kuo for the collaboration and help during my internship. These experiences greatly inspire the scope of my research and it is my great honor to work with these excellent researchers from the industry.

Finally, I would like to thank my parents and my family for their selfless love. Their constant support and encouragement give me the momentum and courage to follow my heart, believe my decisions and achieve my dreams in each step of my life. I express my utmost gratitude to my parents who have devoted their unfailing love and care to me.

# Table of Contents

**Dedication** ii

**Acknowledgements** iii

**List of Tables** viii

**List of Figures** x

**Abstract** xii

**Chapter 1: Introduction** 1
- 1.1 Motivation . . . 1
- 1.2 Challenges and Constrained Conditions . . . 3
- 1.3 Dissertation Overview . . . 5

**Chapter 2: Sparsely Connected and Disjointly Trained Deep Neural Networks for Behavior Classification** 7
- 2.1 Introduction . . . 7
- 2.2 Methodology . . . 9
  - 2.2.1 DNN Training . . . 9
  - 2.2.2 Reduced Feature Dimensionality DNN . . . 10
  - 2.2.3 Sparsely-Connected and Disjointly-Trained DNN . . . 11
  - 2.2.4 Joint Optimization of Sparsely-Connected DNN . . . 12
  - 2.2.5 Local to Session mappings . . . 12
- 2.3 Couple Therapy Corpus . . . 13
- 2.4 Acoustic Feature Extraction . . . 13
  - 2.4.1 Audio Preprocessing . . . 13
  - 2.4.2 Acoustic Feature Extraction . . . 13
- 2.5 Experiment Setup . . . 14
- 2.6 Experiment Results and Discussion . . . 15
- 2.7 Conclusion . . . 18

**Chapter 3: Linking Emotions to Behaviors through Deep Transfer Learning** 20
- 3.1 Introduction . . . 20
- 3.2 Background . . . 23

3.3 Related Work . . . 25
3.4 Proposed Work: Behavioral Primitives . . . 27
3.4.1 Emotion Recognition . . . 28
3.4.2 Behavior Recognition through Emotion-based Behavior Primitives . . . 30
3.5 Datasets . . . 35
3.5.1 Emotion Dataset: CMU-MOSEI Dataset . . . 35
3.5.2 Behavior Dataset: Couples Therapy Corpus . . . 35
3.6 Audio Processing and Feature Extraction . . . 36
3.6.1 Behavioral Dataset Pre-processing . . . 36
3.6.2 Feature Extraction . . . 37
3.7 Experiments and Results Discussion . . . 37
3.7.1 General Settings . . . 37
3.7.2 ER and EC for Emotion Recognition . . . 38
3.7.3 Context-dependent Behavior Recognition . . . 40
3.7.4 Reduced Context-dependent Behavior Recognition . . . 44
3.7.5 Analysis on Behavior Prediction Uncertainty Reduction . . . 46
3.8 Conclusion and Future Work . . . 49

**Chapter 4: Speaker-invariant Affective Representation Learning via Adversarial Training 51**
4.1 Introduction . . . 51
4.2 Related Work . . . 52
4.3 Methodology . . . 53
4.3.1 Model Structure . . . 53
4.3.2 Difference with Prior Work . . . 54
4.3.3 Emotion Representation Adversarial Training . . . 55
4.3.3.1 Training of SC . . . 55
4.3.3.2 Training of ENC and EC . . . 56
4.4 Dateset . . . 57
4.5 Experiment Setup . . . 58
4.5.1 Model Configurations . . . 58
4.6 Results and Discussion . . . 59
4.6.1 Evaluation on IEMOCAP . . . 59
4.7 Conclusion . . . 62

**Chapter 5: Unsupervised Speech Representation Learning for Behavior Modeling using Triplet Enhanced Contextualized Networks 63**
5.1 Introduction . . . 63
5.2 Related Work and Motivation . . . 65
5.3 Unsupervised Speech Representation Learning for Human Behavior Modeling . . . 67
5.3.1 Behavioral Stationarity Assumption . . . 67
5.3.2 Deep Contextualized Network . . . 68
5.3.3 Triplet Enhanced Deep Contextualized Network . . . 68
5.4 Datasets . . . 72
5.4.1 Evaluation Datasets . . . 73

5.5 Experimental Setup . . . 74
5.5.1 Audio Data Preparation . . . 74
5.5.2 Feature Extraction . . . 75
5.5.3 Model Configurations and Parameter Settings . . . 76
5.5.4 Evaluation Method . . . 76
5.5.4.1 Evaluation Method for In-domain Couples Therapy Corpus . . . 76
5.5.4.2 Evaluation Method for Diverse Speech Behavior Corpora . . . 77
5.6 Experimental Results and Discussions . . . 77
5.6.1 Experiment Results of Couple Therapy Corpus . . . 77
5.6.2 Behavioral Trajectory Analysis . . . 81
5.6.3 Experiment Results on Diverse Speech Behavior Corpora . . . 82
5.6.4 Nuisance Factors and Selection of Features . . . 84
5.7 Conclusion . . . 84

**Chapter 6: Conclusion and Future Work** **86**
6.1 Summary of Research . . . 86
6.2 Future Work . . . 87

**Bibliography** **89**

**Appendices** **102**
A Detailed Network Architecture and Training Parameters of Chapter 3 . . . 103

# List of Tables


2.1 Classification accuracy (%) for the two different feature splits: One random instantiation and one knowledge based . . . . . 16

2.2 Classification accuracy (%) with all behavioral recognition systems . . . . . 17

3.1 Description of behaviors . . . . . 38

3.2 Weighted classification accuracy (WA) in percentage for emotion recognition on the CMU-MOSEI dataset. Bold numbers represent the best performing system. . . 39

3.3 Behavior binary classification accuracy in percentage for context-dependent behavior recognition model from emotion labels . . . . . 41

3.4 Behavior binary classification accuracy in percentage for context-dependent behavior recognition model from emotion-embeddings. Bold numbers represent the best performing system. . . . . 43

3.5 Behavior binary classification accuracy in percentage for reduced context-dependent behavior recognition from emotion-informed embeddings. Bold numbers represent the best performing system. . . . . 45

4.1 Model structure and training configuration details . . . . . 59

4.2 Classification accuracy (%) comparison on IEMOCAP . . . . . 59

4.3 Emotion classification accuracy (%) on CMU-MOSEI . . . . . 62

5.1 Description of behavior codes in Couples Therapy Corpus . . . . . 73

5.2 Description of evaluation data in Diverse Speech Behavior corpora . . . . . 74

5.3 Classification accuracy (%) of behavior codes in Couple Therapy Corpus . . . . . 78

5.4 Original annotation ratings and binarized classification labels for each behavior code 81

6.1 Network architecture of ER . . . . . 103

6.2 Network architecture of EC . . . . . 103

6.3 B-BP based context-dependent behavior recognition model framework . . . . . 104

6.4 E-BP based context-dependent behavior recognition model framework . . . . . 105

6.5 E-BP based reduced context-dependent behavior recognition model framework. Those AvgPool1d layers are optional to adjust temporal receptive field size. . . . . 106

# List of Figures


2.1 During training, local reference is assumed to be equal to global as denoted by the green row of $\rho$. During testing the mean rating is assigned as the estimated session-level rating $\rho'$. . . . . 10

2.2 Basic behavior recognition system based on sub feature set . . . . . 11

2.3 Sparsely-Connected and Disjointly-Trained DNN . . . . . 11

2.4 Output of SD-DNN for one sample test session with three behavior codes . . . . . 18

3.1 Illustration of task complexity or age of acquisition for machines and humans. . . . 24

3.2 Models of ER, EC and two kinds of BPs. L is the input feature length. . . . . 29

3.3 B-BP based context-dependent behavior recognition model . . . . . 31

3.4 E-BP based context-dependent behavior recognition model. E-BP_$l$ is the output from $l^{\text{th}}$ pretrained CNN layer. In practice multiple E-BP_$l$ can be employed at the same time through concatenation. In this work we only employ the output of a single layer at a time. . . . . 32

3.5 Illustration of local context awareness and global context reduction. In previous sections, the E-BPs (and B-BPs) are passed to a GRU that preserves their sequences. Here they are processed through pooling and context is removed. . . . . 33

3.6 E-BP reduced context-dependent behavior recognition model. Model (A) has a smaller receptive field while model (B) has a larger receptive field because of the added local average pooling layers. . . . . 34

3.7 Sessions with similar percentage of emotions presence but different behavior label . 41

3.8 PUR optimal value of E-BP based context-dependent and reduced context-dependent models across behaviors . . . . . 48

4.1 Model structure with loss propagating flow . . . . . 54

4.2 t-SNE plot of emotion embedding with both 4 emotion labels (left) and 2 speaker labels (right) for multi-task model and our proposed MEnAN model. . . . . 61

5.1 Behavior representation learning framework via the DCN model. During training, the model encodes neighboring frames with a DCN. The choice of features promotes the behavior information as the extracted common information in the behavior manifold. During evaluation, similarity comparison is made by calculating distance within the behavior manifold. . . . . 69

5.2 Behavior representation learning framework via the TE-DCN model. The model has a triplet structure with shared weights. Audio 1 and 2 can be two temporally-distant regions or two different files. During training, the model is optimized to minimize both reconstruction loss and triplet loss. During evaluation, representation similarity comparison is performed within the behavior manifold. . . . . 70

5.3 One sample session with five behavior score trajectories . . . . . 82

5.4 Confusion matrix of behavior scenario similarity evaluation . . . . . 83

# Abstract

The expression and perception of human behavioral signals play an important role in human interactions and social relationships. However, the computational study of human behavior from speech remains a challenging task since it is difficult to find generalizable and representative features because of noisy and high-dimensional data, especially when data is limited and annotated coarsely and subjectively. This dissertation focuses on the computational study of human behaviors via deep learning techniques.

Deep Neural Networks (DNN) have shown promise in a wide range of machine learning tasks, but for Behavioral Signal Processing (BSP) tasks, their application has been constrained due to limited quantity of data. In the first part of this dissertation, we propose a Sparsely-Connected and Disjointly-Trained DNN (SD-DNN) framework to deal with limited data. First, we break the acoustic feature set into subsets and train multiple distinct classifiers. Then, the hidden layers of these classifiers become parts of a deeper network that integrates all feature streams. The overall system allows for full connectivity while limiting the number of parameters trained at any time and allows convergence possible with even limited data. The results demonstrate the benefits in behavior classification accuracy.

An important cue of behavior analysis is the dynamical changes of emotions during the conversation. In the second part of this dissertation, we employ deep transfer learning to analyze inferential capacity and contextual importance between emotions and behaviors. We first train a network to quantify emotions from acoustic signals and then use information from the emotion recognition network as features for behavior recognition. We treat this emotion-related information as behavioral primitives and further train higher level layers towards behavior quantification.

Through our analysis, we find that emotion-related information is an important cue for behavior recognition. Further, we investigate the importance of emotional-context in the expression of behavior by constraining (or not) the neural networks' contextual view of the data. This demonstrates that the sequence of emotions is critical in behavior expression.

The results suggest that it is feasible to use emotion-related speech representation for behavior quantification and understanding. However, the representation learning for speech emotion recognition is also challenging. There is much variability from input speech signals, human subjective perception of the signals and emotion label ambiguity. In the third part of this dissertation, we propose a machine learning framework to obtain speech emotion representations by limiting the effect of speaker variability in the speech signals. Specifically we propose to disentangle the speaker characteristics from emotion through an adversarial training network in order to better represent emotion. Our method combines the gradient reversal technique with an entropy loss function to remove such speaker information. We show that our method improves speech emotion classification and increases generalization to unseen speakers.

Though we use a range of techniques for dealing with limited resources, domain specific data and entanglement of information, supervised behavioral modeling mostly relies on domain-specific construct definitions and corresponding manually-annotated data, rendering generalizing across domains challenging. In the last part of this dissertation, we exploit the stationary properties of human behavior within interactions and present a representation learning method to capture behavioral information from speech in an unsupervised way. We hypothesize that nearby segments of speech share the same behavioral context and hence map onto similar underlying behavioral representations. We present an encoder-decoder based Deep Contextualized Network (DCN) as well as a Triplet-Enhanced DCN (TE-DCN) framework to capture the behavioral context and derive a manifold representation, where speech frames with similar behaviors are closer while frames of different behaviors maintain larger distances. The models are trained on movie audio data and validated on diverse domains including on a couples therapy corpus and other publicly collected

data (e.g. stand-up comedy). With encouraging results, our proposed framework also shows the feasibility of unsupervised learning within cross-domain behavioral modeling.

# Chapter 1

# Introduction

## 1.1 Motivation

Human communications include a complex dynamic interplay of rich verbal and nonverbal behavior signals, including prosodic cues, spoken language and body language *etc.*. The expression and perception of human behavioral signals play an important role in human interactions and social relationships. These behavioral signals are both causes and consequences of different manifestations from speakers' underlying communicative intents and purposes, emotional states and stances. They not only shape the structure of an interpersonal interaction but also elicit more natural and intellectual human communications. The study of human behaviors is central to the understanding of human interactions and is essential for the advancement of computational affective models.

Signals of expressive human behaviors have also been used in the assessment of mental health and well being. Analysis and classification of human behaviors exhibited in patient interactions have gradually become one of the core tasks among a variety of clinical domains in psychotherapy, in which certain behaviors have been widely used as indicative cues of mental health in observational studies [105]. Observation, evaluation, and identification of domain-specific human behaviors are essential for psychologists to provide effective and specific treatment to patients. In the traditional behavior annotation process, manual behavior rating is a costly and time-consuming process, and the annotated behavior codes often suffer from the variability of the subjective biases

of the annotation experts. Having an automatic behavior quantification framework can greatly enhance the objective assessments of patients, the quality of treatment, and more importantly, it can inform and help the domain experts.

Given the importance of understanding human behaviors, the computational study of human behaviors has attracted interest from a wide variety of disciplines, including psychology, social science, health care, and engineering. Automatic computational approaches that support measurement, analysis, and modeling of human behavior and interactions have been investigated over the last few years. The emerging research field, known as Behavioral Signal Processing (BSP) [105], integrates domain-specific knowledge, machine learning and signal processing methods, and employs acoustic[14, 157], lexical[28, 50], and visual information[97, 158] to model and analyze multimodal human behaviors. Great advances have been made in assessing human state through the technical way in areas such as couples therapy [14, 108], depression [59, 106, 142] and suicide risk assessment [31, 39, 107, 109, 151]. Though promising advances have been achieved, much of the existing research focuses on a standard two-step learning process: Designed acoustic features or lexical features are firstly extracted and then followed by a simple classification process, with models such as Support Vector Machines (SVM) [62] and Hidden Markov Models (HMM) [115]. However, the real human behavioral understanding process is highly complex and nonlinear. It is challenging to design one specific algorithm to simulate complex human annotation processes. Those knowledge-based features or pre-defined machine learning frameworks often limit the learning of underlying representation of complex human behaviors, and many intrinsic behavior properties cannot be derived from data through those learning processes.

Recently, deep learning techniques [58] have been employed and shown promising results in many related areas related to human affective computing such as speech emotion recognition and sentiment analysis [24]. The success of deep learning can be attributed to two main properties: first, the Deep Neural Networks (DNN) can approximate any continuous smoother functions [160], and second, it can learn functions based on large amounts of data. This data-driven learning framework might alleviate the difficulty of the design of effective behavior representations and specific

algorithms to simulate the complex human behavior annotation process. The underlying representations that the DNN identifies might advance the research in the BSP domain.

This dissertation aims to advance the computational human behaviors quantification and understanding, focusing primarily on addressing problems arising from the utilization of large-scale data-driven deep learning techniques under certain constraints in the BSP domain. Our goals are to explore the possibilities of employing deep learning modules in the BSP domain applications, to develop computational methodologies of human behavior modeling, to further facilitate the understanding of different human affect related behaviors from data itself and finally to facilitate the human experts' assessment in the psychotherapy.

## 1.2 Challenges and Constrained Conditions

Automated affect-related human behavior annotation is a complex task. Even for trained human annotators, rating behaviors is a costly and time-consuming process, and the subjective biases in annotated ratings of behavior codes are inevitable. This subjectivity is mainly induced by the complexity property of human communication and interaction itself. For example, human affect can be easily shaped by many factors, such as context, linguistic content and acoustic spectral and prosodic information.

In addition, human behaviors are often manifested over long time periods, they are complex, domain-specific and multimodal. In real scenarios, a range of behavioral cues, from explicit and overt linguistic constructs to implicit paralanguage and nonverbal communication, are encoded in multiple resolutions, time scales, and with different levels of complexity. All these factors make human behavioral integration a highly complex and non-linear process. Thus, it is challenging to design algorithms to simulate such a human annotation process. Regardless of the complexity of human behaviors, compared with other machine learning domains, there are also specific constrained conditions and challenges in the field of BSP.

1. **Limited data resources**

   In BSP, rich behavior signals are usually collected in real scenarios from real patients. One of the biggest challenges is that representative samples of behavior are extremely limited due to the cost of data collection and annotation. Moreover, this data sparsity issue is exacerbated by the difficulty of obtaining data due to privacy constraints [93, 105]. In addition, the large imbalanced data in different behavior codes also exacerbates the issue of limited resources.

2. **Low-resolution low agreement labels**

   When annotators manually annotate interaction sessions, they usually require training in order to give ratings in a consistent manner, unbiased by personal influences. However, because of the inherent complexity of human behavior, disagreement in human annotations is still inevitable [95]. Thus, challenges with subjectivity, low inter-annotator agreement and coarse annotations (both attributed to cost and human contextualization of short-term information) especially in micro and macro annotation, complicate the learning task.

3. **Errors from pre-processing procedures**

   The data collected in real scenarios usually requires a lot of pre-data processing steps before they are ready to use. The typical processing procedures for audio recordings can include denoising for low signal to noise ratio (SNR) sessions, voice activity detection (VAD) for selection of human speech regions, speaker diarization [103] for segmentation of speech regions belonging to each speaker and automated speech recognition for the obtainment of speech transcripts. Errors from all these processing steps can be accumulated and potentially further impact the machine learning performance on automated behavior recognition.

In this dissertation, we mainly take the first two challenges described above as constrained conditions to explore direct and indirect approaches on building computational human behavior modeling from speech signals via deep learning techniques and employ the out-of-domain data to facilitate the domain behavior training.

## 1.3 Dissertation Overview

This proposed work mainly focuses on building a computational behavior quantification system on acoustic modality from speech signals, which is unobtrusively obtainable and known to offer rich behavioral information [108]. Specifically, we mainly utilize the Distressed Couple Therapy Behavior [34] quantification as a case study to probe into the computational approaches for modeling human behaviors under constrained conditions. To facilitate the automated domain behavior recognition, a range of techniques are employed to deal with limited domain-specific data and the entanglement information from speech signals. We show that our proposed models not only improve existing automated human behavior quantification systems, but more importantly, provide an insightful understanding of behaviors from real data.

Here are the main statement and contribution of this dissertation:
Even in a low-resource scenario, with an appropriate designed model structure or the utilization of related out-of-domain resources, data-driven deep learning techniques could demonstrate the outstanding capability of representation formulation in behavior understanding and quantification.

The overview of this dissertation can be briefly summarized as follows: First, we explore the feasibility of directly using feed-forward neural networks on knowledge-based affect related acoustic features. However, due to the data sparsity and variability property of human behaviors, the fully connected DNN system cannot converge well. Even though we reduce the number of neural networks' parameters, add the dropout [137] to prevent overfitting, it is still overfitting to the training data rather than learning behavior related information. Inspired by the dropout technique, we propose a neural network framework which allows for full connectivity while limiting the number of parameters trained at any time and allows for the probability of convergence even with limited data (see Chapter 2).

Then, we study the indirect approaches towards deep behavior analysis and quantification. Since emotions are psychologically known to be linked to behaviors, we link basic emotions to complex behaviors through deep transfer learning, in which we employ short-term emotions as primitives towards complex behavior understanding and investigate the inferential capacity and

contextual importance between emotions and behaviors (see Chapter 3). Once we verify the effectiveness of using emotion related representation for behavior quantification, in Chapter 4, we further focus on the robust affect representation learning from speech. Specifically, we regard the speaker information as one type of domain knowledge, and propose a max-entropy adversarial network to obtain speaker-invariant affective representation. In Chapter 5, the slow varying properties of human behaviors are exploited, and we propose a triplet enhanced contextual encoder-decoder structure to connect context and derive the behavioral manifold in an unsupervised manner.

Finally, in Chapter 6, we summarize this dissertation and provide several potential research directions for future study.

# Chapter 2

# Sparsely Connected and Disjointly Trained Deep Neural Networks for Behavior Classification

## 2.1 Introduction

Observational practice, such as in the field of psychology, relies heavily on analysis of human behaviors based on observable interaction cues. In Couples' Therapy, one fundamental task is to observe, evaluate and identify domain-specific behaviors during couples' interactions. Based on behavioral analyses, psychologists can provide effective and specific treatment.

Rating behaviors by human annotators is a costly and time consuming process. Great advances have been made during last decade on assessing human state through technical way. For example, speech emotion recognition works [45, 127, 153] have shown effectiveness of extracting emotional content from human speech signals. In addition, Deep Neural Networks (DNN) have been employed for many related speech tasks [68, 123, 139]. Han *et al.*[60] and Le *et al.*[82] both utilized DNN to extract high level representative features to improve emotion classification accuracy.

Human emotions can change quickly and frequently in a short time period, thus emotion recognition mainly focuses on very short speech segments (*e.g.,* less than 2s) [104]. Affect recognition models basic emotions and is not domain-specific. For mental health applications, though, experts are more interested in very specific and complex behaviors exhibited over longer time scales. Over

the last few years Behavioral Signal Processing (BSP) [51, 105] has examined the analysis of such complex, domain specific behaviors. Based on machine learning techniques, BSP employed lexical [50], acoustic [14], and visual [97, 158] information to analyze and model multimodal human behaviors. For instance, in couples' therapy domain, Black *et al.*[14] built an automatic human behavioral coding system for couples' interaction by using acoustic features. In [28, 157] the authors employed a top layer HMM to take dynamic behavior state transitions into consideration and thus achieved higher accuracy on session-level behavioral classification.

Despite these efforts, behavior estimation is still a complex task. Session level models combine information at different timescales to estimate a session level rating. In doing so, they ignore non-linear information integration models which are often employed by human raters, such as recency and primacy models. For one thing, how human interpret and integrate behavior information is still not well understood since the process is not a simple linear system; Further, and one of the biggest challenges, is that representative samples of behavior are extremely limited due to privacy constraints, cost of annotation, subjective ground truth, and coarse annotations (both attributed to cost and human contextualization of short-term information).

Deep Neural Networks have shown promise in a wide range of machine learning tasks, especially for their ability to extract high level descriptions from raw data. However, in BSP, due to the limited quantity of data, DNN deployment is difficult. Because of limited data, high-dimensionality acoustic features, high signal variability, and the complication that the same acoustic signal encodes a range of additional information, training DNN systems on such data fails to converge to optimal operating conditions.

To address this problem, we propose a Sparsely-Connected and Disjointly-Trained Deep Neural Networks (SD-DNN) and demonstrate its use for behavioral recognition in Couples' Therapy.

The rest of chapter is organized as follows: Section 2.2 describes the proposed SD-DNN behavior learning system in detail. Section 2.3 provides a brief description of the database used in

experiments. Section 2.4 describes audio pre-processing steps and feature extraction methods employed in our work, after which we design multiple experiments and discuss our results in Section 2.5 and 2.6. Finally, we present our conclusions in Section 2.7.

## 2.2 Methodology

Human experts integrate a range of cues over a wide time interval and significant context to arrive at session-level behavior descriptors. For example, a therapist can observe a couple interacting for an hour and derive an assessment that one of the partners is negative while the other shows acceptance. This, unfortunately, means that we are often left without an instantaneous ground-truth. More often than not, this results in either building session level systems by employing all available data *e.g.,* [14], averaging of local decisions towards session level ratings [50], or creating models of interaction as in [28, 157].

In this work, we will build a system that is able to estimate behaviors over short time frames towards implementing a live behavioral estimation framework. We propose a Sparsely-Connected and Disjointly-Trained Deep Neural Network (SD-DNN), that aims to tackle the data sparsity issues in behavioral analysis.

Due to the lack of ground truth at short time intervals, we will employ session level ratings for training and evaluation. For training, we will assume that every frame in a session shares the same rating as the session level gestalt rating as shown on Fig. 2.1. For evaluation, we will use the average of the macro-coding to estimate session level coding. Finally, we will demonstrate how the system is able to track behavioral trajectories.

### 2.2.1 DNN Training

Employing the usual way of training a DNN system requires significant amounts of data. In our analysis, and with a feature size of 168, this approach always lead to failure during training: DNN

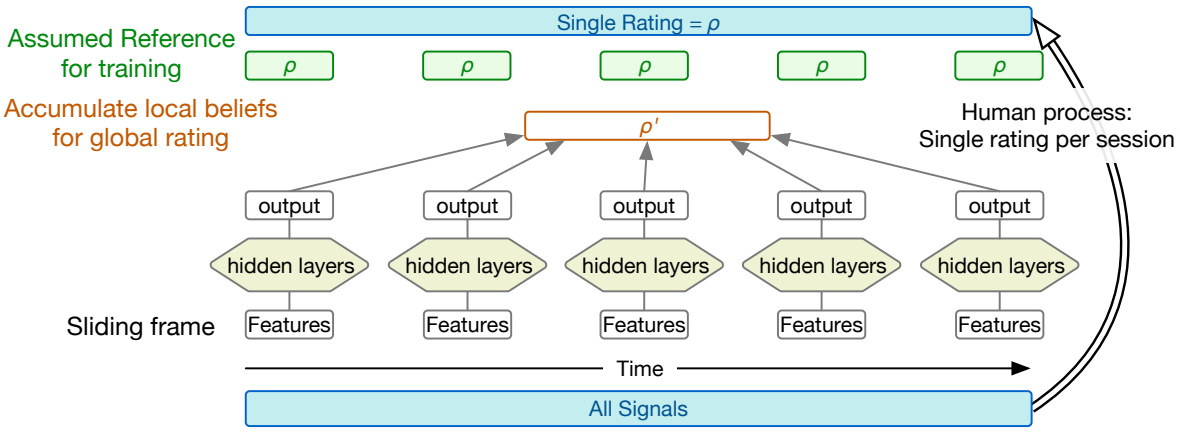



Figure 2.1: During training, local reference is assumed to be equal to global as denoted by the green row of $\rho$. During testing the mean rating is assigned as the estimated session-level rating $\rho'$.

training immediately identifies a local minimum even for small neural networks; while the objective function decreases on the training set, it does not on the development set. Behavioral recognition results during testing are mostly unchanging, and hence uninformative in providing behavioral trajectories. Likely the system converges to different minima relating to other dimensions, such as for instance speaker characteristics.

To minimize overfitting we can add a dropout layer[137] at the input. This feature reduction avoids overfitting to a certain degree, however we still do not obtain the gains we expected from employing a DNN framework.

### 2.2.2 Reduced Feature Dimensionality DNN

One way to avoid overfitting issues is to use a reduced dimensionality input feature set. We can do that through selecting a subset of features and training DNN on those, which means we use these sub-feature-sets to train multiple behavior recognition systems. For each of these systems, the feature dimension is reduced by a significant factor compared to the full feature set, thus number of parameters in the resulting DNN is also decreased. Using same amount of training data, we can obtain a robustly trained DNN. The process of this stage is shown in Figure 2.2.

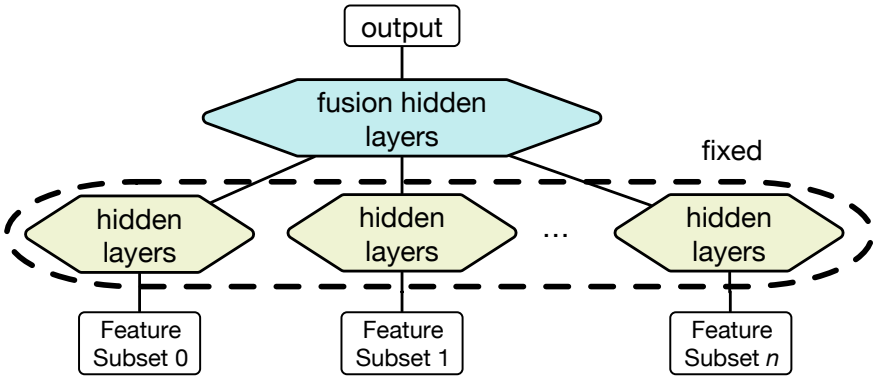



Figure 2.3: Sparsely-Connected and Disjointly-Trained DNN

### 2.2.3 Sparsely-Connected and Disjointly-Trained DNN

To gain both the advantages of small feature sets, which converge to avoid overfitting issues, and to still exploit the redundancy among feature streams, we propose the Sparsely-Connected and Disjointly-Trained DNN (SD-DNN) training framework. In this framework, depicted in Fig. 2.3, we select a sparse feature set, train (as in the Reduced feature dimensionality DNN's) individual DNN systems. Then we fix the parameters of these DNN systems, remove the output layer, connect the top hidden layers together, and add new hidden layers as fusion layers. This framework allows for both *Sparse Connectivity* at the bottom layers (not all features are connected to all hidden layers

above) and *Disjointly Training* the various layers of the DNN thus reducing the degrees of freedom and achieving convergence.

### 2.2.4 Joint Optimization of Sparsely-Connected DNN

The system presented in the previous section and shown in Fig. 2.3 disjointly optimizes the sparse lower layers and top fusion layers. Without increasing the parameter dimensionality of the SD-DNN, we can initialize training from the disjoint optimization point and jointly optimize the system. We will denote this *Sparse, Jointly* optimized system by SJ-DNN.

### 2.2.5 Local to Session mappings

As mentioned earlier, we have only session-level ratings for the couple therapy corpus. This is not unusual in mental health applications given the cost and subjectivity of annotations.

Due to subjectivity and inter-annotator agreement issues we use a binarized subset of the dataset that lies at the top and bottom 20% of the dataset as in [14] for our training. We assign score 1 for high presence and 0 for low presence of one certain behavior. Frame-level training samples are given the same reference as the session level reference as shown on Figure 2.1.

At test time, the output of the DNN system provides a score of the presence of behavior (as in Figure 2.4), but doesn't provide a global rating. While a range of methods exist for fusing decisions (*e.g.,* [28, 83, 157]), in this work we will use the simplest one: Average posteriors. We can treat the output of DNN, $q_i^k$ as a proxy to the posterior probability of the behavior given the frame $i$ for session $k$, and $L_k$ is the number of frames in session $k$. We then average $q_i^k$ to derive the session level confidence score $Q_k$. Mathematically:

$$Q_k = \exp(\frac{1}{L_k}\sum_i \log q_i^k) \tag{2.1}$$

For comparison with the reference session level label, we threshold and binarize $Q_k$. The threshold, $T_k$, is selected by optimization to give the minimum classification error rate on the training data.

## 2.3 Couple Therapy Corpus

The database used in this chapter is provided by UCLA/UW Couple Therapy Research Project [34], in which 134 couples participated in video-taped problem-solving interactions. During each discussion, a relationship-related topic (e.g."why can't you leave my stuff alone?") was selected. Each participant's behaviors was rated separately by human annotators for a set of 33 behavioral codes (e.g. "Blame", "Acceptance" *etc.*) based on the Couples Interaction Rating System (CIRS) [63] and the Social Support Interaction Rating System (SSIRS) [77]. Every human annotator provided a subjective rating scale from 1 to 9, where 1 refers absence of the behavior and 9 indicates a strong presence. For more information about this dataset, please refer to [14, 34].

## 2.4 Acoustic Feature Extraction

### 2.4.1 Audio Preprocessing

In any acoustic behavior classification task, we first need to identify contiguous regions of speech by the interlocutors. This requires a range of pre-processing steps: *Voice Activity Detection* (VAD) to identify spoken regions, *Speaker Diarization* to identify same-speaker regions. Following this, we perform the feature extraction from speech regions. In our work we employ the preprocessing steps described in [14]. In short: We employ all available interactions with a SNR above 5dB, and perform VAD and Diarization. Then we ignore speech segments that are shorter than 1.5 seconds. Speech segments from each session for the same speaker are then used to analyze behaviors.

### 2.4.2 Acoustic Feature Extraction

We extract acoustic features characterizing speech prosody (pitch and intensity), spectral envelope characteristics (MFCCs, MFBs), and voice quality (jitter and shimmer). All these Low-Level-Descriptors (LLDs) are extracted every 10 *ms* with a 25 *ms* Hamming window through *openSMILE*[47] and *PRAAT*[15]. We perform session level feature normalization for each of the speakers as in [14] to reduce the impact of recording conditions and physical characteristics of different speakers.

Unlike [14] we are interested in building a fine-resolution behavioral estimation, rather than session-level classification-only system, and as such we employ features with a sliding frame[1]. Within each frame, we calculate a number of functionals: Min (1st percentile), Max (99th percentile), Range (99th percentile – 1st percentile), Mean, Median, and Standard Deviation.

## 2.5 Experiment Setup

We use leave-one-couple-out cross-validation to separate training and test data. We can thus ensure a fair evaluation where same couple is not seen in the test set. For each behavior code and each gender we use 70 sessions on one extreme of the code (*e.g.,* high blame) and 70 sessions at the other extreme (*e.g.,* low blame)[2]. This is to achieve higher inter-annotator agreement and provide training data with binary class labels.

Temporal variation in behavior is slower than basic emotions' and thus a longer frame window size of speech segment is needed for its analysis. An earlier work[157] compared behavior classification performance on various frame sizes and showed that a 20 s frame was sufficient to estimate meaningful behavioral metrics while maintaining high resolution, we thus choose to use a 20 s window with 1 s shift.

In our experiments we employ 3 of behavioral codes available to us: *Acceptance, Negativity, Blame*. We evaluate using a baseline SVM system and compare with the above proposed DNN based systems.

*In summary:* We use 168 features as discussed in section 2.4.2; classify 3 behavioral codes: *Acceptance, Negativity, Blame*; train a 1s-slide, 20s-length rating system; accumulate beliefs towards binary classification evaluation; and qualitatively evaluate the behavioral trajectories resulting from the proposed system.

[1]Note: arguably this could be converted into an online system if the normalization was done with a slower-varying sliding window, akin to the CMV normalization of ASR systems.

[2]These do not necessarily correspond to matched partners due to the selection of the extreme sessions.

## 2.6 Experiment Results and Discussion

**Baseline SVM:**

The baseline SVM model was built similar to the Static Behavioral Model discussed in [157].

**Fully Connected DNN:**

The fully connected DNN system described in section 2.2.1 did not converge and always kept the first epoch values as the final states. To reduce this issue we had to introduce significant dropout at the input layer. We also had to keep the overall network very small with only one hidden layer of 15 units. We used a mini-batch adaptive gradient optimizer with a mean square error objective function. As seen from Table 2.2, the fully connected DNN gains were modest.

**Reduced dimensionality DNN:**

To create smaller DNNs that may converge easier, we divided features into 5 parts: (a) knowledge-based split by feature type: pitch, MFCCs, MFBs, jitter and shimmer, intensity. (b) Randomly. Then for each feature subset we train a DNN with the same configuration as in the fully connected DNN, *i.e.,* one hidden layer with 15 units.

With these reduced and shallow neural nets we immediately observe good training characteristics and convergence. Further from the results of Table 2.1 we can observe that even the reduced feature size can often outperform the baseline SVM, which suggests potential gains from employing DNNs for behavior recognition. We also note that even the random split can perform quite well in fusion compared to the baseline. Due to the randomness in this feature selection, different splits may even be able to improve, however due to the lack of a development set we decided not to perform such an optimization. The knowledge-based feature selection has a less uniform classification accuracy due to the feature-size imbalance as expected, but we obtain better performance on SD-DNN fusion described next, so we use knowledge-based feature split in all following experiments.

| One random feature split instantiation | | | | | | | |
|---|---|---|---|---|---|---|---|
| SVM (Baseline) | Subset 0 | Subset 1 | Subset 2 | Subset 3 | Subset 4 | Fusion | SD-DNN |
| 68.57 | 70.36 | 72.85 | 72.14 | 67.50 | 67.50 | 70.00 | 75.00 |
| Knowledge-based feature split | | | | | | | |
| SVM (Baseline) | Pitch | MFCCs | MFBs | Intensity | Jitter & Shimmer | Fusion | SD-DNN |
| 68.57 | 66.07 | 71.07 | 66.78 | 61.43 | 61.79 | 72.14 | 75.36 |

Table 2.1: Classification accuracy (%) for the two different feature splits: One random instantiation and one knowledge based

**SD-DNN:**

We thus proceed to construct our SD-DNN system by fixing the parameters of the reduced dimensionality DNN systems and connecting their hidden layers ($15 \times 5$) to another layer of DNN. In our experiment, we utilize additional two hidden layers with 30 and 10 units respectively, and use the same optimizer and objective function as before. As we can see from the last column of Table2.1 the performance of the SD-DNN is significantly better than that of the fusion of the individual reduced dimensionality DNN's.

**SJ-DNN:**

To relax the disjoint optimization constraint we also train jointly reduced feature DNNs at the front layers and the top fusion DNNs of the above model. The parameter space of the model is identical to the SD-DNN except all parameters are initialized on SD-DNN values but jointly trained. Table 2.2 shows that despite the two models being identical, the joint optimization of a larger set of parameters reduces the performance of the SJ-DNN model versus the SD-DNN.

**Fully Connected DNN, SD-DNN Initialized ($\mathbf{DNN_{SD\text{-}init}}$):**

After achieving a better performing system, we attempt once again to reduce sparseness, and hence increase the parameter space of the model, by fully connecting all inputs/hidden layers. We employ

the SD-DNN model as initialization instead of using random initialization on DNN. This model is initialized with the weights of the SD-DNN, or zero if the connection did not exist before.

All results of experiments are shown in Table 2.2, in general, the SD-DNN system has higher accuracy rate than SVM baseline and plain DNN system. We obtain the greatest improvement for *Acceptance* behavior from 68.57% to 75.36%, which shows benefits in employing DNN and reducing connectivity of DNN because of sparse data.

| Behavior Code | SVM | Fully connected DNN | SD-DNN | SJ-DNN | $\text{DNN}_{\text{SD-init}}$ |
|---|---|---|---|---|---|
| Acceptance | 68.57 | 71.79 | **75.36** | 73.57 | 71.43 |
| Negativity | 73.21 | 74.64 | **77.14** | 75.36 | 74.29 |
| Blame | 73.21 | 73.93 | **75.71** | 74.29 | 73.93 |

Table 2.2: Classification accuracy (%) with all behavioral recognition systems

In summary we can observe that both reduction of the total number of parameters via sparseness but also reduction of the trainable parameters at any time via disjoint training can help in dealing with limited data. Specifically by observing the fully connected DNN and $\text{DNN}_{\text{SD-init}}$ results, for most behavioral codes, we can see that any increase in the system's number of parameters (reduction of sparseness) results in reduction of the performance, even if the initialization point is a good one. We can also see that increasing the number of simultaneously and jointly trainable parameters, as visible by comparing SD- and SJ-DNN's, also damages performance.

**Online Behavioral Trajectories:**

One of the advantages of moving to an estimation, rather than classification framework, is that we can now provide domain experts with behavioral trajectories. These are becoming increasingly necessary, especially in new behavioral analysis paradigms where patients are instrumented continuously in-lab, at-home, and *in-situ*. The resulting datasets are vast, even though training data is limited, and behavioral trajectories can help identify specific behaviors over time. One sample

behavior dynamic change trajectories is shown in Fig. 2.4. From this figure, we can see behavior *Negativity* and *Blame* are highly correlated, and have opposite trend with *Acceptance*, which is in agreement with our intuition and previous research work[14].

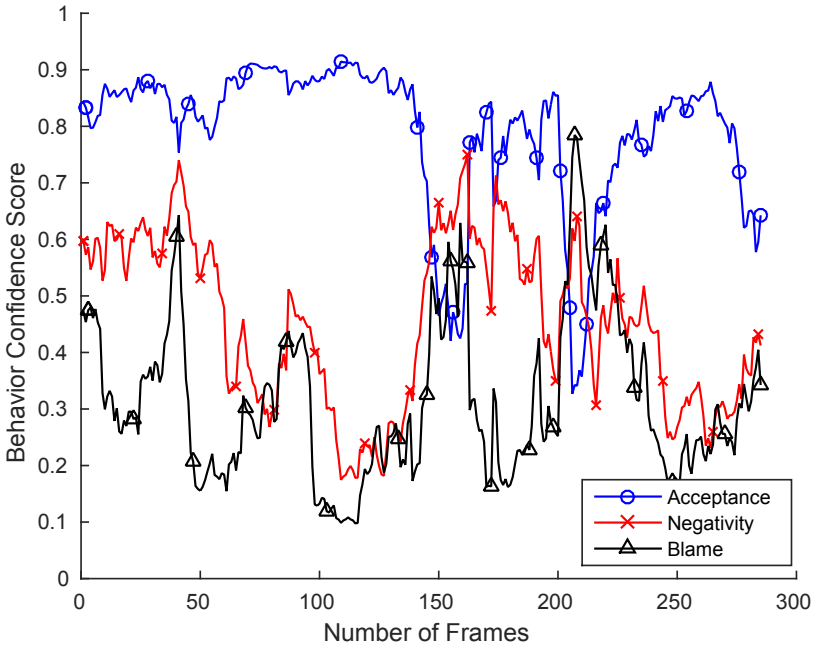



Figure 2.4: Output of SD-DNN for one sample test session with three behavior codes

Overall, results suggest that a Sparsely-Connected, Disjointly-Trained DNN framework provides the most promise in employing DNNs into the limited data BSP domain.

## 2.7 Conclusion

Compared to other DNN based machine learning tasks, data sparsity is a critical issue in BSP domain due to its costly and complicated data generating process. Through *Sparsely Connected* and *Disjoint Training* we can train more complex architecture DNN systems with limited dataset, achieve increased session-level performance, and importantly obtain continuous in time and rating annotations of our data. For future work, we can tune the SD-DNN architecture and parameters. For instance, different reduced dimensionality DNN learning system can use different DNN architecture.

At the same time, we notice that the low resource constrained condition limits the use of large-scale data-driven machine learning in BSP domain. Thus, in the next chapter, we plan explore the "indirect" approaches to employ mutual or shared information between different behavior codes. Also, we utilize the link between basic emotion and complex behaviors into behavioral analysis..

# Chapter 3

# Linking Emotions to Behaviors through Deep Transfer Learning

## 3.1 Introduction

Human communication includes a range of cues from lexical, acoustic and prosodic, turn taking and emotions to complex behaviors. Behaviors encode many domain-specific aspects of the internal user state, from highly complex interaction dynamics to expressed emotions. These are encoded at multiple resolutions, time scales, and with different levels of complexity. For example, a short speech signal or a single uttered word can convey basic emotions [43, 44]. More complex behaviors require domain specific knowledge and longer observation windows for recognition. This is especially true in task specific behaviors of interest in observational treatment for psychotherapy such as in couples' therapy [34] and suicide risk assessment [39]. Behaviors encompass a rich set of information that includes the dynamics of interlocutors and their emotional states, and can often be domain specific. The evaluation and identification of domain specific behaviors (e.g. blame, suicide ideation) can facilitate effective and specific treatments by psychologists. During the observational treatment, annotation of human behavior is a time consuming and complex task. Thus, there have been efforts on automatically recognizing human emotion and behavior states, which resulted in vibrant research topics such as affective computing [113, 118, 144], social signal processing [152], and behavioral signal processing (BSP) [50, 105]. In the task of speech emotion

recognition (SER), researchers are combining machine learning techniques to build reliable and accurate affect recognition systems [129]. In the BSP domain, through domain-specific focus on areas such as human communication, mental health and psychology, research targets advances of understanding of higher complexity constructs and helps psychologists to observe and evaluate domain-specific behaviors.

However, despite these efforts on automatic emotion and behavior recognition (see section3.3), there has been less work on examining the relationship between these two. In fact, many domain specific annotation manuals and instruments [63, 66, 77] have clear descriptions that state specific basic emotions can be indicators of certain behaviors. Such descriptions are also congruent with how humans process information. For example, when domain experts attempt to quantify complex behaviors, they often employ affective information within the context of the interaction at varying timescales to estimate behaviors of interest [105, 148].

Moreover, the relationship between behavior and emotion provides an opportunity for (i) transfer learning by employing emotion data, that is easier to obtain, annotate, and less subjective, as the initial modeling task; and (ii) employing emotional information as building blocks, or primitive features, that can describe behavior.

The purpose of this work is to explore the relationship between emotion and behavior through deep neural networks, and further the employ emotion-related information towards behavior quantification. There are many notions of what an "emotion" is. For the purpose of this paper and most research in the field [45, 129], the focus is on *basic emotions*, which are defined as cross-culturally recognizable. One commonly used discrete categorization is by Ekman [43, 44], in which six basic emotions are identified as anger, disgust, fear, happiness, sadness, and surprise. According to theories [119, 120], emotions are states of feeling that result in physical and psychological changes that influence our behaviors.

Behavior, on the other hand, encodes many more layers of complexity: the dynamics of the interlocutors, their perception, appraisal, and expression of emotion, their thinking and problem-solving intents, skills and creativity, the context and knowledge of interlocutors, and their abilities

towards emotion regulation [8, 9]. Behaviors are also domain dependent. In addiction [6], for example, a therapist will mostly be interested in the language which reflects changes of addictive habits. In suicide prevention [39], reasons for living and emotional bond are more relevant. In doctor-patient interactions, empathy or bedside manners are more applicable.

In this paper, we will first address the task of basic emotion recognition from speech. Thus we will discuss literature on the notion of emotion (see section 3.2) and prior work on emotion recognition (see section 3.3). We will then, as our first scientific contribution, describe a system that can label emotional speech (see section 3.4.1).

The focus of this paper however is to address the more complex task of behavior analysis. Given behavior is very related to the dynamics, perception, and expression of emotions [119], we believe a study is overdue in establishing the degree to which emotions can predict behavior. We will therefore introduce more analytically the notion of behavior (see section 3.2) and describe prior work in behavior recognition (see section 3.3), mainly from speech signals. The second task of this paper will be in establishing a model that can predict behaviors from basic emotions. We will investigate the emotion-to-behavior aspects in two ways: we will first assume that the discrete emotional labels directly affect behavior (see section 3.4.2). We will further investigate if an embedding from the emotion system, representing behaviors but encompassing a wider range of information, can better encode behaviorally meaningful information (see section 3.4.2).

In addition, the notion that behavior is highly dependent on emotional expression also raises the question of how important the sequence of emotional content is in defining behavior. We will investigate this through progressively removing the context from the sequence of emotions in the emotion-to-behavior system (see. section 3.4.2) and study how this affects the automatic behavior classification performance.

## 3.2 Background

**Emotions:** There is no consensus in the literature on a specific definition of emotion. An "emotion" is often taken for granted in itself and, most often, is defined with reference to a list of descriptors such as anger, disgust, happiness, and sadness *etc.*[23]. Oatley and Jenkins [111] distinguish emotion from mood or preference by the duration of each kind of state. Two emotion representation models are commonly employed in practice [129]. One is based on the discrete emotion theory, where six basic emotions are isolated from each other, and researchers assume that any emotion can be represented as a mixture of the basic emotions [38]. The other model defines emotions via continuous values corresponding to different dimensions which assumes emotions change in a continuous manner and have strong internal connections but blurred boundaries between each other. The two most common dimensions are arousal and valence [121].

In our work, following related literature, we will refer to basic emotions as emotions that are expressed and perceived through a short observation window. Annotations of such emotions take place without context to ensure that time-scales, back-and-forth interaction dynamics, and domain-specificity is not captured.

**Behavior:** Behavior is the output of information and signals including but not limited to those: (i) manifested in both overt and covert multimodal cues ("expressions"); and (ii) processed and used by humans explicitly or implicitly ("experience" and "judgment") [9, 105]. Behaviors encompass significant degrees of emotional perception, facilitation, thinking, understanding and regulation, and are functions of dynamic interactions [8]. Further, such complex behaviors are increasingly domain specific and subjective.

**Link between emotions and behavior:** Emotions can change frequently and quickly in a short time period [44, 104]. They are internal states that we perceive or express (*e.g.,* through voice or gesture) but are not interactive and actionable. Behaviors, on the other hand, include highly

complex dynamics, information from explicit and implicit aspects, are exhibited over longer time scales, and are highly domain specific.

For instance, "happiness", as one of the emotional states, is brought about by generally positive feelings. While within couples therapy domain, behavior "positivity" is defined in [63, 77] as "Overtly expresses warmth, support, acceptance, affection, positive negotiation".

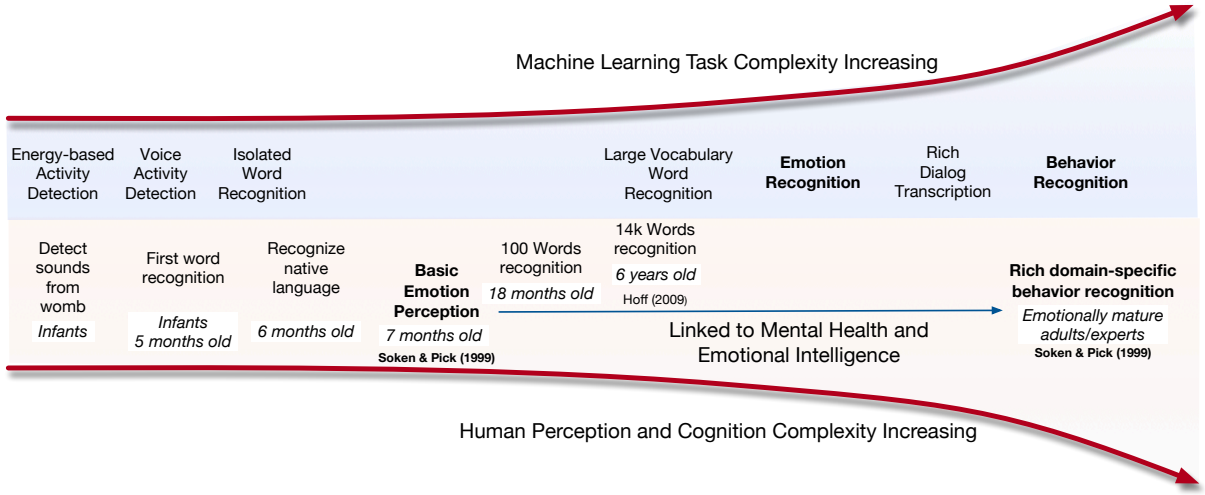



Figure 3.1: Illustration of task complexity or age of acquisition for machines and humans.

Those differences apply to both human cognition and machine learning aspects of speech capture, emotion recognition and behavior understanding as shown in Figure 3.1 [69, 133]. The increased complexity and contextualization of behavior can be seen both in humans as well as machines. For example, babies start to develop basic emotion perception at the age of seven months [133]. However, it takes emotionally mature and emotionally intelligent humans and often trained domain experts to perceive domain-specific behaviors. In Figure 3.1, we illustrate the complexity for machine processing along with the 340 age-of-acquisition for humans. We see a parallel in the increase in demands of identifying behavior in both cases.

**Motivations and goals of this work:** The relationship between emotion and behavior is usually implicit and highly nonlinear. Investigating explicit and quantitative associations between behavior and emotions is thus challenging.

In this work, based on the deep neural networks' (DNNs) underlying representative capability [11, 12], we try to analyze and interpret the relationship between emotion and behavior information through data-driven methods. We investigate the possibility of using transfer learning by employing emotion data as emotional related building blocks, or primitive features, that can describe behavior. Further, we design a deep learning framework that employs a hybrid network structure containing context dependent and reduced contextualization causality models to quantitatively analyze the relationship between basic emotions and complex behaviors.

## 3.3 Related Work

Researchers are combining machine learning techniques to build reliable and accurate emotion and behavior recognition systems. Speech emotion recognition (SER) systems, of importance in human-computer interactions, enable agents and dialogue systems to act in a more human-like manner as conversational partners [129]. On the other hand, in the domain of behavior signal processing (BSP), efforts have been made in quantitatively understanding and modeling typical, atypical, and distressed human behavior with a specific focus on verbal and non-verbal communicative, affective, and social behaviors [105]. We will briefly review the related work in the following aspects.

**Emotion quantification from speech** A dominant modality for emotion expression is speech cowie2003describing. Significant efforts [10, 45, 127] have focused on automatic speech emotion recognition. Traditional emotion recognition systems usually rely on a two-stage approach, in which the feature extraction and classifier training are conducted separately. Recently, deep learning has demonstrated promise in emotion classification tasks [60, 82]. Convolutional neural networks (CNNs) have been shown to be particularly effective in learning affective representations directly from speech spectral features [2, 5, 73, 94, 166]. Mao et al. [94] proposed to learn CNN filters on spectrally whitened spectrograms by an auto-encoder through unsupervised manners.

Aldeneh and Provost [2] showed that CNNs can be directly applied to temporal low-level acoustic features to identify emotionally salient regions. Anand and Verma [5] and Huang et al. [73] compared multiple kinds of convolutional kernel operations, and showed that the full-spectrum temporal convolution is more favorable for speech emotion recognition tasks. In addition, models with hidden Markov model (HMM) [124], recurrent neural networks (RNNs) [85, 98, 155] and the hybrid neural network combining CNNs and RNNs [72, 91] have also been employed to model emotion affect.

**Behavior quantification from speech** Behavioral signal processing (BSP) [51, 105] can play a central role in informing human assessment and decision making, especially in assisting domain specialists to observe, evaluate and identify domain-specific human behaviors exhibited over longer time scales. For example, in couples therapy [14, 108], depression [59, 106, 138, 142] and suicide risk assessment [39, 107, 109, 151], behavior analysis systems help psychologists observe and evaluate domain-specific behaviors during interactions. Li et al. [86] proposed sparsely connected and disjointly trained deep neural networks to deal with the low-resource data issue in behavior understanding. Unsupervised [87] and out-of-domain transfer learning [149] have also been employed on behavior understanding tasks. Despite these important and encouraging steps towards behavior quantification, obstacles still remain. Due to the end-to-end nature of recent efforts, low-resource data becomes a dominant limitation [37, 67, 86, 134]. This is exacerbated in BSP scenario by the difficulty of obtaining data due to privacy constraints [93, 105]. Challenges with subjectivity and low interannotator agreement [20, 148], especially in micro and macro annotation complicate the learning task. Further, and importantly such end-to-end systems reduce interpretability generalizability and domain transfer [130].

**Linking emotion and behavior quantification** As mentioned before, domain experts employ information within the context of the interaction at varying timescales to estimate the behaviors of interest [105, 148]. Specific short-term affect, *e.g.,* certain basic emotions, can be indicators of some complex long-term behaviors during manual annotation process [63, 66, 77]. These vary

according to the behavior; for example, negativity is often associated with localized cues [25], demand and withdrawal require more context [64], and coercion requires a much longer context beyond a single interaction [48]. Chakravarthula et al. [30] analyzed behaviors, such as "anger" and "satisfaction", and found that negative behaviors could be quantified using short observation length whereas positive and problem solving behaviors required much longer observation.

In addition, Baumeister et al. [8, 9] discussed two kinds of theories: the direct causality model and inner feedback model. Both models emphasize the existence of a relationship between basic emotion and complex behavior. Literature from psychology [19, 42] and social science [136] also showed that emotion can have impacts and further shape certain complex human behaviors. To connect basic emotion with more complex affective states, Carrillo et al. [26] identified a relationship between emotional intensity and mood through lexical modality. Khorram et al. [79] verified the significant correlation between predicted emotion and mood state for individuals with bipolar disorder on acoustic modality. All these indicate that the aggregation and link between basic emotions and complex behaviors is of interest and should be examined.

## 3.4 Proposed Work: Behavioral Primitives

Our work consists of three studies for estimation of behavior through emotion information as follows:

1. **Context-dependent behavior from emotion labels:** Basic emotion affect labels are directly used to predict long-term behavior labels through a recurrent neural network. This model is used to investigate whether the basic emotion states can be sufficient to infer behaviors.

2. **Context-dependent behavior from emotion-informed embeddings:** Instead of directly using the basic emotion affect labels, we utilize emotion-informed embeddings towards the prediction of behaviors.

3. **Reduced context-dependent behavior from emotion-informed embeddings:** Similar to (2) above, we employ emotion-informed embeddings. In this case, however, we investigate the importance of context, by progressively reducing the context provided to the neural network in predicting behavior.

For all three methods, we utilize a hybrid model of convolution and recurrent neural networks that we will describe in more detail below.

Through our work, both emotion labels and emotionally informed embeddings will be regarded as a type of behavior primitive, that we call ***Basic Affect Behavioral Primitive Information*** (or *Behavioral Primitives* for short, BP).

An important step in obtaining the above BP is the underlying emotion recognition system. We thus first propose and train a robust ***Multi-Emotion Regression Network*** (ER) using convolutional neural network (CNN), which is described in detail in the following subsection.

### 3.4.1 Emotion Recognition

In order to extract emotionally informed embeddings and labels, we propose a CNN based ***Multi-Emotion Regression Network*** (ER). The ER model has a similar architecture as [2], except that we use one-dimensional (1D) CNN kernels and train the network through a regression task. The CNN kernel filter should include entire spectrum information per scan, and shift along the temporal axis, which performs better than other kernel structures according to Huang et al. [73].

Our model has three components: (1) stacked 1D convolutional layers; (2) an adaptive max pooling layer; (3) stacked dense layers. The input acoustic features are first processed by multiple stacked 1D convolution layers. Filters with different weights are employed to extract different information from the same input sample. Then, one adaptive max pooling layer is employed to further propagate 1D CNN outputs with the largest value. This is further processed through dense layers to generate the emotional ratings at short-term segment level. The adaptive max pooling layer over time is one of the key components of this and all following models: First, it can cope with variable length signals and produce fixed size embeddings for subsequent dense layers; Second, it

only returns the maximum feature within the sample to ensure only the more relevant emotionally salient information is propagated through training.

We train this model as one regression model which predicts the annotation ratings of all emotions jointly. Analogous to the continuous emotion representation model [121], this multi-emotion joint training framework can utilize strong bonds but blurred boundaries within emotions to learn the embeddings. Through this joint training process, the model can integrate the relationship across different emotions, and hopefully obtain an affective-rich embedding.

In addition, to evaluate the performance of proposed ER, we also build multiple binary, single-emotion, classification models (***Single-Emotion Classification Network*** (EC)). The EC model is modified based on pre-trained ER by replacing the last linear layer with new fully connected layers to classify each single emotion independently. During training, the back propagation only updates the newly added linear layers without changing the weights of pre-trained ER model. In this case, the loss from different emotions is not entangled and the weights will be optimized towards each emotion separately. More details of experiments and results comparison are described in 3.7.

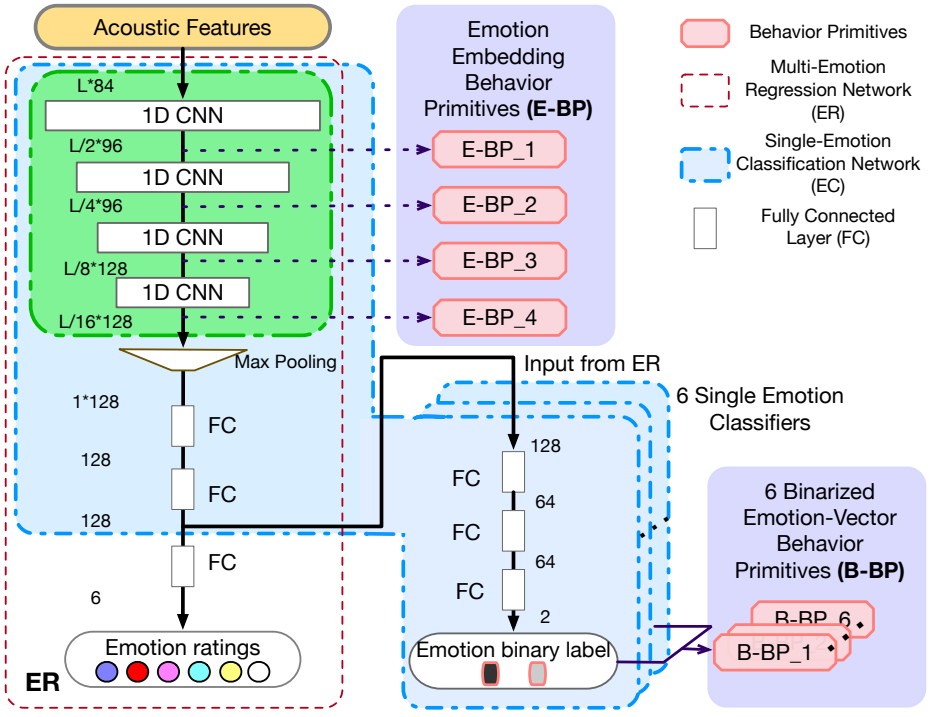



Figure 3.2: Models of ER, EC and two kinds of BPs. L is the input feature length.

As mentioned before, we employ *two kinds of behavioral primitives* in order to investigate the relationship between emotions and behaviors, and the selection of these two kinds of BP arises through the discrete, EC, and continuous, ER, emotion representation models. The two kinds of BP are: (1) The discrete vector representation of predicted emotion labels, denoted as B-BP_k, from the ***Single-Emotion Classification Network*** (EC), where $k$ means $k^{\text{th}}$ basic emotion; and (2) The output embeddings of the CNN layers, denoted as E-BP_$l$, from the ***Multi-Emotion Regression Network*** (ER) system, where $l$ represents the output from $l^{\text{th}}$ CNN layer. All these are illustrated in Figure 3.2.

### 3.4.2 Behavior Recognition through Emotion-based Behavior Primitives

We now describe three architectures for estimating behavior through ***Basic Affect Behavioral Primitive Information*** (or *Behavioral Primitives* for short, BP). The three methods employ full context of the emotion labels from the ***Single-Emotion Classification Network*** (EC), the full context from the embeddings of the ***Multi-Emotion Regression Network*** (ER) system, and increasingly reduced context from the ***Multi-Emotion Regression Network*** (ER) system.

#### Context-dependent behavior recognition from emotion labels

In this approach, the binarized predicted labels from the EC system are employed to predict long-term behaviors via sequential models in order to investigate relationships between emotions and behaviors. Such a design can inform the degree to which short-term emotion can influence behaviors. It can also provide some interpretability of the employed information for decision making, over end-to-end systems that generate predictions directly from the audio features.

We utilize the ***Single-Emotion Classification Network*** (EC) described in the previous section to obtain the predicted ***Binarized*** *Emotion-Vector* ***Behavior Primitives*** (B-BP) on shorter speech segment windows as behavioral primitives. These are extracted from the longer signals that describe the behavioral corpus and are utilized, preserving sequence, hence context, within

a recurrent neural network for predicting the behavior labels. Figure 3.3 illustrates the network architecture and B-BP_* means the concatenation of all B-BP_k, where k ranges from 1 to 6.

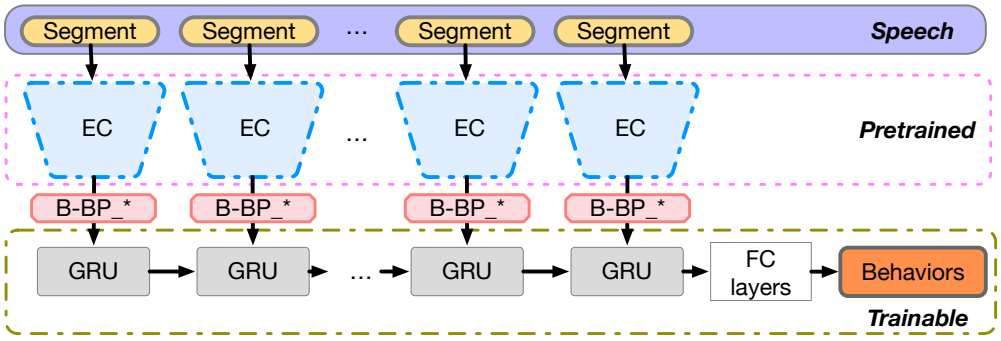



Figure 3.3: B-BP based context-dependent behavior recognition model

In short, the B-BP vectors are fed into a stack of gated recurrent units (GRUs), followed by a densely connected layer which maps the last hidden state of the top recurrent layer to behavior label outputs. GRUs were introduced by Chung et al. [35] as one attempt to alleviate the issue of vanishing gradient in standard vanilla recurrent neural networks and to reduce the number of parameters over long short-term memory (LSTM) neurons. GRUs have a linear shortcut through timesteps which avoids the decay and thus promotes gradient flow. In this model, only the sequential GRU components and subsequent dense layers are trainable, while the EC networks remain fixed.

**Context-dependent behavior recognition from emotion-embeddings**

It is widely understood that information closer to the output layer is more tied to the output labels while closer to the input layer information is less constrained and contains more information about the input signals. In our ER network, the closer we are to the output, the more raw information included in the signal is removed and the more we are constrained to the basic emotions. Given that we are not directly interested in the emotion labels, but in employing such relevant information for behavior, it makes sense to employ layers below the last output layer to capture more behavior-relevant information closer to its raw form. Thus, instead of using the binary values representing

the absence or existence of the basic emotions, we can instead employ *Emotion-**Embedding Behavior Primitives** (E-BP)* as the input representation.

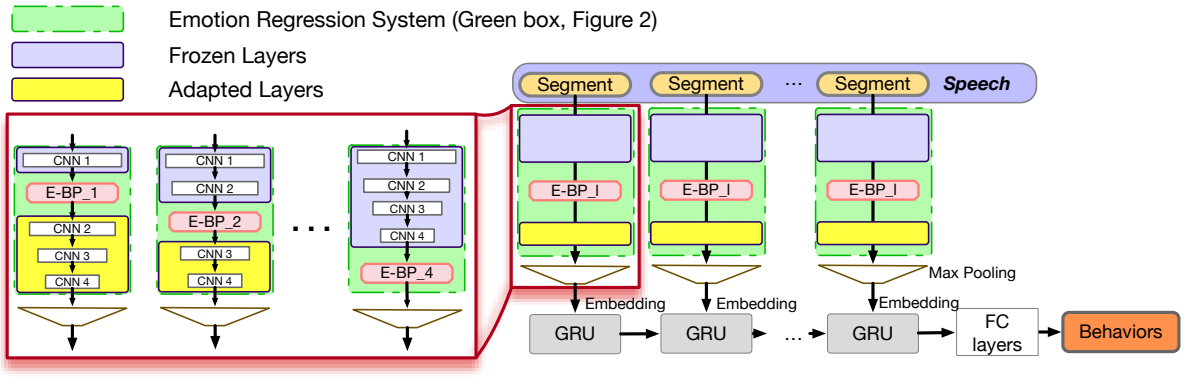



Figure 3.4: E-BP based context-dependent behavior recognition model. E-BP_*l* is the output from $l^{\text{th}}$ pretrained CNN layer. In practice multiple E-BP_*l* can be employed at the same time through concatenation. In this work we only employ the output of a single layer at a time.

The structure of the system is illustrated in Figure 3.4. After pretraining the ER, we keep some layers of that system fixed, and employ their embeddings as the *Emotion-**Embedding Behavior Primitives** (E-BP)*. We will discuss the number of fixed layers in the experiments section. This E-BP serves as the input of the subsequent, trainable, convolutional and recurrent networks.

The overall system is trained to predict the longer-term behavior states. By varying the number of layers that remain unchanged in the ER system and using different embeddings from different layers for the behavior recognition task we can identify the best embeddings under the same overall number of parameters and network architecture.

The motivation of the above is that the fixed ER encoding module is focusing on learning emotional affect information, which can be related but not directly linked with behaviors. By not using the final layer, we are employing a more raw form of the emotion-related information, without extreme information reduction, that allows for more flexibility in learning by the subsequent behavior recognition network. This allows for transfer learning [146] from one domain (emotions) to another related domain (behaviors). Thus, this model investigates the possibility of using transfer learning by employing emotional information as "building blocks" to describe behavior.

**Reduced context-dependent behavior recognition from emotion-informed embeddings**

In the above work, we assume that the sequence of the behavior indicators (embeddings or emotions) is important. To verify the need for such an assumption, in this section, we propose varying the degree of employed context. Through quantification, we analyze the time-scales at which the amount of sequential context affects the estimation of the underlying behavioral states.

In this proposed model, we design a network that can only preserve local context. The overall order of the embeddings extracted from the different local segments is purposefully ignored so we can better identify the impact of de-contextualizing information as shown in Figure 3.5.

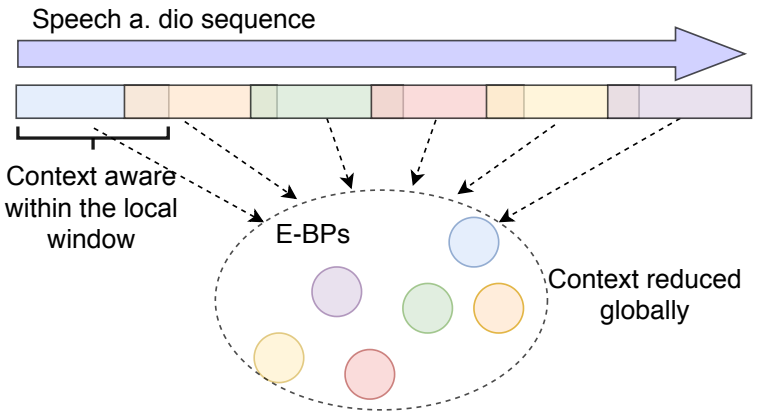



Figure 3.5: Illustration of local context awareness and global context reduction. In previous sections, the E-BPs (and B-BPs) are passed to a GRU that preserves their sequences. Here they are processed through pooling and context is removed.

In practice, this reduced-context model is built upon the existing CNN layers as in the E-BP case. We will create this reduced context system by employing only the E-BP embeddings. The E-BP embeddings are extracted from the same emotion system as before. In this case, however, instead of being fed to a recursive layer with full-session view, we eliminate the recursive layer and incorporate a variable number of CNN layers and local average pooling functions in between to adjust context view. Since the final max-pooling layer ignores the order of the input, the largest context is determined by the receptive field view of the last layer before this max-pooling. We can thus investigate the impact of context by varying the length of the CNN receptive field.

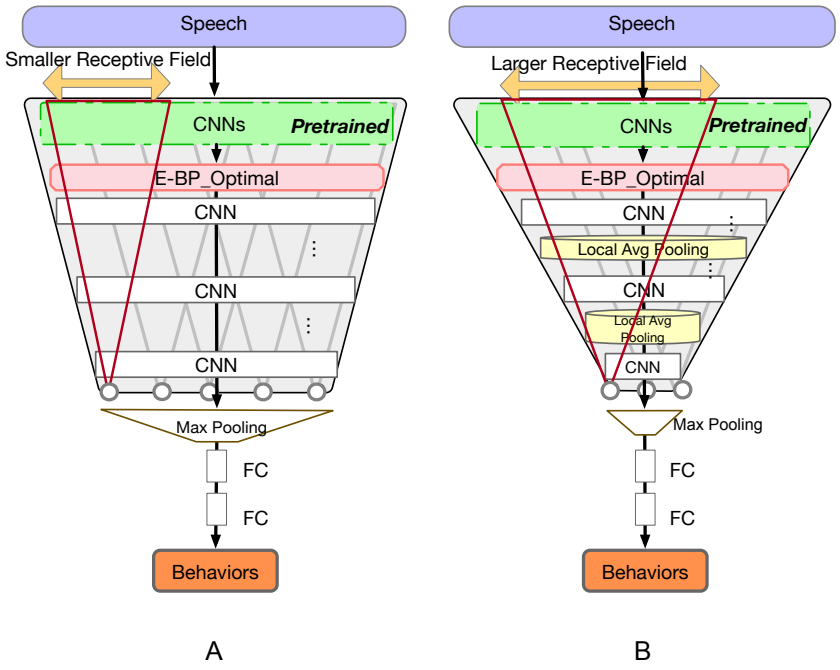



Figure 3.6: E-BP reduced context-dependent behavior recognition model. Model (A) has a smaller receptive field while model (B) has a larger receptive field because of the added local average pooling layers.

Figure 3.6 illustrates the model architecture. We extract the optimal E-BP based on the results of previous model, and then employ more CNN layers with different receptive field sizes to extract high-dimensional representation embeddings, and finally input them to the adaptive max-pooling along the time axis to eliminate the sequential information. Within each CNN receptive field, shown as red triangles in the figure, the model still has access to the full receptive field context. The max pooling layer removes context across the different receptive windows.

Furthermore, the receptive field can be large enough to enable the model to capture behavioral information encoded over longer timescales. In contrast a very small receptive area, *e.g.,* at timescale of phoneme or word, sensing behaviors should be extremely difficult [9] and can even be challenging to detect emotions [104]. The size of the receptive field is decided by the number of CNN layers, corresponding stride size, and the number of local average pooling layers in between. In our model, we adjust the size of the receptive field by setting different number of local average pooling layers under which the overall number of network parameters is unchanged.

## 3.5 Datasets

### 3.5.1 Emotion Dataset: CMU-MOSEI Dataset

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) [162] contains video files carefully chosen from YouTube. Each sample is a monologue with verified quality video and transcript. This database includes 1000 distinct speakers with 1000 kinds of topics, and are gender balanced with an average length of 7.28 seconds. Speech segments are already filtered during the data collection process, thus all speech segments are monologues of verified audio quality.

For each speech segment, six emotions (Happiness, Sadness, Anger, Fear, Disgust, Surprise) are annotated on a [0,3] Likert scale for the presence of each emotion. (0: no evidence; 1: weak evidence; 2: evidence; and 3: high evidence of emotion). This, after averaging ratings from 3 annotators, results in a 6-dimensional emotional rating vector per speech segment. CMU-MOSEI ratings can also be binarized for each emotion: if a rating is greater than 0 it is considered that there is some presence of emotion, hence it is given a true presence label, while a zero results in a false presence of the emotion.

The original dataset has 23,453 speech segments and each speech segment may contain more than one emotion presence label. Through our experiments, we use the segments with available emotion annotations and standard speaker independent split from dataset SDK [161]: Overall we have true presence in 12465 segments for happiness, 5998 for sadness, 4997 for anger, 2320 for surprise, 4097 for disgust and 1913 for fear. Due to the imbalance, accurate estimation of some emotions will be challenging. The training set consists of 16331 speech segments, while the validation set and test set consist of 1871 and 4662 sentences respectively.

### 3.5.2 Behavior Dataset: Couples Therapy Corpus

The Couples Therapy dataset is employed to evaluate complex human behaviors. The corpus was collected by researchers from the University of California, Los Angeles and the University of

Washington for the Couple Therapy Research Project [34]. It includes a longitudinal study of 2 years of 134 real distressed couples. Each couple has been recorded at multiple instances over the 2 years. At the beginning of each session, a relationship-related topic (e.g. "why can't you leave my stuff alone?") was selected and the couple interacted about this topic for 10 minutes. Each participant's behaviors were rated by multiple well-trained human annotators based on the Couples Interaction [63] and Social Support Interaction [77] Rating Systems. 31 behavioral codes were rated on a Likert scale of 1 to 9, where 1 refers absence of the given behavior and 9 indicates a strong presence. Most of the sessions have 3 to 4 annotators, and annotator ratings were averaged to obtain the final 33-dimensional behavioral rating vector. The employed part of the dataset includes 569 coded sessions, totaling 95.8 hours of data across 117 unique couples.

## 3.6 Audio Processing and Feature Extraction

### 3.6.1 Behavioral Dataset Pre-processing

For preprocessing the couples therapy corpus we employ the procedure described in [14]. The main steps are Speech Activity Detection (SAD) and diarization. Since we only focus on acoustic features extracted for speech regions, we extract the speech parts using the SAD system described in [53], and only keep sessions with an average SNR greater than 5 dB (72.9% of original dataset). Since labels of behavior are provided per-speaker, accurate diarization is important in this task. Thus, for diarization we employ the manually-transcribed sessions and a forced aligner in order to achieve high quality interlocutor-to-audio alignment. This is done using the recursive ASR-based procedure of alignment of the transcripts with audio by *SailAlign* [78].

Speech segments from each session for the same speaker are then used to analyze behaviors. During testing phase, a leave-test-couples-out process is employed to ensure separation of speaker, dyad, and interaction topics. More details of the preprocessing steps can be found in [14].

After the processing procedure above, the resulting corpus has a total of 48.5 hours of audio data across 103 unique couples and a total of 366 sessions.

### 3.6.2 Feature Extraction

In this work, we focus only on the acoustic features of speech. We utilize Log-Mel filterbank energies (Log-MFBs) and MFCCs as spectrogram features. Further, we employ pitch and energy. These have been shown in past work to be the most important features in emotion and behavior related tasks. These features are extracted using Kaldi [114] toolkit with a 25 ms analysis window and a window shift of 10 ms. The number of Mel-frequency filterbanks and MFCCS are both set to 40. For pitch, we use the extraction method in [52], in which 3 features, normalized cross correlation function (NCCF), pitch ($f_0$), the delta of pitch, are included for each frame.

After feature extraction, we obtain an 84-dimensional feature per frame (40 log-MFB's, 40 MFCC's, energy, $f_0$, delta of $f_0$, and NCFF).

## 3.7 Experiments and Results Discussion

### 3.7.1 General Settings

For emotion-related tasks, we utilize the CMU-MOSEI dataset with the given standard train, validation, test data split from [161].

For the behavior related tasks, we employ the couple therapy corpus and use leave-4-couples-out cross-validation. Note that this results in 26 distinct neural-network training-evaluation cycles for each experiment. During each fold training, we randomly split 10 couples out as a validation dataset to guide the selection of the best trained model and prevent overfitting. All these settings ensure that the behavior model is speaker independent and will not be biased by speaker characteristics or recording and channel conditions.

In our experiments, we employ five behavioral codes: *Acceptance, Blame, Positivity, Negativity* and *Sadness*, each describing a single interlocutor in each interaction of the couples therapy corpus. Table 3.1 lists a brief description[1] of these behaviors from the annotation manuals [63, 77] .

[1]Full definitions are too long to insert in this manuscript and reader is encouraged to look into [63, 77]

| Behavior | Description |
|---|---|
| Acceptance | Indicates understanding, acceptance, respect for partner's views, feelings and behaviors |
| Blame | Blames, accuses, criticizes partner and uses critical sarcasm and character assassinations |
| Positivity | Overtly expresses warmth, support, acceptance, affection, positive negotiation |
| Negativity | Overtly expresses rejection, defensiveness, blaming, and anger |
| Sadness | Cries, sighs, speaks in a soft or low tone, expresses unhappiness and disappointment |

Table 3.1: Description of behaviors

Following the same setting of [13] to reduce effects of interannotator disagreement, we model the task as a binary classification task of low- and high- presence of each behavior. This also enables balancing for each behavior resulting in equal-sized classes. This is especially useful as some of the classes, *e.g.,* Sadness, have an extremely skewed distribution towards low ratings. More information on the distribution of the data and impact on classification can be found in [50]. Thus, for each behavior code and each gender, we filter out 70 sessions on one extreme of the code (e.g., high blame) and 70 sessions at the other extreme (e.g., low blame).

Since due to the data cleaning process, some sessions may be missing some of the behavior codes, we use a mask and train only for the available behaviors. Moreover, the models are trained to predict the binary behavior labels for all behaviors together. The loss is calculated by averaging 5 behavioral classification loss with masked labels. Thus, this loss is not optimizing for any specific behavior but it is focusing on the general, latent, link between emotions and behaviors.

### 3.7.2 ER and EC for Emotion Recognition

Both the ***Multi-Emotion Regression Network*** (ER) and the ***Single-Emotion Classification Network*** (EC) are trained using the CMU-MOSEI dataset.

The ***Multi-Emotion Regression Network*** (ER) system consists of 4 layers of 1D CNN layers, adaptive max-pooling layer and followed by 3 fully connected layers with ReLU activation function. During the training, we randomly choose a segment from each utterance and represent the label of the segment using the utterance label. In our work, we employ a segment length of 1 second.

The model is trained jointly with all six emotions by optimizing the mean square error (MSE) regression loss for all emotions ratings together using Adam optimizer [80].

In a stand-alone emotion regression task, a separate network that can optimize per-emotion may be needed (through higher-level disconnected network branches), however in our work, as hypothesized above, this is not necessary. Our goal is to extract as much information as possible from the signal relating to any and all available emotions. We will, however, investigate optimizing per emotion in the EC case.

Further to the ER system, we can optimize per emotion through the ***Single-Emotion Classification Network*** (EC). This is trained for each emotion separately by replacing the pre-trained ER's last linear layer with three emotion-specific fully connected layers. We use the same binary labeling setting as described in [162]: Within each emotion, for samples with original rating value larger than zero, we assign the label 1 by considering the presence of that emotion; for samples with rating 0, we assign label 0. During training, we randomly choose 1-second segments as before. During evaluation, we segment each utterance into one-second segments and the final utterance emotion label is obtained via majority voting. In addition, the CMU-MOSEI dataset has a significant data imbalance issue: the true label in each emotion is highly under-represented. To alleviate this, during training, we balance the two classes by subsampling the 0 label esence class in every batch.

| Emotions | Anger | Disgust | Fear | Happy | Sad | Surprise |
|---|---|---|---|---|---|---|
| Methods in CMU-MOSEI [162] | 56.4 | 60.9 | **62.7** | 61.5 | 62.0 | 54.3 |
| Proposed EC | **61.2** | **64.9** | 57.0 | **63.1** | **62.5** | **56.2** |

Table 3.2: Weighted classification accuracy (WA) in percentage for emotion recognition on the CMU-MOSEI dataset. Bold numbers represent the best performing system.

In our experiments, in order to correctly classify most of the relevant samples, the model is optimized and selected based on average weighted accuracy (WA) as used in [162]. WA is defined in [145]: Weighted Accuracy = $(TP \times N/P + TN)/2N$, where $TP$ (resp. $TN$) is true positive (resp. true negative) predictions, and $P$ (resp. $N$) is the total number of positive (resp. negative) examples.

As shown in Table 3.2, we present WA of each EC system and compare them with the state-of-art results from [162].

Compared with [162], our proposed 1D CNN based emotion recognition system achieves comparable results and thus the predicted binary emotion labels can be considered satisfactory for further experiments. More importantly, our results indicate that the pre-trained ER embedding captures sufficient emotion related information and can thus be employed as a behavior primitive.

### 3.7.3 Context-dependent Behavior Recognition

The main purpose of the experiments in this subsection is to verify the relationship between emotion-related primitives and behavioral constructs. We employ both B-BP and E-BP as described below. Before that, we first use examples to illustrate the importance of context information in behavior understanding.

#### Importance of context information in behavior understanding

Prior to presenting the behavior classification results, we use two sessions from couple therapy corpus to illustrate the importance of context information in behavior understanding. Once the ***Single-Emotion Classification Network*** (EC) systems are trained, a sequence of emotion label vectors can be generated by applying the EC systems on each speech session. We choose two sessions and plot those sequences of emotion presence vectors of the first 100 seconds as an example in Figure 3.7, in which each dot represents the emotion presence (*i.e.,* predicted label equals to 1) at the corresponding time. For each emotion, the percentage of emotion presence segments is calculated by dividing the number of emotion presence segments by the total number of segments.

These two sessions are selected as an example since they have similar audio stream length and percentage of emotion presence segments but different behavior labels: the red represents one session with "strong presence of negativity" while blue represents another session with "absence of negativity". This example reveals the fact that, as we expected, the behaviors are determined not only by the percentage of affective constructs but also the contextual information. As shown in

the left Figure 3.7 (A-F), the emotion presence vectors exhibit different sequential patterns within two sessions, even though no significant distribution difference can be observed in Figure 3.7 (G).

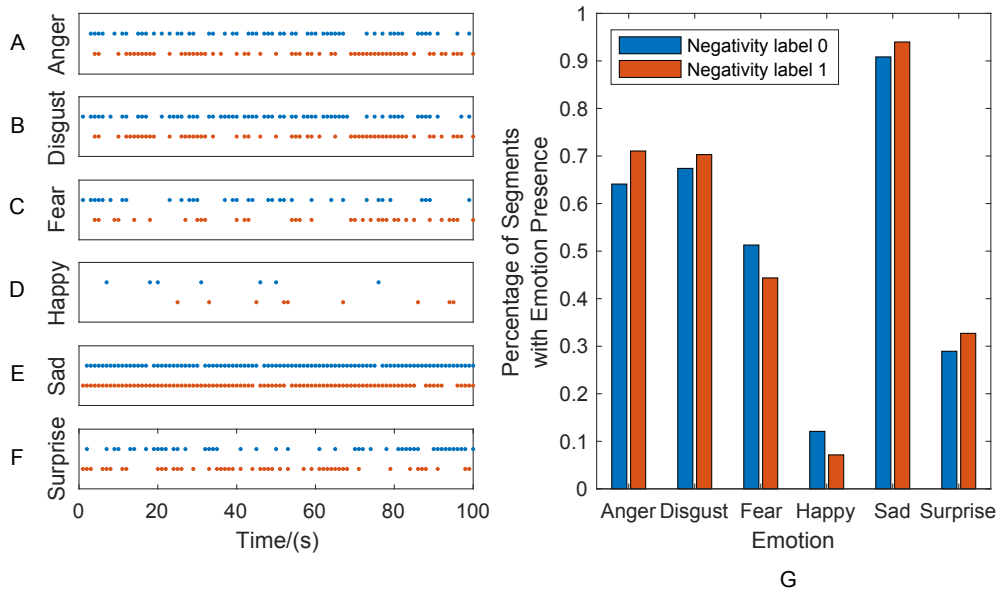



Figure 3.7: Sessions with similar percentage of emotions presence but different behavior label

### B-BP based context-dependent behavior recognition

***Binarized*** *Emotion-Vector* ***Behavior Primitives*** (B-BP) are generated by applying the ***Single-Emotion Classification Network*** (EC) systems on the couple therapy data: For each session, a sequence of emotion label vectors is generated as $E = [\mathbf{e_1}, \mathbf{e_2}, ..., \mathbf{e_T}]$, where each element $\mathbf{e_i}$ is the 6 dimensional B-BP binary label vector at time $i$. That means that $e_{ij}$ represents the presence, through a binary label 0 or 1, of emotion $j$ at time $i$. Such B-BP are the input of the context-dependent behavior recognition model that has two layers of GRUs followed by two linear layers as illustrated in Fig. 3.3.

| **Average** | Acceptance | Blame | Positivity | Negativity | Sadness |
|---|---|---|---|---|---|
| 60.43 | 61.07 | 63.21 | 59.64 | 59.29 | 58.93 |

Table 3.3: Behavior binary classification accuracy in percentage for context-dependent behavior recognition model from emotion labels

As shown in Table 3.3, the average binary classification accuracy of these five behaviors is 60.43%. Considering that the classification accuracy can reach up to 50% by chance with balanced data, our results show that behavioral states can be weakly inferred from the emotion label vector sequences. Further, we perform the McNemar test, and the results above and throughout the paper are statistically significant with $p < 0.01$. Despite the low accuracy of the behavior positivity, these results suggest a relationship between emotions and behaviors that we investigate further below.

**E-BP based context-dependent behavior recognition**

The simple binary emotion vectors (as B-BP) indeed link emotions and behaviors. However, they also demonstrate that the binarized form of B-BP limits the provided information bandwidth to higher layers in the network, and as such limits the ability to predict the much more complex behaviors. These are reflected in the low accuracies in Table 3.3 .

This further motivates the use of the *Emotion-**Embedding Behavior Primitives** (E-BP)*. As described in Figure 3.4, we construct input of the E-BP context-dependent behavior recognition system using the pretrained ***Multi-Emotion Regression Network*** (ER). These E-BP embeddings capture more information than just the binary emotion labels. They potentially capture a higher abstraction of emotional content, richer paralinguistic information, conveyed through a non-binarized version that doesn't limit the information bandwidth, and may further capture other information such as speaker characteristics or even channel information.

We employ embeddings from different layers of the ER network. The layers before the employed embedding are in each case frozen and only the subsequent layers are trained as denoted in Figure 3.4. The trainable part of the network includes several CNN layers with max pooling and subsequent GRU networks. The GRU part of the network is identical to the ones used by the context-dependent behavior recognition from E-BP.

The use of different depth embeddings can help identify where information loss becomes too specific to the ER loss objective versus where there is too much unrelated information to the behavior task.

| | **Average** | Acceptance | Blame | Positivity | Negativity | Sadness |
|---|---|---|---|---|---|---|
| None-E-BP model (Baseline) | 58.86 | 62.86 | 62.50 | 57.86 | 60.00 | 51.07 |
| E-BP_1 model | 59.79 | 64.29 | 62.86 | 60.00 | 61.07 | 50.71 |
| E-BP_2 model | 60.79 | 61.79 | 63.93 | 62.86 | 63.57 | 51.79 |
| E-BP_3 model | 65.00 | 66.07 | 69.29 | **65.36** | 69.29 | 55.00 |
| E-BP_4 model | **69.00** | **72.50** | **71.79** | **65.36** | **76.07** | **59.29** |

Table 3.4: Behavior binary classification accuracy in percentage for context-dependent behavior recognition model from emotion-embeddings. Bold numbers represent the best performing system.

In Table 3.4, the none-E-BP model, as the baseline, means all parameters are trained from random initialization instead of using the pretrained E-BP input. While E-BP_$l$ model means the first $l$ layers of the pretrained ER network are fixed and their output is used as the embedding E-BP for the subsequent system. As seen in the second column of the table, all of E-BP based models perform significantly better than the B-BP based model, which achieves an improvement of 8.57% on average and up to 16.78% for Negativity.

These results, further support the use of basic emotions as constructs of behavior. In general, for all behaviors, the higher-level E-BP s, which are closer to the ER loss function, can capture affective information and obtain better performance in behavior quantification compared with lower-level embeddings. From the description in Table 3.1, some behaviors are closely related to emotions. For example, negativity is defined in part as "Overtly expresses rejection, defensiveness, blaming, and anger", and sadness[2] is defined in part as "expresses unhappiness and disappointment". This shows that these behaviors are very related to emotions such as anger and sad, thus it's expected that an embedding closer to the ER loss function will behave better. Note that these are not at all the same though: a negative behavior may mean that somewhere within the 10 min interaction or through unlocalized gestalt information the expert annotators perceived negativity; in contrast a negative emotion has short-term information (on average 7s segment) that is negative.

[2]Which isn't necessarily perfectly aligning with the basic emotion "sad" but follows the SSIRS manual

An interesting experiment is what happens if we use a lower-ratio of emotion (out-of-domain) vs. behavior (couples-in-domain) data. To perform this experiment we use only half of the CMU-MOSEI data[3] to train another ER system, and use this less robust ER system and corresponding E-BP representations to reproduce the behavior quantification as in Table 3.4. What we observe is that the reduced learning taking place on emotional data requires the in-domain system to have prefer embeddings closer to the feature. Specifically Negativity performs equally well with layers 3 or 4 at 71.43%. Positivity performs best with layer 3 at 64.64%, Blame and Acceptance perform best with layer 2 at 71.07% and 72.86% respectively while Sadness performs best through layer 1 at 56.07%.

In the reduced data case we observe that best performing layer is not consistently layer 4. Employing the full dataset as in Table 3.4 provides better performance than using less data and in that case layer 4 (E-BP_4) is always the best performing layer, thus showing that more emotion data provides better ability of transfer learning.

### 3.7.4 Reduced Context-dependent Behavior Recognition

In the previous two sections we demonstrate that there is a benefit to transfer emotion-related knowledge to behavior tasks. We show that the wider bandwidth information transfer through an embedding E-BP is beneficial to a binarized B-BP representation. We also show that depending on the degree of relationship of the desired behavior to the signal or to the basic emotion, different layers that are closer to the input signal or closer to the output loss, may be more or less appropriate. However, in all the above cases we assume that the sequence and contextualization of the extracted emotion information was needed. That is captured and encoded through the recursive GRU layers.

We conduct an alternative investigation into how much contextual information is needed. As discussed in section 3.4.2 and shown on Figure 3.6 we can reduce context through changing the receptive field of our network prior to removing sequential information via max pooling.

[3] 11875 samples from commit:
https://github.com/A2Zadeh/CMU-MultimodalSDK/commit/f0159144f528380898df8093381c8d83fd7cc475

In this section we select the best E-BP based on average results in Table 3.4, *i.e.,* E-BP-4, as the input of the reduced context-dependent behavior recognition model. Based on E-BP-4 embeddings, the reduced context-dependent model employs 4 more CNN layers with optional local average pooling layers in between, and is followed by an adaptive max pooling layer and three fully connected layers to predict the session level label directly without sequential modules.

Since the number of parameters of this model is largely increased, dropout [137] layers are also utilized to prevent overfitting. Local average pooling layers with kernel size 2 and stride 2 are optionally added between newly added CNN layers to adjust the final size of the receptive field: The more average pooling layers we use, the larger temporal receptive field can be obtained for the same number of network parameters. We endure that the overall number of trainable parameters is the same for the different receptive field settings, which provides a fair comparison of the resulting systems. The output of these CNN/local pooling layers is passed to an adaptive max pooling before the fully connected layers as in Figure 3.6.

| | **Average** | Acceptance | Blame | Positivity | Negativity | Sadness |
|---|---|---|---|---|---|---|
| Receptive_field_ 4s | 63.43 | 65.00 | 70.00 | 58.92 | 67.50 | 55.71 |
| Receptive_field_ 8s | 62.71 | 65.00 | 69.64 | 56.79 | 66.07 | 56.07 |
| Receptive_field_ 16s | 63.36 | 63.57 | 69.64 | 60.71 | 66.42 | **56.43** |
| Receptive_field_ 32s | **66.36** | **68.21** | **73.21** | **63.21** | 71.43 | 55.71 |
| Receptive_field_ 64s | 65.57 | 66.43 | 72.86 | 62.50 | **71.79** | 54.29 |

Table 3.5: Behavior binary classification accuracy in percentage for reduced context-dependent behavior recognition from emotion-informed embeddings. Bold numbers represent the best performing system.

In Table 3.5, each model has a different temporal receptive window ranging from 4 seconds to 1 minute. For most behaviors, we observe a better classification as the receptive field size increases, especially in the range from 4 seconds to 32 seconds, demonstrating a need for longer observations for behaviors.

Furthermore, the results suggest different behaviors require different observation window length to be quantified, which is also observed by Chakravarthula et al. [30] using lexical analysis. By comparing results with different receptive window sizes, we can indirectly obtain the appropriate behavior analysis window size for each behavior code. As shown in Table 3.5, sadness has a

smaller optimal receptive field size than behaviors such as acceptance, positivity and blame. This is in good agreement with the behavior descriptions. For example, behaviors of acceptance, positivity and blame often require relatively longer observations since they relate to understanding and respect for partner's views, positive negotiation, and accusation respectively, which often require multiple turns in a dialog and context to be captured. On the other hand, sadness which can be expressed via emitting a long, deep, audible breath, and is also related to short-term expression of unhappy affect, can be captured with shorter windows.

Moreover, we find the classification of negativity reaches high accuracy when using a large receptive field. This might be contributed by the fact that the negative behavior in the couple therapy domain is complex, which is not only revealed by short term negative affect but also related to context based negotiation and hostility, and is captured through gestalt perception of the interaction.

In addition, the conclusion that most of the behaviors do not benefit much from longer than 30 seconds[4] windows matched existing literature on thin slices [3], which refer to excerpts of an interaction that can be used to arrive at a similar judgment of behavior to as if the entire interaction had been used.

### 3.7.5 Analysis on Behavior Prediction Uncertainty Reduction

Besides the verification of the improvement from B-BP based model to E-BP based models, in this section, we further analyze the importance of context information for each behavior by comparing results between E-BP based context-dependent and reduced context-dependent models. This analysis calls into question that which behavior is more context involved and to what degree.

Classification accuracy is used as the evaluation criterion in previous experiments. More generally, this number can be regarded as a probability of correct classification when a new session

[4]Note that this does not make any claims on interlocutor dynamics, talk time, turn-taking *etc.*, but just single person acoustics

comes to measure. Inspired by entropy from information theory, we define one metric named Prediction Uncertainty Reduction (PUR) and use it to indicate the relative behavior prediction and interpretation improvement among different models for each behavior.

Suppose $p_m(x) \in [0,1]$ is the probability of correct classification for behavior $x$ with model $m$. We define the uncertainty of behavior prediction as:

$$I_m(x) = -p_m(x)log_2(p_m(x)) - (1-p_m(x))log_2(1-p_m(x))$$

if $p_m(x)$ is equal to 1, $I_m(x) = 0$ there is no improvement possibility; if $p_m(x)$ is equal to 0.5, same as random prediction accuracy, the uncertainty is the largest. We further define the Prediction Uncertainty Reduction (PUR) value of behavior $x$ from model $m$ to model $n$ as:

$$R_{m \to n}(x) = I_m(x) - I_n(x)$$

We use this value to indicate improvements between different models.

We use PUR to sense the relative improvement from E-BP based context-dependent and E-BP based reduced context-dependent models respectively, to the baseline B-BP based context-dependent model. The larger value of PUR suggests the clear improvement of behavior prediction. For each behavior, for each E-BP based model, we choose the best performance model (the bold number from Table 3.4 and Table 3.5) to calculate PUR value from baseline B-BP context-dependent model.

In Figure 3.8, as expected, for most behaviors the positive PUR values verify the improvement from using informative E-BP to simple binary B-BP. In addition, the results support the hypothesis that the sequential order of affective states is one non-negligible factor of behavior analysis since the PRU of context-dependent (blue color) model is better than that of reduced context one (red color) for most behaviors.

More interestingly, for each behavior, the difference between two bars (*i.e.,* PUR difference) can imply the necessity and importance of the sequential and contextual factor of quantifying that

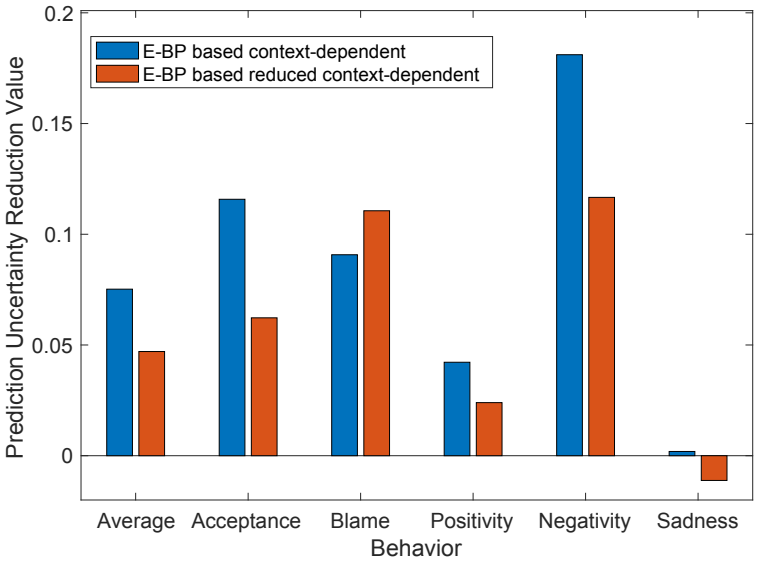



Figure 3.8: PUR optimal value of E-BP based context-dependent and reduced context-dependent models across behaviors

behavior. We notice that for "positive" or more "complex problem solving" related behaviors (*e.g.,* Acceptance, Positivity), the context based model can achieve better performance than the reduced context model. While the PUR differences from "negative" related behaviors (*e.g.,* Blame, Negativity) varies from different behaviors. For example, the behavior of acceptance, with a large PUR difference, it is more related to "understanding, respect for partner's views, feelings and behaviors", which could involve more turns in a dialog and context information. In addition, positivity requires the monitoring of consistent positive behavior, since a single negative instance within a long positive time interval would still reduce positivity to a very low rating.

In contrast, we see that although blame can still benefit from a larger contextual window, there is no benefit to employing the full context. This may infer that blame expression is more localized.

Furthermore, our findings are also congruent with many domain annotation processes: Some behaviors are potentially dominated by salient information with short range, and one short duration appearance can have a significant impact on the whole behavior rating, while some behaviors need longer context to analyze [63, 77].

However, among all behaviors, "sadness" is always the hardest one to predict with high accuracy, and there is little improvement after introducing different BPs. This could be resulting from

the extremely skewed distribution towards low ratings as mentioned in above and [14, 50], which leads to a very blurred binary classification boundary compared to other behaviors.

## 3.8 Conclusion and Future Work

In this work, we explored the relationship between emotion and behavior states, and further employed emotions as behavioral primitives in behavior classification. In our designed systems, we first verified the existing connection between basic emotions and behaviors, then further verified the effectiveness of utilizing emotions as behavior primitive embeddings for behavior quantification through transfer learning. Moreover, we designed a reduced context model to investigate the importance of context information in behavior quantification.

Through our models, we additionally investigated the empirical analysis window size for speech behavior understanding, and verified the hypothesis that the order of affective states is an important factor for behavior analysis. We provided experimental evidence and systematic analyses for behavior understanding via emotion information.

To summarized, we investigated three questions and we concluded:

1. Can the basic emotion states infer behaviors?

   The answer is yes. Behavioral states can be weakly inferred from emotions states. However behavior requires richer information than just binary emotions.

2. Can emotion-informed embeddings be employed in the prediction of behaviors?

   The answer is yes. The rich emotion involved embedding representation helps the prediction of behaviors. They also do so much better than the information-bottlenecked binary emotions.

3. Is the contextual (sequential) information important in defining behaviors?

   The answer is yes. We verify the importance of context of behavior indicators for all behaviors. Some behaviors benefit from incorporating the full interaction (10 minutes) length

while others require as little as 16 seconds of information, but all perform best when given contextual information.

Moreover, the proposed neural network systems are not limited to the datasets and domains of this work, but potentially provides a path for investigating a range of problems, such as local versus global, sequential versus non-sequential comparisons in many related areas. In addition to the relationship of emotions to behaviors, a range of other cues can also be incorporated towards behavior quantification. Moreover, many other aspects of behavior, such as entrainment, turn-taking duration, pauses, non-verbal vocalizations, and influence between interlocutors, can be incorporated. Many such additional features can be similarly developed on different data and employed as primitives; for example entrainment measures can be trained through unlabeled data [109].

Furthermore, we expect that the results of behavior classification accuracy maybe be further improved through improved architectures, parameter tuning, and data engineering for each behavior of interest. In addition, behavior primitives, *e.g.,* from emotions, can also be employed via the lexical and visual modalities.

# Chapter 4

# Speaker-invariant Affective Representation Learning via Adversarial Training

## 4.1 Introduction

Human speech signals contain rich linguistic and paralinguistic information. Linguistic information is encoded at different temporal scales ranging from phoneme to sentence and discourse levels. More importantly, speech signal encodes speaker characteristics and affective information. All information above is jointly modulated and intertwined in the human-produced speech acoustics and it is difficult to dissociate these various components simply from features, such as those from the time waveform or its transformed representations e.g., Mel filterbank energies.

Representation learning of speech [36, 87, 112], i.e., the transformation from low-level acoustic descriptors to higher-level representations, has received significant attention recently. Traditional methods focus on using supervised learning, specifically multi-task learning [27] to extract specialized representations of particular targets. However, target representations are easily contaminated by undesired factors, such as noise, channel or source (speaker) variability. These are difficult to eliminate due to the complexity and entanglement of information sources in the speech signal.

Emotion recognition systems are further greatly affected by source variability, be that speaker, ambient acoustic conditions, language, or socio-cultural context [129]. Limited domain data and labeling costs have resulted in many systems that are only evaluated within domain and are not

robust to such variability. For example mismatch between training and evaluation sets, such as speaker variations [164] and domain condition incongruity [49], make it challenging to obtain robust emotion representations across different speakers and domains.

In this work, we propose an adversarial training framework to learn robust speech emotion representations. We specifically aim to remove speaker information from the representation, as that is one of the most challenging and confounding issues in speech emotion recognition (SER). Note that many SER systems have addressed this issue through normalization of features, but these ad-hoc solutions lack generalization within complex learning frameworks [22, 131].

In our work, inspired by the domain adversarial training (DAT) [49], we propose a neural network model and an adversarial training framework with an entropy-based speaker loss function to relieve speaker variability influences. Considering the adversarial training strategy and entropy-based objective function, we name our model *Max-Entropy Adversarial Network* (*MEnAN*). We demonstrate the effectiveness of the proposed framework in SER within- and across-corpora. We show that MEnAN can successfully remove the speaker-information from extracted emotion representations, and this disentanglement can further effectively improve speech emotion classification on both the IEMOCAP and CMU-MOSEI datasets.

## 4.2 Related Work

Robust representations of emotions in speech signals have been investigated via pre-trained denoising autoencoders [54], end-to-end learning from raw audio [147], unsupervised contextual learning [87], and multi-task learning [165] etc. In a different way, we apply GANs based adversarial training to generate robust representations across domains (speaker to be specific) for speech emotion recognition. Among previous work on SER, GANs are mainly utilized to learn discriminative representations [32] and conduct data augmentation [116]. Our method is different in that we aim to disentangle speaker information and learn speaker-invariant representations for SER.

Recently, within speech applications, domain adversarial training (DAT) techniques have been applied on cross-corpus speaker recognition [154], automatic speech recognition [96, 132, 140] and SER [1, 150] to deal with the domain mismatch problems. Compared to the two most related studies [1, 150], our proposed MEnAN is different from DAT: 1) we argue that simple gradient reversal layer in DAT may not guarantee domain-invariant representation: simply flipping the domain labels can also fool the domain classifier however the learned representation is not necessary to be domain invariant. 2) we propose a new entropy-based loss function for domain classifier to induce representations that maximize the entropy of the domain classifier output, and we show the learned representation is better than DAT for speech emotion recognition.

## 4.3 Methodology

Our goal is to obtain an embedding from a given speech utterance, in which emotion-related information is maximized while minimizing the information relevant to speaker identities. This is achieved by our proposed adversarial training procedure with designed loss functions which will be introduced in this section.

### 4.3.1 Model Structure

Our proposed model is built based on a multi-task setting with three modules: the representation encoder (ENC), the emotion category classification module (EC) and the speaker identity classification module (SC). The structure of our model is illustrated in Figure 4.1.

The ENC module has three components: (1) stacked 1D convolutional layers; (2) recurrent neural layers and (3) statistical pooling layers. The sequence of acoustic spectral features is first input to multiple 1D CNN layers. The CNN kernel filters shift along the temporal axis and include the entire spectrum information per scan, which is proven to have better performance than other kernel structure settings by [73]. CNN filters with different weights are utilized to extract different information from same input features and followed by recurrent layers to capture context and

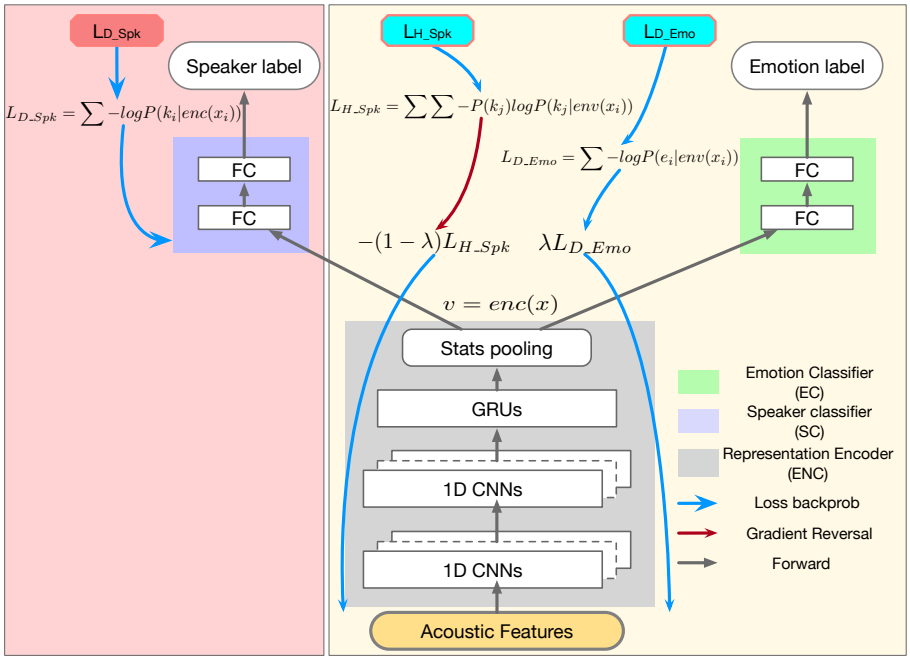



Figure 4.1: Model structure with loss propagating flow

dynamic information within each speech segment. Then, we add the statistical pooling functions, including maximum, mean and standard deviation in our model, to map a variable speech segment into an embedding with a fixed dimension.

This fixed dimension representation embedding, as the output of ENC, is further connected with two sub-task modules: the emotion classifier (EC) and speaker classifier (SC), which are both built with stacked fully connected layers.

With normal training settings, our model can be regard as a multi-task learning framework. Moreover, our model can be regarded as a speech emotion recognition system if we only keep the EC and ENC components.

### 4.3.2 Difference with Prior Work

In domain adversarial training [1, 49, 150], one gradient reversal layer is usually added to the domain classifier (SC in our case) to reverse the gradient flow in order to generate the domain-invariant (speaker-invariant) features. The usage of the gradient reversal layer ensures the (desired) lower performance of domain classifier (SC in our case), however, it often fails to guarantee the

domain (speaker) information has been fully removed [92]. For instance, in this approach, it is highly likely that even a lower performance of SC will only map a particular speaker to other target speakers with similar sounds instead of properly removing the speaker identity information, likely picking up the second-best speaker match.

Our proposed training method is different from the existing strategy as it attempts to completely remove all speaker information.

### 4.3.3 Emotion Representation Adversarial Training

We now describe the adversarial training strategy and the designed loss function in detail. Our training dataset $\mathscr{D} = \{(x_1, e_1, k_1), ..., (x_N, e_N, k_N)\}$ contains $N$ pairs of $(x_i, e_i, k_i) \in (\mathscr{X}, \mathscr{E}, \mathscr{K})$, in which speech segment $x_i$ is produced by the speaker $k_i$ with emotion $e_i$. $\mathscr{X}$, $\mathscr{E}$ and $\mathscr{K}$ are the sets of whole speech utterances, emotion labels and all speakers respectively.

Our training strategy is similar to generative adversarial networks (GANs) [57]. The system has two output paths. On one path (left Fig.4.1), we attempt to accurately estimate the speaker information (loss $L_{D\_Spk}$). On the other path (right Fig.4.1), we attempt to estimate the emotion label (loss $L_{D\_Emo}$) and remove speaker information (loss $L_{H\_Spk}$). Both estimators (SC and EC) employ the same representation encoder (ENC) but that is only updated from the right-side loss back propagation.

The output of ENC, denoted as $v$, is the speaker-invariant emotion representation we try to obtain. We have $v = enc(x)$.

#### 4.3.3.1 Training of SC

The speaker classifier (SC) can be regarded as a discriminator which is trained to distinguish speaker identities based on a given encoder output $v$ and has no influence in the training of $v$. The SC is trained by minimizing $L_{D\_Spk}$, the cross entropy function as in (4.1):

$$L_{D\_Spk} = \sum_{(x_i, e_i, k_i) \in \mathscr{D}} -logP(k_i | enc(x_i)) \tag{4.1}$$

In this training step, weights of ENC and EC are frozen. Only parameters of SC are optimized to achieve higher speaker classification accuracy from a given representation $v$.

#### 4.3.3.2 Training of ENC and EC

Under adversarial training we need to ensure the ENC output contains emotion-related information, while it is also optimized to confuse and make it difficult for SC to distinguish speaker identities.

Thus, we need to optimize ENC to increase the uncertainty or randomness of SC's outputs. Mathematically, we want to maximize the entropy value of SC's output. The entropy of SC's output, denoted as $L_{H\_Spk}$, is defined as

$$L_{H\_Spk} = \sum_{(x_i,e_i,k_i)\in\mathscr{D}} \sum_{k_j\in\mathscr{K}} -P(k_j) log P(k_j|env(x_i)) \tag{4.2}$$

Maximizing entropy would promote equal likelihood for all speakers:

$$P(k_j|env(x_i)) = P(k_q|env(x_i)) \quad \forall k_j, k_q \in \mathscr{K} \tag{4.3}$$

This differs, as mentioned above, from simply picking up a different speaker, since that may lead into selecting a "second-best" similar sounding speaker. Our proposed loss function removes all (correct or wrong) speaker information.

In addition, the performance of emotion classifier is optimized by minimizing the cross entropy loss $L_{D\_Emo}$ from EC's output:

$$L_{D\_Emo} = \sum_{(x_i,e_i,k_i)\in\mathscr{D}} -log P(e_i|env(x_i)) \tag{4.4}$$

To combine these two objective functions above together, we flip the sign of $L_{H\_Spk}$ to do a gradient reversal and minimize the weighted overall loss sums. The final objective loss function is written as:

$$L(\theta_{ENC}, \theta_{EC}) = \lambda L_{D\_Emo} - (1-\lambda) L_{H\_Spk} \tag{4.5}$$

where $\lambda \in (0, 1)$ is a parameter adjusting the weighting between two types of loss functions.

In this training step, weights of SC are frozen. Only parameters of ENC and EC are optimized. Modules with corresponding loss back propagation flows are shown in Fig.4.1. With this iterative training scheme, we expect the proposed model can ultimately relieve the impact of speaker variability thus improve the SER performance.

## 4.4 Dateset

Two datasets are employed to evaluate the proposed MEnAN based emotion representation learning in our work:

The **IEMOCAP** dataset [21] consists of five sessions of speech segments with categorical emotion annotation, and there are two different speakers (one female and one male) in each session. In our work, we use both improvised and scripted speech recordings and merge *excitement* with *happy* to achieve a more balanced label distribution, a common experiment setting in many studies such as [54, 55, 110]. Finally, we obtain 5,531 utterances selected from four emotion classes (1,103 angry, 1,636 happy, 1,708 neutral and 1,084 sad).

The **CMU-MOSEI** dataset [162] contains 23,453 single-speaker video segments carefully chosen from YouTube. This database includes 1000 distinct speakers, and are gender balanced with an average length of 7.28 seconds. Each sample has been manually annotated with a [0,3] Likert scale on the six basic emotion categories: happiness, sadness, anger, fear, disgust, and surprise. The original ratings are also binarized for emotion classification: for each emotion, if a rating is greater than zero, it is considered that there is presence of that emotion, while a zero results in a false presence of that emotion. Thus, each segment can have multiple emotion presence labels.

IEMOCAP provides a relatively large number of samples within each combination across different speakers and emotions, making it feasible to train our speaker-invariant emotion representation. We mainly use CMU-MOSEI for evaluation purposes, since it includes variable speaker identities, and to establish cross-domain transferability of MEnAN.

## 4.5 Experiment Setup

**Feature extraction:** In this work we utilize 40 dimensional Log-Mel Filterbank energies (Log-MFBs), pitch and energy. All these features are extracted with a 40 ms analysis window with a shift of 10 ms. For pitch, we employ the extraction method in [52], in which the normalized cross correlation function (NCCF) and pitch (f0) are included for each frame. We do not perform any per-speaker/sample normalization.

**Data augmentation:** To enrich the dataset, we perform data augmentation on IEMOCAP. Similar to [143], we create multiple data samples for training by slightly modifying the speaking rate with four different speed ratios, namely 0.8, 0.9, 1.1 and 1.2.

**General settings:** To obtain a reliable evaluation of our model, we need to ensure unseen speakers for both validation and testing. Thus, we conduct 10 fold leave-one-speaker out cross-validation scheme. More specifically, we use 8 speakers as the training set, and for the remaining session (two speakers), we select one speaker for validation and one for testing. We then repeat the experiment with the two speakers switched.

In addition, considering the variable length of utterances, we only extract the middle 14s to calculate acoustic features for utterance whose duration is longer than 14s (2.07% of the total dataset) [40], since important dynamic emotional information is usually located in the middle part and lengthy inputs would have negative effect on emotion prediction [74]. For utterances shorter than 14s, we use the cycle repeat mode [74] to repeat the whole utterance to the target duration. The idea of this cycle repeat mode is to make emotional dynamic cyclic and longer, which facilitates the training process of utterances of variable duration.

### 4.5.1 Model Configurations

The detailed model parameters and training configurations are shown in Table 4.1.

| | |
|---|---|
| Training details: | Adam optimizer (lr=0.001) + polynomial learning rate decay ;<br>batch size=16; epochs=300; $\lambda$=0.5 |
| ENC | Conv1D(in_ch=43,out_ch=32, kernel size=10, stride=2, padding=0), PReLU<br>Conv1D(in_ch=32,out_ch=32, kernel size=5,stride=2, padding=0), PReLU<br>GRU(in_size=32, hidden_size=32, num_layers=1)<br>Linear(in=32, out=32), PReLU<br>Statistical Pooling[Mean, Std, Max] |
| EC | Linear(in=32, out=32) PReLU<br>Linear(in=32, out=10) PReLU<br>Linear(in=10, out=4 ) |
| SC | Linear(in=32, out=32) PReLU<br>Linear(in=32, out=10) PReLU<br>Linear(in=10, out=8) |

Table 4.1: Model structure and training configuration details

## 4.6 Results and Discussion

### 4.6.1 Evaluation on IEMOCAP

For comparison purposes, we also train the EC only model, multi-task learning model and DAT model [150] with regular cross entropy loss under the same configuration. Both the weighted accuracy (WA, the number of the correctly classified samples divided by the total number of samples) and the unweighted accuracy (UA, the mean value of the recall for each class) are reported. The Table 4.2 shows the emotion classification accuracy (%) on both validation and testing, and we also include their differences ($\Delta$).

| | WA | | | UA | | |
|---|---|---|---|---|---|---|
| | Val | Test | $\Delta$ | Val | Test | $\Delta$ |
| EC only model | 58.50 | 55.92 | -2.56 | 59.94 | 57.45 | -2.49 |
| Multi-task model | 59.24 | 55.90 | -3.34 | 60.52 | 57.28 | -3.24 |
| DAT model | 58.28 | 56.68 | -1.60 | 60.16 | 58.48 | -1.68 |
| Proposed MEnAN | 58.85 | **58.62** | -0.23 | 60.24 | **59.91** | -0.33 |

Table 4.2: Classification accuracy (%) comparison on IEMOCAP

First, we observe that our model achieves the best classification accuracy in the test case among all models. To the best of our knowledge, the best results from the literature on IEMOCAP with

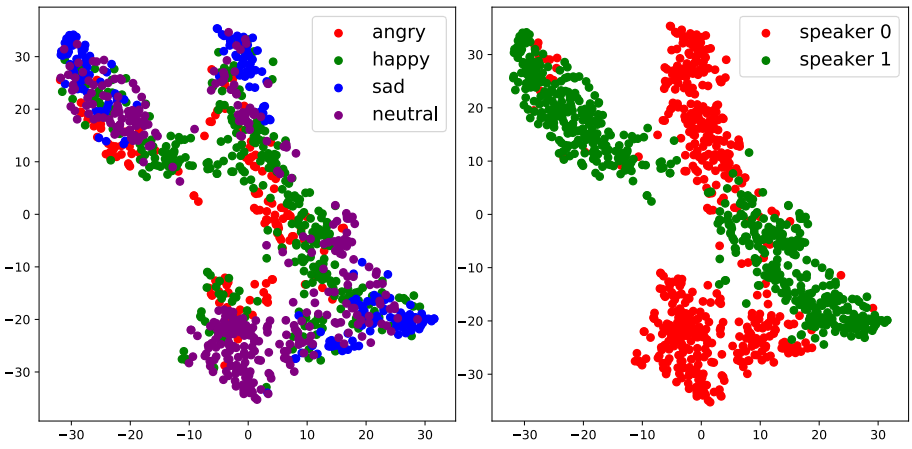
similar settings are generally around 60% [110, 156]. We achieve a UA of 59.91% which is comparable with the state of the art results. However, strict comparisons remain difficult because there are no standardized data selection rules or train/test splits. For example, some did not use speaker independent split [110] or only used improvised utterances. Some did not clearly specify which speaker in each session was used for validation and testing respectively [85] or performed per-speaker normalization in advance [22, 117].

Second, we notice that there is a large difference of $\Delta$ value among all four models. Compared with others, we find that the multi-task learning model can achieve a better performance on the validation set. However, the extra gain from speaker information can also lead a significant mismatch during the evaluation of unseen speakers, as indicated by the large value of $\Delta$. Compared with DAT model, our MEnAN model gains better classification accuracy with smaller $\Delta$. This supports our claim of the MEnAN's advantage over DAT. The small $\Delta$ in our model suggests our embedding has better generalization ability and is more robust to unseen speakers. To illustrate this, we plot t-SNE of emotion representation, i.e., $enc(x)$, on two unseen speakers.

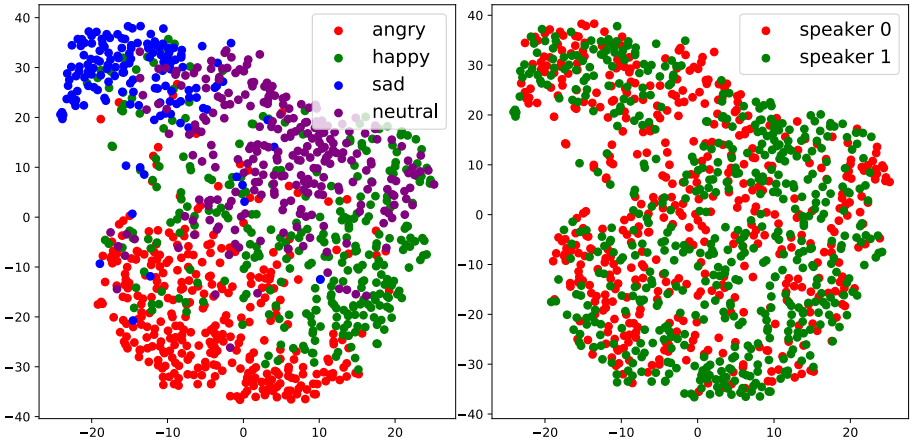
As shown in Fig.4.2, in the multi-task learning setting, it is obvious that the speaker's characteristics and emotion information are entangled with each other, which makes this representation less generic on unseen speakers. For our proposed MEnAN, the speaker representations of different speakers on the 2D space are well mixed and independent of speaker labels; while different emotion segments are more separable in the embedding manifold. These results further demonstrates the effectiveness and robustness of proposed model.

**Evaluation on CMU-MOSEI**

In addition, we test our system on the CMU-MOSEI dataset (cross-corpus setting). As mentioned before, the CMU-MOSEI has a large variability in speaker identities, which is a suitable corpus to evaluate our model's generalization ability on unseen speakers. It also introduces a challenge stemming from the different annotation methodology and the inherent effect on the interpretation of labels.




(a) Multi-task learning model




(b) MEnAN model

Figure 4.2: t-SNE plot of emotion embedding with both 4 emotion labels (left) and 2 speaker labels (right) for multi-task model and our proposed MEnAN model.

To match emotion labels of IEMOCAP, we only consider samples with positive ratings in the categories of "happiness", "sadness" and "anger". Samples with zero ratings of all six emotion categories are also included with the label "Neutral". Finally, 22,323 samples are selected and four-class emotion classification evaluation are performed. The prediction is considered to be correct if the rating of that predicted emotion originally has a positive value. In Table 4.3, we report the mean, minimum and maximum of the classification accuracy (%) evaluated on the pretrained model of each fold from the 10-fold cross validation of IEMOCAP.

We observe that MEnAN model has the best performance among all three models, and it achieves better classification accuracy with 1.89% improvement on the mean value and with 4.89%

| | mean | min | max |
|---|---|---|---|
| EC only model | 31.35 | 27.14 | 34.96 |
| DAT model | 32.34 | 25.91 | 37.85 |
| Proposed MEnAN | **33.24** | **28.84** | **39.85** |

Table 4.3: Emotion classification accuracy (%) on CMU-MOSEI

on the best model compared with the EC only model. Considering that all speakers of these evaluation samples are not seen during the training, these results suggest our adversarial training framework can provide more robust emotion representation with better speaker-invariant property and achieve improved performance in the emotion recognition task.

## 4.7 Conclusion

Compared with other representation learning tasks, the extraction of speech emotion representation is challenging considering the complex hierarchical information structures within the speech, as well as the practical low-resource (labeled) data issue. In our work, we use an adversarial training strategy to generate speech emotion representations while being robust to unseen speakers. Our proposed framework MEnAN, however, is not limited to the emotion recognition task, and it can be easily applied to other domains with similar settings e.g., cross-lingual speaker recognition. For further work, we plan to combine the domain adaption techniques with our proposed model to employ training samples from different corpora. For example, we can utilize speech utterances from speaker verification tasks to obtain more robust speaker information.

# Chapter 5

# Unsupervised Speech Representation Learning for Behavior Modeling using Triplet Enhanced Contextualized Networks

## 5.1 Introduction

Understanding human behaviors [105] through observational study has been one of the core themes in fields such as psychology and sociology. Human behaviors encompass rich information: from emotional expression, processing, and regulation to the intricate dynamics of interactions including the context and knowledge of interlocutors and their thinking and problem-solving intent [88]. Furthermore, the behavioral constructs of interest are often dependent on the domain of interaction [105]. Hence characterization of human behavior usually requires domain-specific knowledge and adequate windows of observation. Notably, across psychological health science and practice [16] such as couple therapy [34], suicide cognition evaluation [18] and addiction counseling [159], this is exemplified in the definition and derivation of a variety of domain-specific behavior constructs (e.g., blame and affect patterns exhibited by partners, suicidal ideation of an individual at risk, and empathy expressed by a therapist) to support specific subsequent plan of action.

Human speech offers rich information about the mental state and traits of the talkers. Vocal cues, including speech and spoken language as well as nonverbal vocalizations and disfluency patterns, have been shown to be informationally relevant in the context of human behavior (*e.g.,* in marital interaction [7], in motivational interviewing [4, 75, 102]). Many automatic computational

approaches that support measurement, analysis, and modeling of human behaviors from speech have been investigated in affective computing [84], social signal processing [152] and behavioral signal processing (BSP) [105].

Automated behavior modeling from speech however remains a challenging domain. Behavior annotations used for (supervised) modeling are usually obtained from well-trained human annotators, in a process that is both complex and expensive. Moreover, the prevalence of many specific behaviors of interest in a given interaction inherently tend to be low. As a result, the amount of annotated training data available for supervised behavior modeling are relatively small compared to other speech related training tasks.

In addition, behavior analyses tend to be guided by target domain needs. For example, in looking for markers of behavior change in addiction, therapists look for language which reflects changes of addictive habits [6]. In suicide prevention [39], behavioral patterns related to reasons for living and emotional bonds are deemed relevant. Thus, behavior models built with domain-specific constructs and data may not be directly and easily adaptable across domains.

Recently, unsupervised and self-supervised learning [33, 81] have shown the benefits of using large amounts of unlabelled data to extract informative representations. Given the low availability of annotated behavioral data sets, representation learning through unsupervised ways can provide a promising avenue for behavioral modeling. This becomes especially relevant where unlabelled or weakly-labelled speech is often the only available resource.

In unsupervised representation learning, context information has been used for a range of applications [41, 56]. For example, in Natural Language Processing (NLP), word and sentence embedding methods attempt to compress the shared structural information between neighboring words, phrases or sentences. Such compressed structural information, referred as the *context*, resides at a longer scale than either of just two neighboring isolated words, phrases or sentences. In behavior analysis, context information is important: When domain experts attempt to evaluate behaviors, a large observation window is often employed to observe the context e.g., a full interaction session. Within the frame of observation, the target behavior is assumed to remain constant.

In this paper, we describe unsupervised methods to extract behavior related representations from speech under the behavioral stationarity assumption. In addition, we also employ metric learning techniques to improve representation learning directly in the behavior related manifold space. We investigate whether out-of-domain data corpora can be employed for behavior representation learning and, for quantification and analysis of target behavioral constructs. Moreover, we show the proposed unsupervised model can provide domain experts with dynamic behavior change trajectories, which is helpful to facilitate the annotation process and indicate salient behavior regions. To evaluate our proposed methods, we use a couple therapy dataset comprising audio recordings of problem solving interactions as well as speech files from a variety of application domains such as talk shows to show the similarity in the learned behavior manifolds.

## 5.2 Related Work and Motivation

The human speech audio includes information about the state and trait of the talkers ranging at varying levels of linguistic scales, *e.g.,* phonemic, prosodic, and discourse, to the level of the larger socio-emotional communication context.

Traditional supervised behavior recognition systems mainly depend on two aspects: one is the representative feature of the target behavior and, the other is the choice of the classification model. To capture the vocal cues for behavior recognition, traditional computational approaches [14, 86, 108, 126, 157] use a range of hand-crafted low-level descriptors (LLDs) (*e.g.,* f0, intensity, MFCCs (Mel-Frequency Cepstral Coefficients) *etc.*) with statistical functionals (*e.g.,* mean, median, standard deviation, *etc.*) to represent segment- or utterance-level features. Based on these raw acoustic LLDs and their functionals, classifiers such as Support Vector Machines (SVM), k-Nearest Neighbors (kNN) and Hidden Markov Models (HMM) *etc.* have been employed [45, 71, 125, 157, 163].

Over the last few years, many affect and behavior recognition systems have employed Deep Neural Network (DNN) models to extract intermediate representations [60, 86, 90]. Further, sequential models [85, 88] have been used to account for the context effect. However, the success of DNN models heavily relies on the availability of large-scale datasets. A large amount of training data with annotated labels are usually unavailable in the human behavioral related domain, which largely inhibits the use of DNN based supervised frameworks in behavioral modeling tasks [86].

Different from supervised approaches, in this paper, we focus on context-rich techniques for extracting behavior representations in an unsupervised manner. Contextual information has played a significant role in unsupervised representation learning for a range of applications. For example, in NLP, contextual information is employed to generate general word or sentence embeddings (*e.g.,* Word2Vec [56, 99, 100], BERT [41] *etc.*) for downstream tasks. In speech representation learning [81], unsupervised techniques such as autoregressive modeling [36] and self-supervised modeling [101, 112, 141] employ temporal context information for extracting speech representation. In our prior behavior modeling work, an unsupervised representative learning framework was proposed [87], which showed the promise of learning behavior representations based on the behavior stationarity hypothesis that nearby segments of speech share the same behavioral context. A similar framing was used by [109] to evaluate interpersonal entrainment through an unsupervised turn-level distance measure .

In addition, metric learning is often employed to directly learn representations with an appropriate distance metric. For instance, siamese networks [17] and triplet networks [70] are neural networks suitable for direct representation learning by minimizing the contrastive loss or triplet loss calculated in the latent embedding space. These techniques have shown promising results in face verification and identification [122] as well as in speech tasks such as speaker diarization and verification [76, 135].

The goal of this work is to identify, in an unsupervised manner, a latent manifold in which behavior characteristics are retained while other unrelated information are minimized. We believe

the unsupervised representation learning under the behavioral stationarity assumption can take advantage of diverse out-of-domain datasets for improving behavioral modeling.

## 5.3 Unsupervised Speech Representation Learning for Human Behavior Modeling

We present two frameworks for unsupervised behavior modeling. The first one is the *Deep Contextualized Network* (DCN) initially introduced by [87], and the second is a new hybrid approach enhanced by a triplet loss, referred to as *Triplet Enhanced Deep Contextualized Network* (TE-DCN). The overarching goal is to build a function that can map behavior related information from raw acoustic features into the behavioral manifold, where similar behaviors can be clustered closer than they are in the original acoustic feature space, while distinct behavior types can maintain larger distances between one another.

### 5.3.1 Behavioral Stationarity Assumption

Toward designing the unsupervised modeling, we wish to invoke some domain knowledge about human behaviors. An important observation is that complex human behaviors often manifest over longer time scales, and remain relatively constant within a sufficiently long temporal window, and in fact need a sufficiently long observation time for human annotation of target behavioral constructs (*e.g.,* ranging from 30 seconds to 10 minutes [63, 66]). For example, in couples therapy interaction behaviors associated with constructs such as sadness and blame can last over several conversational exchanges.

Based on these observations, we make the *behavior stationarity assumption*: Human behaviors are deemed to remain constant within a sufficiently long window (*i.e.,* behavior stationary region). This means that by observing target behaviors within a long observation window (*e.g.,* 30 seconds), it is likely that the same or similar behavioral states are observed.

### 5.3.2 Deep Contextualized Network

The Deep Contextualized Network (DCN) has an encoder-decoder structure, similar to an autoencoder. But in contrast, rather than just training to reconstruct the input itself, the proposed DCN model is trained to reconstruct neighboring frames sharing the same behavioral context. The overall framework is shown in Figure 5.1.

For the $i^{th}$ frame of acoustic feature $x_i$, the reconstruction frame $x_j$ is selected from $i$-$k$ to $i$+$k$ excluding the $i^{th}$ frame, where $k$ is the maximum sampling shift size within the behavior stationary region, in which we assume the behavioral context to remain constant. During the training, we optimize the network to minimize the reconstruction loss:

$$\begin{aligned} \mathscr{L}_{DCN}(x_i, x_j) &= \sum_{(x_i,x_j)\in\mathscr{D}} \left\| f_{DCN}(x_i) - x_j \right\|_2^2 \\ &= \sum_{(x_i,x_j)\in\mathscr{D}} \left\| \widehat{x_i} - x_j \right\|_2^2 \end{aligned} \tag{5.1}$$

where the training dataset $\mathscr{D}$ consists of input tuples $(x_i, x_j)$, and $\widehat{x_i}$ is the output of DCN. After training, the hidden bottleneck layer's output is used as the behavior representation for evaluation.

The representation from the hidden layers of DCN compresses the shared information between input and output. Once we input behavior-relevant acoustic features into the DCN, the trained encoder can be regarded as a feature extractor for obtaining the shared information between input and output. The choice of features and the model's structure can promote behavior as common information. Assuming the adequacy of the behavioral stationarity assumption and the behavioral information contained in the input features, the model will ensure bottleneck embedding features that are relevant to the relatively constant factors, *i.e.,* the behavior related features.

### 5.3.3 Triplet Enhanced Deep Contextualized Network

In this section, we introduce the Triplet Enhanced Deep Contextualized Network (TE-DCN), in which we use metric learning techniques to improve the performance of DCN.

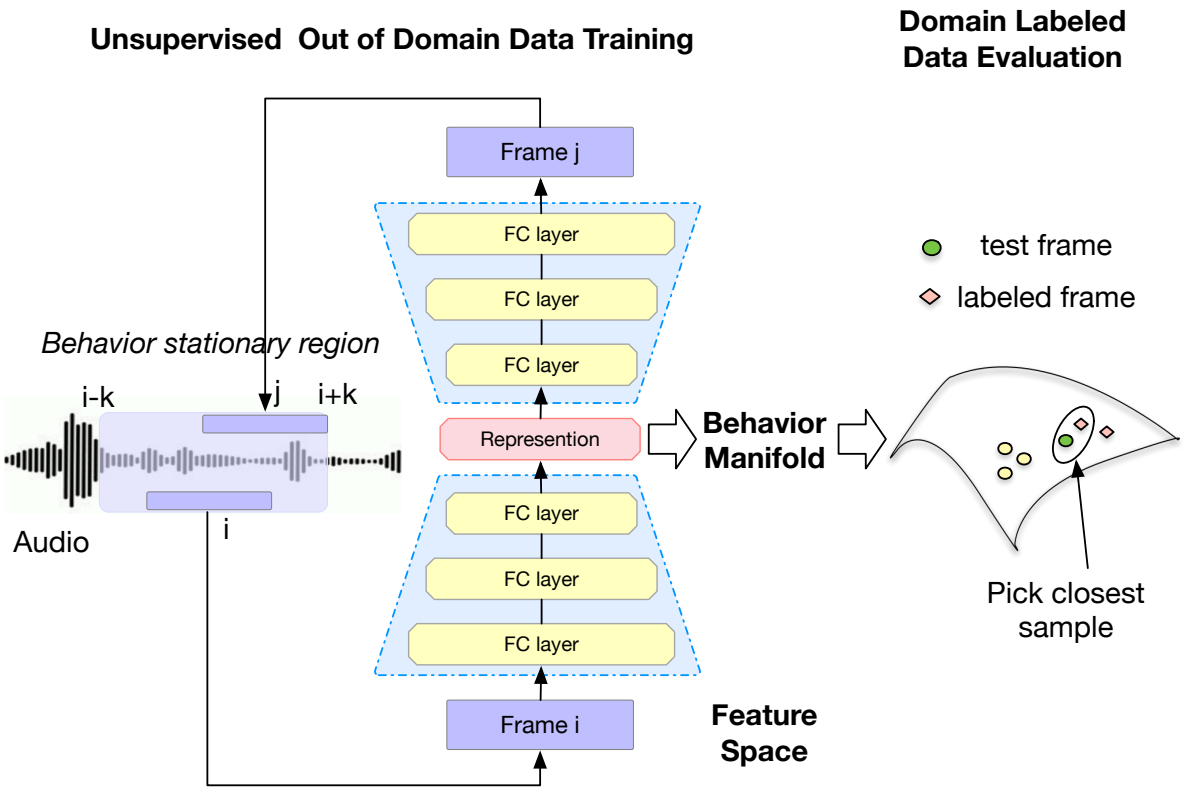



Figure 5.1: Behavior representation learning framework via the DCN model. During training, the model encodes neighboring frames with a DCN. The choice of features promotes the behavior information as the extracted common information in the behavior manifold. During evaluation, similarity comparison is made by calculating distance within the behavior manifold.

Metric learning aims to find an input-output mapping function over a vector space and is explicitly trained to build distance metrics among vectors. Triplet loss enables neural networks to keep the embeddings belonging to the same class close to each other, while moving embeddings with different classes far apart. It is used for representation learning by direct optimizing in the latent embedding space. Suppose the training dataset $\mathscr{D}$ consists of input tuples $(x_a, x_p, x_n)$: one anchor $x_a$, one positive sample $x_p$ which belongs to the same class as the anchor and one negative sample $x_n$ from a different class. The corresponding embedding $(e_a, e_p, e_n)$ is generated by neural networks, and the model is trained to minimize the following loss function:

$$\mathscr{L}_{triplet}(x_a, x_p, x_n) = \sum_{(x_a, x_p, x_n) \in \mathscr{D}} \max[0, m + D(e_a, e_p) - D(e_a, e_n)] \tag{5.2}$$

where $D(\cdot,\cdot)$ denotes the distance metric and $m$ is the parameter of margin value. This objective function targets to ensure that in the embedding space, the anchor sample is closer to the positive sample than it is to the negative sample by at least a margin $m$.

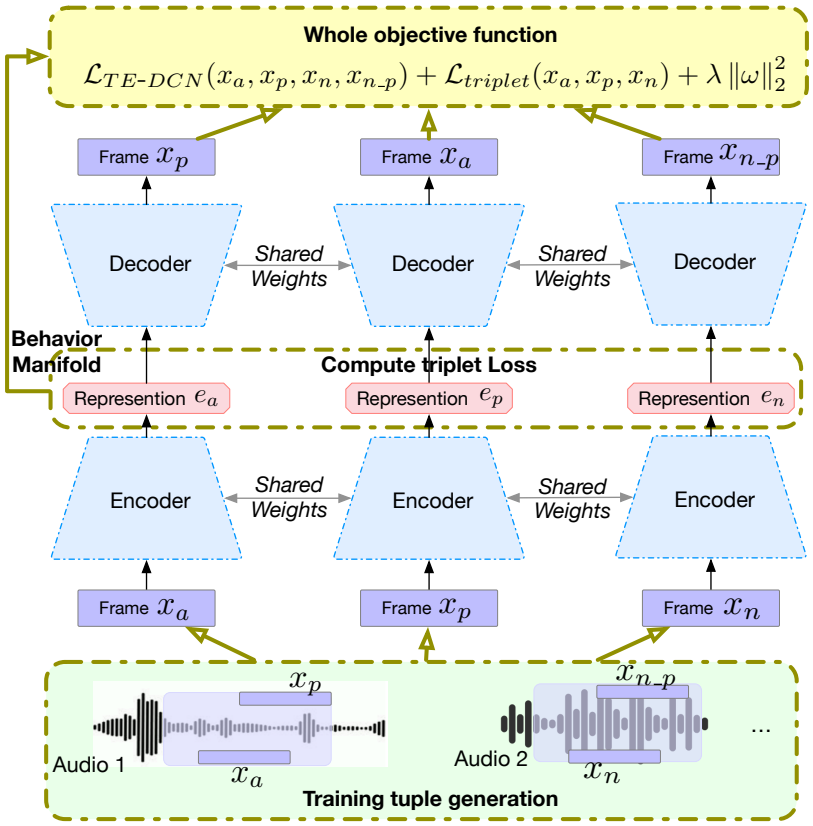



Figure 5.2: Behavior representation learning framework via the TE-DCN model. The model has a triplet structure with shared weights. Audio 1 and 2 can be two temporally-distant regions or two different files. During training, the model is optimized to minimize both reconstruction loss and triplet loss. During evaluation, representation similarity comparison is performed within the behavior manifold.

The architecture of the proposed TE-DCN is shown in Figure 5.2. The added triplet loss is motivated by a similar idea, in which we want to keep shared behavioral information between neighboring frames while disregarding other nuisance factors (with respect to the target behavioral construct), such as speaker and channel information. The DCN is only trained by frames with behavioral similarity in the stationary region, while the triplet loss requires frames from difference regions, which likely have different behaviors as well as distinct acoustic and speaker information.

The model takes tuples of four elements $(x_a, x_p, x_n, x_{n\_p})$ for training, where $(x_a, x_p)$ are two frames within a behavioral observation window (assuming stationarity within it). Let $(x_n, x_{n\_p})$ be another frame pair selected from a temporally-distant region, and is potentially more likely to contain a different target behavior. Given this tuple, we have the input-output pair for our TE-DCN model: $(x_a, x_p, x_n)$ and $(x_p, x_a, x_{n\_p})$. Each branch of the model can be regarded as a DCN framework with shared parameters to reconstruct context frames. Note the model also employs triplet loss on the intermediate embedding in the behavior manifold space.

We simultaneously optimize two objective functions (5.1) and (5.2) jointly. Thus, the overall loss function is:

$$\begin{aligned} \mathscr{L}_{total}(x_a, x_p, x_n, x_{n\_p}) = \mathscr{L}_{TE\text{-}DCN}(x_a, x_p, x_n, x_{n\_p}) \\ + \mathscr{L}_{triplet}(x_a, x_p, x_n) + \lambda \left\|\omega\right\|_2^2 \end{aligned} \tag{5.3}$$

where the reconstruction loss $\mathscr{L}_{TE\text{-}DCN}$ in Equation (5.3) is defined as:

$$\begin{aligned} \mathscr{L}_{TE\text{-}DCN}(x_a, x_p, x_n, x_{n\_p}) = \mathscr{L}_{DCN}(x_a, x_p) + \mathscr{L}_{DCN}(x_p, x_a) \\ + \mathscr{L}_{DCN}(x_n, x_{n\_p}) \end{aligned} \tag{5.4}$$

Since the model can take advantage of practically available (potentially) unlimited amount of unlabelled corpora, to prevent ovefitting in the training domain, as shown in Equation (5.3), we amend the objective function with an $L_2$ regularization term.

The encoder of the model tries to map the frame's acoustic features to a "behavioral manifold". On the one hand, the neighboring frames are trained to cluster together while on the other, frames from different regions are trained to be farther away in the representation space.

We propose the TE-DCN model to provide improvement to the DCN model in the following aspects:

1. **Introduce discriminative information.**

   The metric learning is employed in the TE-DCN, which enables the model not only capture the behavioral contextual information within neighboring frames but also preserve the

discriminative information by imposing triplet constraints. By adding the triplet loss, the model not only reconstructs the frames pairs within behavioral stationary regions, but also uses "negative" frames to reduce nuisance factor effects, such as speaker characteristics and channel information, in the behavior manifold construction.

2. **Select the behavioral embedding layer explicitly.**

   In DCN, the smallest bottleneck layer's output embedding is used as the behavior representation for evaluation. However, it is not guaranteed that this is the optimal choice among all the hidden layers. In TE-DCN, though we choose the same smallest bottleneck embedding, we add specific constraints and optimization on that selected layer, and can ensure that it preserves more targeted information compared to other embeddings.

3. **Uniformize the distance metric for training and evaluation.**

   After learning from unlabelled data, during testing, we use distance metrics to evaluate the similarity within the behavioral manifold. However, the choice of distance measures is not specified within the DCN model. For example, either Cosine distance or Euclidean distance can be selected. While in TE-DCN, an explicit distance metric is used during training, which makes it easy to use for testing.

## 5.4 Datasets

For the unsupervised training process, the training data should be easily acquired, and should include rich behavioral content and diverse set of conversations as much as possible. In this work, we collected around 400 hours of audio from 225 movies[1]. Many of the selected movies include rich and diverse affective content reflecting a range of behaviors. This training corpus is treated at the generic data set outside the target behavioral modeling domains. In previous behavior modeling corpora [29, 31, 88], there are at most 90 hours of original recording speech. Thus, compared to

[1]The list of collected movies can be found at github.com/haoqi/beh2vec

in-domain supervised behavior modeling tasks, we have a significantly larger amount of training data. We will use it to show the feasibility of proposed model within cross-domain behavioral modeling.

### 5.4.1 Evaluation Datasets

The proposed model is first tested using a clinical behavioral dataset from psychotherapy, in which conversations are characterized with clinically-relevant behavioral descriptors. Second, to evaluate the model's generalizability and domain robustness, we also test on curated corpus of several out-of-domain speech files, which contains diverse sources of speech from different scenarios such as comedy shows and debates.

| Behavior Code | Brief description |
|---|---|
| Acceptance | Indicates understanding, acceptance, respect for partner's views, feelings and behaviors |
| Blame | Blames, accuses, criticizes partner, and uses critical sarcasm and character assassinations |
| Humor | Includes jokingly making fun of self, lightly teasing the spouse, or making a reference to a mutually shared joke. |
| Negativity | Overtly expresses rejection, defensiveness, blaming, and anger |
| Positivity | Overtly expresses warmth, support, acceptance, affection, positive negotiation |

Table 5.1: Description of behavior codes in Couples Therapy Corpus

**Couple therapy dataset**

The first dataset we employ is the couples therapy corpus collected by the researchers in the UCLA/UW Couple Therapy Research Project [34], in which 134 real couples were involved in a longitudinal study of 2 years for the evaluation of complex human behaviors related to marital therapy. In each session, a relationship-related topic (*e.g.,* "Why cannot you leave my stuff alone?") was initiated and the couple had a conversation about this topic for 10 minutes.

For evaluation purposes, we employ the annotation labels. In this couple therapy corpus, each participant's behaviors were evaluated based on the Couples Interaction [63] and Social Support

| Category | ID | Brief description |
|---|---|---|
| Comedy show | 1 | George Carlin |
| | 2 | Steve Hofstetter |
| Political debate | 3 | Final Republican Presidential Debate, 2015 |
| | 4 | Vice Presidential Debate 2012 |
| TED Talk | 5 | TEDtalk: Kevin Slavin |
| | 6 | TEDtalk: Christopher Steiner |
| Eulogy | 7 | Eulogy for a Son (youtube) |
| | 8 | Mr. Li Hongyi's Eulogy for the late Mr. Lee Kuan Yew |

Table 5.2: Description of evaluation data in Diverse Speech Behavior corpora

Rating Systems [77]. The original 31 behavior codes were rated on a scale of 1-9, where 1 indicates the absence of the given behavior and 9 refers a strong presence. Similar to a previous study [14], we utilize five of the behaviors by binarizing the top and bottom 20% of the original rating scores. A brief description of the behavior codes used in this work is listed in Table 5.1.

**Curated speech data from different scenarios**

To further test the domain robustness of unsupervised behavior modeling method, we collected audio files representing a variety of other human spoken interaction domains. We manually collected audio files from two distinct speakers from four different scenarios: stand-up comedy routines, political debates, TED talks and eulogies. The audio names are listed in Table 5.2 and the duration of each audio is around 10 minutes.

## 5.5 Experimental Setup

### 5.5.1 Audio Data Preparation

For the training data, the audio files are directly extracted from movie video and combined into one single audio channel. We do not perform any pre-processing procedures (*e.g.,* VAD and diarization) on the training data. Thus, the audio frames of movie can include conversations, silence, background music, and changing of speaker regions.

For couples therapy data, since each session consists of a dyadic conversation and the behavior ratings are provided for each spouse individually, we need to diarize the interactions to obtain the speech regions for each person. We employ the pre-processing procedures described in the work [14]. In short, we select sessions with an Signal-to-noise ratio (SNR) above 5dB, and conduct Voice Activity Detection (VAD) and Speaker diarization. Speech regions from each session for the same speaker are used to analyze behaviors. The corpus has around 48 hours of audio data after these processing procedures. More details of the data processing steps can be found in [14].

### 5.5.2 Feature Extraction

We extract acoustic features, including speech prosody (pitch, intensity and their derivatives), spectral envelope characteristics (MFCCs, MFBs, LPCs and their derivatives), and voice quality (jitter, shimmer and their derivatives). These Low-Level Descriptors (LLDs) are extracted using a 25ms Hamming window with 10 ms shift. Within each analysis frame, we compute functionals of these acoustic features including Min (1st percentile), Max (99th percentile), Range (99th percentile – 1st percentile), Mean, Median, and Standard Deviation using openSMILE toolkit [46]. These features are widely used and have shown effectiveness in many affect related tasks such as speech emotion recognition [129].

The size of analysis frame for target behaviors herein are larger than other shorter duration affective states (*e.g.,* of expressed emotions which can be reliably observed within a few seconds [128], one sentence [162] or a speaker turn [21]). Previous behavioral annotation manuals [63, 66] and computational analysis [88] report that the length of observation window for target behaviors is generally around 30 seconds or even longer. Based on these studies, in this work, the analysis frame size is set to 20 seconds, the same as in previous works [86, 87, 157]. Under the feature configuration described above, for each frame window, we have a feature dimension of 420.

### 5.5.3 Model Configurations and Parameter Settings

The training pairs are from movie audio, within stationary region, the maximum sampling shift size $k$ is set to 6. For each frame $x_a$, we randomly select 4 context frames from neighboring segments as reconstruction frames $x_p$. While the frame pair $(x_n, x_{n\_p})$ is randomly selected from one stationary neighboring window in a different movie.

In our experiment, the encoder-decoder structure of DCN and TE-DCN contains six hidden layers connected by PReLU [61] activation function. The dimension of the hidden layers are 300, 200, 64, 200, 300 respectively. The output of bottleneck embedding layer with 64 dimensions is regarded as behavior related representation that we are interested in. We use the Euclidean distance as the distance metric $D(\cdot,\cdot)$ in Equation (5.2). The model is trained with the Adam optimizer [80] using a learning rate of 0.001 and a decay of 0.1 every 10 epochs. The triplet loss is optimized with a margin of $m$=2 and regularization weight of $\lambda$=0.01. We utilize different movie pairs as the validation set to terminate training with early stopping.

### 5.5.4 Evaluation Method

#### 5.5.4.1 Evaluation Method for In-domain Couples Therapy Corpus

Considering the inter-annotator agreement, we binarize the original behavior ratings to model the evaluation task as a binary classification task of low- and high- presence of each behavior as in [14]. For each behavior code and each gender, we selected 70 sessions on one extreme of the code (e.g., high blame) and 70 sessions at the other extreme (e.g., low blame). This also enables balancing for each behavior resulting in classes of equal size. As mentioned in section 5.4, the couples therapy corpus only has session-level behavior code ratings. With these session-level labels, we evaluate the model in a supervised manner, though the behavior representation is trained in an unsupervised way with an out-of-domain movie corpus.

For each frame, once we obtain the latent behavioral manifold representation, we use the k-nearest neighbors algorithm to find a "reference label". In our case we choose $k$=1 and use Euclidean distance to find the nearest frame among all remaining labeled frames from different sessions. In addition, we also ensure that speaker characteristics information is not involved during testing by using leave-one-couple-out cross validation. Finally, majority voting is employed to generate session-level binary labels from multiple frame-level labels.

#### 5.5.4.2 Evaluation Method for Diverse Speech Behavior Corpora

This evaluation is targeted to reflect different behavior or scenario styles. For example, as listed in Table 5.2, the behavioral style from a stand-up comedy show is expected to be similar across performers, but expected to be different from those in a speech during a eulogy. Instead of focusing on scenario classification of whole speech regions, we are interested in the level of similarity across different scenarios. With this expectation, we calculate the results obtained by frame clustering with nearest neighbor, *i.e.,* which frame is close to which, as a percentage. This percentage score can be regarded as an indicator of style similarity among audio frames.

## 5.6 Experimental Results and Discussions

### 5.6.1 Experiment Results of Couple Therapy Corpus

The performance of couples' behavior classification results across different models is shown in Table 5.3. Besides the DCN and TE-DCN models, we further compare the results with four other models.

#### Baseline Model

For each behavior code, the number of behavior presence and absence sessions are balanced. Thus, a weak baseline of classification accuracy is 50%. In this work, we use a better baseline model, which is built through the nearest neighbor classification in the original acoustic feature space.

| Behavior | Baseline | DCN | Triplet network | TE-auto-encoder | TE-DCN | Supervised training in [88] |
|---|---|---|---|---|---|---|
| Acceptance | 57.14 | 66.43 | 60.71 | 65.71 | **68.21** | 72.50 |
| Blame | 55.00 | 61.07 | 63.21 | 61.43 | **64.64** | 71.79 |
| Humor | 54.29 | 55.00 | 56.79 | **60.36** | **60.36** | - |
| Negativity | 63.92 | 63.93 | 61.79 | 60.71 | **66.43** | 76.07 |
| Positivity | 50.71 | 65.00 | 58.57 | 61.43 | **65.35** | 65.36 |
| Average | 56.212 | 62.286 | 60.214 | 61.928 | **64.998** | 71.43 |

Table 5.3: Classification accuracy (%) of behavior codes in Couple Therapy Corpus

Similarly, the session-level label is obtained by majority voting. The average classification accuracy of five behavior codes is 56.212%, which is slightly better than the weak baseline (random guess). These results indicate that further representation learning process is necessary to extract behavior information from high dimensional acoustic features [87].

**DCN Model**

In Table 5.3, for all behavior codes, the DCN model outperforms the the baseline and achieves an average classification accuracy of 62.29%. With the McNemaar test, compared with the baseline, the results are statistically significant with $p < 0.01$. Further details of the DCN model can be found in our previous work [87]. These preliminary results verify the possibility of using out-of-domain data for low-resource domain behavior modeling. Through it is not guaranteed that the extracted representations remove all other nuisance factors and only contain target behavior information, the results from the DCN model show that affect related information are captured in the proposed manifold space.

**Triplet Network Model**

As a comparison, we also perform the experiment with the triplet network model. Different from reconstruction of neighboring frames in DCN model, the triplet model only uses discriminative distance metric to directly optimize the representations within behavioral manifold. Compared

with the TE-DCN, the model does not contain the decoder parts. Thus, the contextual reconstruction loss from the decoder is not considered during the training and we only optimize the triplet loss from the outputs of encoders.

The experiment is conducted with similar settings as before, and we observe that the triplet model outperforms the baseline with average classification accuracy of 60.21%. We notice that, for most behaviors, the DCN model achieves slightly better performance than this triplet model. In addition, we also tried negative sampling strategies [65, 122] in the selection of triplet pair during training, however, we find that there is no improvement in terms of the domain data classification accuracy.

Considering the complexity of the training data, one reasonable explanation of the lower average performance of the triplet model might be the importance of the "generative" property of decoder. The representation of the behavioral manifold is trained to have the ability of encompassing and reconstructing the acoustic features of its neighbor frames, which are highly related to affect related information. The triplet network is only trained to discriminate samples with distance metric. Such a model might be failing to ensure that the optimized embeddings are highly relevant to capturing behavioral information, resulting a lower performance on the behavior modeling tasks.

**TE-autoencoder Model**

Further, we test the TE-autoencoder model, a variant of the TE-DCN model. In TE-autoencoder model, we replace the TE-DCN's contextual encoder-decoder structure with an autoencoder. Thus, once we have the training input pair $(x_a, x_p, x_n)$, the corresponding reconstruction pair is $(x_a, x_p, x_n)$ rather than previous context-based $(x_p, x_a, x_{n\_p})$. The autoencoder is used to compress the original acoustic features and obtain representations with a same reduced dimension. Under this setting, the model can preserve the property of feature compression while ignoring the contextual information. Similarly, this behavioral representation is optimized through both reconstruction loss and triplet loss in the target manifold. We find that the results of the average performance of TE-autoencoder

is worse than TE-DCN's. This further supports the importance of contextual information, and also validates the behavior stationarity assumption.

### TE-DCN Model

The TE-DCN is built upon DCN, and the extracted behavior representation is enhanced by the discriminative metric under the behavior stationarity assumption. From the classification results, we can observe that there is an improvement, from the 56.21% of baseline to 64.99% of TE-DCN model in terms of the averaged classification accuracy. Under the McNemar test, these results of proposed TE-DCN are statistically significant with $p < 0.01$. The TE-DCN model shows best performance across all models. In addition, compared with both DCN and triplet models, for all five behavior codes, a consistent improvement is obtained.

Moreover, we notice the complementary nature of DCN and triplet models in behavior modeling. By combining these two, TE-DCN shows that both metric learning and context information can contribute to the overall unsupervised behavior modeling performance. These results are encouraging considering only unsupervised approaches are utilized with unlabeled, out-of-domain data in TE-DCN.

### Supervised Training Method

The last column of the table indicates the classification results generated from a context-aware model via utilizing emotion related representation as behavioral primitives to facilitate the behavior quantification. Details of this supervised training approach can be found in [88].

These supervised classification results can be regarded as an upper bound performance of the supervised versus the unsupervised methods. Moreover, it is necessary to mention that due to the complexity of human behavior and the subjectivity in annotation process, even for human annotators, the inter-annotator agreement can only reach about Krippendorff's $\alpha = 0.8$ [148]. Thus,

although worse than the supervised method, the TE-DCN's performance is encouraging considering the fact that classification is obtained by a completely unsupervised method with simple majority vote.

### 5.6.2 Behavioral Trajectory Analysis

In scenarios such as psychotherapy, instead of obtaining session-level classification labels, domain experts might be more interested in dynamic behavior change trajectories. These trajectories can help the psychologists quickly locate the most salient regions and potentially reduce the workload of manual annotation. In this subsection, we use the couple therapy corpus as an example to illustrate that our unsupervised behavior modeling method can potentially provide such behavioral trajectories.

Suppose we use the labeled frame samples as reference, and select the top $N$ nearest samples in the behavior manifold space. Among the top $N$ reference frames, we can calculate the percentage of samples labeled with the presence of a certain behavior code label (samples with label 1 in our case). For each test frame, the percentage value can indirectly imply the behavior ratings at some level. Figure 5.3 shows an example with one sample session's behavior dynamic change trajectories among five behaviors, and we set $N = 60$ in this case. In Table 5.4, we provide the original averaged human annotation ratings and the automatically assigned behavior classification labels of this session.

| Behavior | Binarized label (0:absence; 1: presence) | Manual rating (ranging from 1-9) |
|---|---|---|
| Acceptance | 0 | 2.33 |
| Blame | 1 | 7.66 |
| Humor | 0 | 1.0 |
| Negativity | 1 | 6.25 |
| Positivity | 0 | 1.5 |

Table 5.4: Original annotation ratings and binarized classification labels for each behavior code

Although the corpus does not provide utterance- or frame- level annotations, from this figure, we can notice the correlations among different predicted behavior code ratings. We observe that

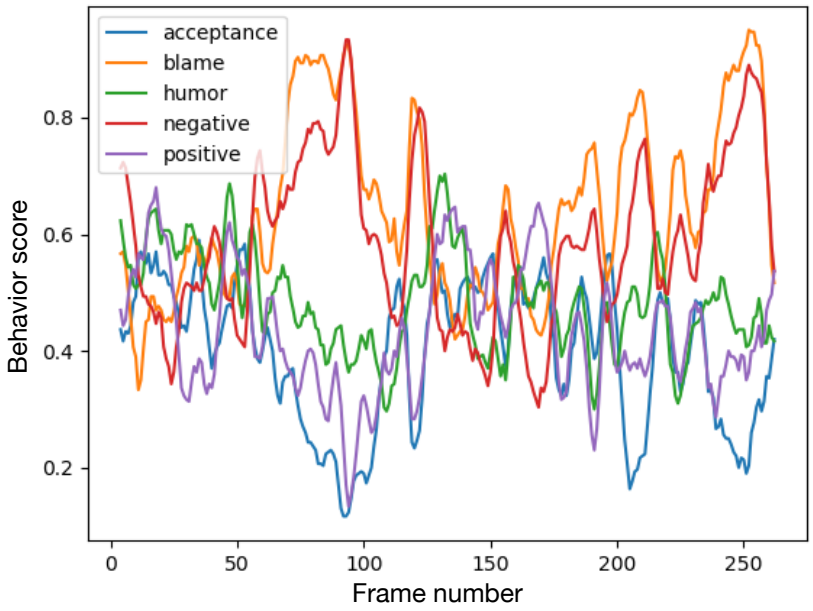



Figure 5.3: One sample session with five behavior score trajectories

behaviors Blame and Negativity are highly correlated, and behavior Positivity, Acceptance and Humor tend to have a similar trend. In addition, "positive" related and "negative" related behaviors have the opposite trend, which is in agreement with our intuition and previous supervised modeling research work [14, 86]. From the plot, we can also observe this session shows more presence of "negative" behaviors (with higher scores) and less degree of "positive" behaviors (with lower scores), which is in agreement with the human ratings listed in Table 5.4.

In real world scenarios, it is often the case that the amount of annotated data might not be adequate to train a supervised behavior recognition system well. Through our unsupervised behavior modeling approach, if we need to annotate a newly collected session, this behavioral trajectory can quickly indicate salient behavior regions and help domain experts to locate and annotate the corresponding regions efficiently.

### 5.6.3 Experiment Results on Diverse Speech Behavior Corpora

In this subsection, we use collected out-of-BSP domain data to evaluate the generalizability of TE-DCN model. As listed in Table 5.2, We collect two audio files from different speakers for each

category. The results of similarity evaluation among different scenarios is shown in Figure 5.4. As described in Section 5.5.4, in this table, each entry is calculated by dividing the number of nearest frames in each selected file by the total number of frames in the input audio. This normalized percentage value is used to evaluate the behavior similarity.

| Selected | | Comedy | | Debate | | Ted talk | | Eulogy | |
|---|---|---|---|---|---|---|---|---|---|
| Input | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Comedy | 1 | 0.00 | **0.52** | 0.06 | 0.06 | 0.06 | 0.24 | 0.04 | 0.02 |
| | 2 | **0.40** | 0.00 | 0.07 | 0.08 | 0.16 | 0.12 | 0.10 | 0.07 |
| Debate | 3 | 0.24 | 0.08 | 0.00 | 0.10 | **0.33** | 0.10 | 0.12 | 0.04 |
| | 4 | 0.08 | 0.10 | 0.06 | 0.00 | **0.27** | 0.18 | 0.24 | 0.06 |
| Ted talk | 5 | 0.06 | 0.11 | 0.17 | 0.11 | 0.00 | **0.32** | 0.01 | 0.09 |
| | 6 | 0.12 | 0.12 | 0.08 | 0.12 | **0.34** | 0.00 | 0.15 | 0.07 |
| Eulogy | 7 | 0.07 | 0.13 | 0.13 | 0.10 | 0.17 | 0.19 | 0.00 | **0.22** |
| | 8 | 0.05 | 0.13 | 0.02 | 0.08 | 0.18 | 0.21 | **0.32** | 0.00 |

Figure 5.4: Confusion matrix of behavior scenario similarity evaluation

From the similarity confusion matrix, we observe that for comedy, TED talk and eulogy categories, audio files exhibit high similarity scores within same category, and have lower scores for less related scenarios, as we expected. However, for the debate files, they are mostly confused with the ted talk files while also showing large similarity values with other scenarios. The reason for this might be the fact that during the debate, different politicians employed different kinds of debate skills and behaviors vary among different situations and topics. In general, we find eight out of ten files are classified correctly based on the majority vote on frame-level clustering. Moreover, from the results table, we also can observe the similarity under different degrees among the different scenarios considered. These promising results underscore the domain robustness and applications of the proposed unsupervised behavior modeling.

### 5.6.4 Nuisance Factors and Selection of Features

The TE-DCN model tends to preserve the shared behavioral information between the input frame and its neighboring frames. We acknowledge that the design of the model combined with the behavioral stationarity assumption may have a potential complication: the neighboring frames could also encode speaker characteristics as well as acoustic conditions such as of the environment and channel conditions.

To minimize the effect of these nuisance factors, the choice of input feature is critical in our proposed model. In addition to the triplet loss, the input features are designed to ensure that the unsupervised behavior modeling focuses more on affect related aspects rather than only employing the contextual information itself. As described in section 5.5.2, we directly use affect related hand-crafted features as input rather than extracting intermediate representations from raw spectrum features (*e.g.,* MFCC or MFB coefficients). We further replace the encoder-decoder structure of TE-DCN with CNN layers to input lower level raw spectrum features directly. Based on the experiments, we find it is still challenging to extract behavioral representation exclusively, if inputs are lower level raw spectrum features, which largely contain other acoustically encoded information.

## 5.7 Conclusion

The availability of adequate labelled data has been a critical bottleneck for supervised behavior modeling. Obtaining relevant behavioral data for such modeling often suffers from not only expensive data collection but varied and low human inter-annotation agreements. These constraints not only impact the modeling performance, but also limit the generalizability of the obtained behavioral models across domains.

In this work, we explore unsupervised learning for computational behavior modeling. We propose the TE-DCN model to extract behavioral representations in an unsupervised way. The results suggest that the reconstruction with context information and metric learning are complementary

methods within unsupervised behavior modeling. As a case study of unsupervised behavior modeling from speech using couples therapy data, our framework is shown to extract target behaviors from audio signals and achieves promising behavioral quantification results. Although there is scope for improvement compared with the supervised method, our work provides possible solutions for the computational human behavior modeling: transfer information from out-of-domain data which are easily obtainable, and then adapt the model to specific domain applications. We also note that information encoded in the speech unrelated to the target behaviors being modeled can cause negative effects in the representation learning.

In the future, we plan to further computationally disentangle and reduce the speaker characteristics and other complex acoustic nuisance factors in the behavior representation. We plan to consider adversarial training to obtain more speaker-invariant and environment-robust behavior representations [89]. Moreover, we also plan to investigate the feasibility of representation adaptation for downstream tasks by adding additional domain-specific supervised tuning.

# Chapter 6

# Conclusion and Future Work

## 6.1 Summary of Research

This dissertation explored the computational approaches of human behavior quantification via emerging deep learning techniques under certain constraints. Meanwhile, we tried to capture dynamics and relevant important behavioral representations directly from speech signals.

Through our work, first, we presented a neural network model on the binary classification of behaviors from distressed couples in therapy using acoustic features. The proposed SD-DNN framework employs the multi-stage training and limits the number of training parameters at all stages to prevent overfitting and achieves well convergences. As shown in Chapter 2, this method shows the promising benefits of introducing DNN into the behavior signal processing domain. The results achieved better in behavior classification when compared to traditional machine learning approaches, such as SVM. Even though our model did not specifically optimize the dynamic behavior changes within the session, it provided behavioral trajectories within the session, and the classification results are better than existing HMM dynamic models.

Meanwhile, we noticed the limitation of data resources in the utilization of large-scale data-driven machine learning approaches on BSP applications. Thus, in the second part of the research, we investigated the possibility of employment of out-of-domain data to facilitate the domain behavior understanding.

In Chapter 3, we analyzed the link between emotions and behaviors through transfer learning. Through this, we experimentally verified the existing relationship and used emotional information as constructs of behavioral understanding. The importance of temporal dynamics and the context information of short-term affect states in shaping behaviors was also addressed in our experiment results. The results suggested different behaviors require different observation window length to be quantified. Moreover, we found different behavior exhibits different characters. Some are more closely related to emotions, while some are more related to speech signals or lexical descriptions. Most importantly, most of our experimental findings are congruent with existing research work [25, 63, 77] from psychology or social science. The importance of emotion involved behavior primitive motived the working of investigating more robust emotion representations. Human speech signals encompass rich information such as linguistic features, speaker characteristics, and affective states. In Chapter 4, we explored the disentanglement of affective information from speech representations using adversarial training. Specifically, we eliminated speaker-related information and obtain a speaker-invariant affect embedding from speech. In Chapter 5, we further reduced restrictions of out of domain data utilization. We exploited the slow varying properties of human behavior and proposed a deep contextualized encoder-decoder structure to connect behavioral context and derived the behavioral manifold in an unsupervised manner. By introducing the idea of metric learning into our model, we proposed the triplet enhanced contextualized networks. The results are extremely encouraging and promise improved behavioral quantification in an unsupervised manner.

## 6.2 Future Work

There are several potential research directions with emerging challenges in the study of behavior quantification and understanding. As the well-being of mental health is gradually attracting considerable interest, the application of BSP is extending from the assistance of treatment in psychotherapy to the facilitation of well-being of people in daily life. Recently, many wearable commercial

products, such as Apple Watch and Amazon Halo, make it easier to collect more behavior signals (e.g. speech or heart rate signals) with fewer efforts. The evolution of hardware and emphasis on mental health potentially produce many opportunities in BSP areas. However, data collected in real-life scenarios is usually much more complex than data collected under well-designed experimental settings. For example, it may include conversations ranging from different topics and speech from multiple speakers. Thus, how to adapt the existing BSP techniques to complex daily life scenarios usage can be an interesting and challenging research topic to further investigate. In addition, since human behavioral information is encoded in multiple modalities, the relationship and link across different modalities, the interpretation of dynamic changes of behavior trajectories can be explored.

# Bibliography


[1] Mohammed Abdelwahab and Carlos Busso. Domain adversarial for acoustic emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(12):2423–2435, 2018.

[2] Zakaria Aldeneh and Emily Mower Provost. Using regional saliency for speech emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2741–2745. IEEE, 2017.

[3] Nalini Ambady and Robert Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological bulletin*, 111(2):256, 1992.

[4] Paul C Amrhein, William R Miller, Carolina E Yahne, Michael Palmer, and Laura Fulcher. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of consulting and clinical psychology*, 71(5):862, 2003.

[5] Namrata Anand and Prateek Verma. Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data. In *Technical Report*. Stanford University, 2015.

[6] John S Baer, Elizabeth A Wells, David B Rosengren, Bryan Hartzler, Blair Beadnell, and Chris Dunn. Agency context and tailored training in technology transfer: A pilot evaluation of motivational interviewing training for community counselors. *Journal of substance abuse treatment*, 37(2):191–202, 2009.

[7] Brian R Baucom, David C Atkins, Lorelei E Simpson, and Andrew Christensen. Prediction of response to treatment in a randomized clinical trial of couple therapy: a 2-year follow-up. *Journal of Consulting and Clinical Psychology*, 77(1):160, 2009.

[8] Roy F Baumeister, Kathleen D Vohs, C Nathan DeWall, and Liqing Zhang. How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and social psychology review*, 11(2):167–203, 2007.

[9] Roy F Baumeister, C Nathan DeWall, Kathleen D Vohs, and Jessica L Alquist. Does emotion cause behavior (apart from making people do stupid, destructive things). *Then a miracle occurs: Focusing on behavior in social psychological theory and research*, pages 12–27, 2010.

[10] R Beale and C Peter. *Affect and emotion in human-computer interaction*. Springer, 2008.

[11] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. *ICML Unsupervised and Transfer Learning*, 27:17–36, 2012.

[12] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35 (8):1798–1828, 2013.

[13] Matthew Black, Athanasios Katsamanis, Chi-Chun Lee, Adam C Lammert, Brian R Baucom, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. Automatic classification of married couples' behavior using audio features. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[14] Matthew P Black, Athanasios Katsamanis, Brian R Baucom, Chi-Chun Lee, Adam C Lammert, Andrew Christensen, Panayiotis G Georgiou, and Shrikanth S Narayanan. Toward automating a human behavioral coding system for married couples' interactions using speech acoustic features. *Speech communication*, 55(1):1–21, 2013.

[15] Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5 (9/10):341–345, 2002.

[16] Daniel Bone, Chi-Chun Lee, Theodora Chaspari, James Gibson, and Shrikanth Narayanan. Signal processing and machine learning for mental health research and clinical applications. *IEEE Signal Processing Magazine*, 34(5):189–196, September 2017.

[17] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[18] Craig J Bryan, M David Rudd, Evelyn Wertenberger, Neysa Etienne, Bobbie N Ray-Sannerud, Chad E Morrow, Alan L Peterson, and Stacey Young-McCaughon. Improving the detection and prediction of suicidal behavior among military personnel by measuring suicidal beliefs: An evaluation of the suicide cognitions scale. *Journal of affective disorders*, 159:15–22, 2014.

[19] Bethany A Burum and Marvin R Goldfried. The centrality of emotion to psychological change. *Clinical Psychology: Science and Practice*, 14(4):407–413, 2007.

[20] Carlos Busso and Shrikanth S Narayanan. The expression and perception of emotions: Comparing assessments of self versus others. In *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[21] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[22] Carlos Busso, Angeliki Metallinou, and Shrikanth S Narayanan. Iterative feature normalization for emotional speech detection. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011.

[23] Michel Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.

[24] Erik Cambria. Affective computing and sentiment analysis. *IEEE intelligent systems*, 31 (2):102–107, 2016.

[25] Dana R Carney, C Randall Colvin, and Judith A Hall. A thin slice perspective on the accuracy of first impressions. *Journal of Research in Personality*, 41(5):1054–1072, 2007.

[26] Facundo Carrillo, Natalia Mota, Mauro Copelli, Sidarta Ribeiro, Mariano Sigman, Guillermo Cecchi, and Diego Fernandez Slezak. Emotional intensity analysis in bipolar subjects. *arXiv preprint arXiv:1606.02231*, 2016.

[27] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[28] S. N. Chakravarthula, R. Gupta, B. Baucom, and P. Georgiou. A language-based generative model framework for behavioral analysis of couples' therapy. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 2090–2094, April 2015.

[29] Sandeep Nallan Chakravarthula, Haoqi Li, Shao-Yen Tseng, Maija Reblin, and Panayiotis Georgiou. Predicting behavior in cancer-afflicted patient and spouse interactions using speech and language. *arXiv preprint arXiv:1908.00908*, 2019.

[30] Sandeep Nallan Chakravarthula, Brian RW Baucom, Shrikanth Narayanan, and Panayiotis Georgiou. An analysis of observation length requirements for machine understanding of human behaviors from spoken language. *Computer Speech & Language*, page 101162, 2020.

[31] Sandeep Nallan Chakravarthula, Md Nasir, Shao-Yen Tseng, Haoqi Li, Tae Jin Park, Brian Baucom, Craig J Bryan, Shrikanth Narayanan, and Panayiotis Georgiou. Automatic prediction of suicidal risk in military couples using multimodal interaction cues from couples conversations. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6539–6543. IEEE, 2020.

[32] Jonathan Chang and Stefan Scherer. Learning representations of emotional speech with deep convolutional generative adversarial networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[33] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[34] Andrew Christensen, David C Atkins, Sara Berns, Jennifer Wheeler, Donald H Baucom, and Lorelei E Simpson. Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples. *Journal of consulting and clinical psychology*, 72(2):176, 2004.

[35] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[36] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. An unsupervised autoregressive model for speech representation learning. *Proc. Interspeech 2019*, pages 146–150, 2019.

[37] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

[38] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 18(1):32–80, 2001.

[39] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

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

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

[42] Sally Dunlop, Melanie Wakefield, and Yoshi Kashima. Can you feel it? negative emotion, risk, and narrative in health communication. *Media Psychology*, 11(1):52–75, 2008.

[43] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[44] Paul Ekman. Are there basic emotions? 1992.

[45] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[46] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010.

[47] Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2404-5.

[48] Mark E Feinberg, Marni L Kan, and E Mavis Hetherington. The longitudinal influence of coparenting conflict on parental negativity and adolescent maladjustment. *Journal of Marriage and Family*, 69(3):687–702, 2007.

[49] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[50] Panayiotis G. Georgiou, Matthew P. Black, Adam C. Lammert, Brian R. Baucom, and Shrikanth S. Narayanan. "that's aggravating, very aggravating": Is it possible to classify behaviors in couple interactions using automatically derived lexical features? In Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin, editors, *Affective Computing and Intelligent Interaction*, pages 87–96, Berlin, Heidelberg, October 2011. Springer Berlin Heidelberg.

[51] Panayiotis G. Georgiou, Matthew P. Black, and Shrikanth S. Narayanan. Behavioral signal processing for understanding (distressed) dyadic interactions: Some recent developments. In *Proceedings of the 2011 Joint ACM Workshop on Human Gesture and Behavior Understanding*, J-HGBU '11, pages 7–12, New York, NY, USA, 2011. ACM, ACM. ISBN 978-1-4503-0998-1.

[52] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal, and Sanjeev Khudanpur. A pitch extraction algorithm tuned for automatic speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 2494–2498. IEEE, 2014.

[53] Prasanta Kumar Ghosh, Andreas Tsiartas, and Shrikanth Narayanan. Robust voice activity detection using long-term signal variability. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(3):600–613, 2011.

[54] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Representation learning for speech emotion recognition. In *Interspeech*, pages 3603–3607, 2016.

[55] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, and Emily Mower Provost. Progressive neural networks for transfer learning in emotion recognition. *Proc. Interspeech 2017*, pages 1098–1102, 2017.

[56] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[57] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[58] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[59] Rahul Gupta, Nikolaos Malandrakis, Bo Xiao, Tanaya Guha, Maarten Van Segbroeck, Matthew Black, Alexandros Potamianos, and Shrikanth Narayanan. Multimodal prediction of affective dimensions and depression in human-computer interactions. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, pages 33–40. ACM, 2014.

[60] Kun Han, Dong Yu, and Ivan Tashev. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech*, pages 223–227, 2014.

[61] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[62] Marti A. Hearst, Susan T Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[63] C Heavey, D Gill, and A Christensen. Couples interaction rating system 2 (cirs2). *University of California, Los Angeles*, 7, 2002.

[64] Christopher L Heavey, Andrew Christensen, and Neil M Malamuth. The longitudinal impact of demand and withdrawal during marital conflict. *Journal of consulting and clinical psychology*, 63(5):797, 1995.

[65] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[66] Richard E Heyman. Rapid marital interaction coding system (rmics). In *Couple observational coding systems*, pages 81–108. Routledge, 2004.

[67] Richard E Heyman, Bushra R Chaudhry, Dominique Treboux, Judith Crowell, Chiyoko Lord, Dina Vivian, and Everett B Waters. How much observational data is enough? an empirical test using marital interaction coding. *Behavior therapy*, 32(1):107–122, 2001.

[68] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, Nov 2012. ISSN 1053-5888. doi: 10.1109/MSP.2012.2205597.

[69] Erika Hoff. Language development at an early age: Learning mechanisms and outcomes from birth to five years. *Encyclopedia on early childhood development*, pages 1–5, 2009.

[70] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[71] Hao Hu, Ming-Xing Xu, and Wei Wu. Gmm supervector based svm with spectral features for speech emotion recognition. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–413. IEEE, 2007.

[72] Che-Wei Huang and Shrikanth Shri Narayanan. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 583–588. IEEE, 2017.

[73] Che-Wei Huang, Shrikanth Narayanan, et al. Characterizing types of convolution in deep convolutional recurrent neural networks for robust speech emotion recognition. *arXiv preprint arXiv:1706.02901*, 2017.

[74] Jian Huang, Ya Li, Jianhua Tao, Zhen Lian, et al. Speech emotion recognition from variable-length inputs with triplet loss function. In *Interspeech*, pages 3673–3677, 2018.

[75] Zac E Imel, Jacqueline S Barco, Halley J Brown, Brian R Baucom, John S Baer, John C Kircher, and David C Atkins. The association of therapist empathy and synchrony in vocally encoded arousal. *Journal of counseling psychology*, 61(1):146, 2014.

[76] Arindam Jati and Panayiotis Georgiou. Neural predictive coding using convolutional neural networks toward unsupervised learning of speaker characteristics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(10):1577–1589, 2019.

[77] J Jones and A Christensen. Couples interaction study: Social support interaction rating system. *University of California, Los Angeles*, 7, 1998.

[78] Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and S Narayanan. Sailalign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[79] Soheil Khorram, Mimansa Jaiswal, John Gideon, Melvin McInnis, and Emily-Mower Provost. The priori emotion dataset: Linking mood to emotion detected in-the-wild. *Proc. Interspeech 2018*, pages 1903–1907, 2018.

[80] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[81] Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Junaid Qadir, and Björn W Schuller. Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378*, 2020.

[82] D. Le and E. M. Provost. Emotion recognition from spontaneous speech using hidden markov models with deep belief networks. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 216–221, Dec 2013.

[83] Chi-Chun Lee, Athanasios Katsamanis, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Based on isolated saliency or causal integration? toward a better understanding of human annotation process using multiple instance learning and sequential probability ratio test. In *Proceedings of InterSpeech*, 2012.

[84] Chul Min Lee and Shrikanth S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 13(2):293–303, mar 2005.

[85] Jinkyu Lee and Ivan Tashev. High-level feature representation using recurrent neural network for speech emotion recognition. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[86] Haoqi Li, Brian Baucom, and Panayiotis Georgiou. Sparsely connected and disjointly trained deep neural networks for low resource behavioral annotation: Acoustic classification in couples' therapy. In *Proceedings of Interspeech*, San Francisco, CA, September 2016.

[87] Haoqi Li, Brian Baucom, and Panayiotis Georgiou. Unsupervised latent behavior manifold learning from acoustic features: Audio2behavior. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5620–5624. IEEE, 2017.

[88] Haoqi Li, Brian Baucom, and Panayiotis Georgiou. Linking emotions to behaviors through deep transfer learning. *PeerJ Computer Science*, 6:e246, 2020.

[89] Haoqi Li, Ming Tu, Jing Huang, Shrikanth Narayanan, and Panayiotis Georgiou. Speaker-invariant affective representation learning via adversarial training. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7144–7148. IEEE, 2020.

[90] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 312–317. IEEE, 2013.

[91] Wootaek Lim, Daeyoung Jang, and Taejin Lee. Speech emotion recognition using convolutional and recurrent neural networks. In *Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific*, pages 1–4. IEEE, 2016.

[92] Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In *International Conference on Machine Learning*, pages 4013–4022, 2019.

[93] Samuel D Lustgarten. Emerging ethical threats to client privacy in cloud communication and data storage. *Professional Psychology: Research and Practice*, 46(3):154, 2015.

[94] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. *IEEE Transactions on Multimedia*, 16(8):2203–2213, 2014.

[95] Gayla Margolin, Pamella H Oliver, Elana B Gordis, Holly Garcia O'hearn, Anna Marie Medina, Chandra M Ghosh, and Leslie Morland. The nuts and bolts of behavioral observation of marital and family interaction. *Clinical child and family psychology review*, 1(4): 195–213, 1998.

[96] Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gang, and Biing-Hwang Juang. Speaker-invariant training via adversarial learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[97] A. Metallinou, R. B. Grossman, and S. Narayanan. Quantifying atypicality in affective facial expressions of children with autism spectrum disorders. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6, July 2013.

[98] Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Bjorn Schuller, and Shrikanth Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Transactions on Affective Computing*, 3(2):184–198, 2012.

[99] T Mikolov and J Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 2013.

[100] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[101] Benjamin Milde and Chris Biemann. Unspeech: Unsupervised speech context embeddings. *Proc. Interspeech 2018*, pages 2693–2697, 2018.

[102] William R Miller, R Gayle Benefield, and J Scott Tonigan. Enhancing motivation for change in problem drinking: a controlled comparison of two therapist styles. *Journal of consulting and clinical psychology*, 61(3):455, 1993.

[103] Mohammad Hossein Moattar and Mohammad M Homayounpour. A review on speaker diarization systems and approaches. *Speech Communication*, 54(10):1065–1103, 2012.

[104] E. Mower and S. Narayanan. A hierarchical static-dynamic framework for emotion classification. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2372–2375. IEEE, May 2011.

[105] Shrikanth Narayanan and Panayiotis G Georgiou. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, 101(5): 1203–1233, 2013.

[106] Md Nasir, Arindam Jati, Prashanth Gurunath Shivakumar, Sandeep Nallan Chakravarthula, and Panayiotis Georgiou. Multimodal and multiresolution depression detection from speech and facial landmark features. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 43–50. ACM, 2016.

[107] Md Nasir, B. R. Baucom, Craig J Bryan, Shrikanth Narayanan, and Panayiotis Georgiou. Complexity in speech and its relation to emotional bond in therapist-patient interactions during suicide risk assessment interviews. In *Interspeech*, Stockholm, Sweden, August 2017.

[108] Md Nasir, Brian Robert Baucom, Panayiotis Georgiou, and Shrikanth Narayanan. Predicting couple therapy outcomes based on speech acoustic features. *PloS one*, 12(9):e0185123, 2017.

[109] Md Nasir, Brian Baucom, Shrikanth Narayanan, and Panayiotis Georgiou. Towards an unsupervised entrainment distance in conversational speech using deep neural networks. In *Interspeech / arXiv:1804.08782*, 2018.

[110] Michael Neumann and Ngoc Thang Vu. Improving speech emotion recognition with unsupervised representation learning on unlabeled speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

[111] Keith Oatley and Jennifer M Jenkins. *Understanding emotions.* Blackwell publishing, 1996.

[112] Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte, and Yoshua Bengio. Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*, 2019.

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

[114] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society, 2011.

[115] Lawrence Rabiner and B Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

[116] Saurabh Sahu, Rahul Gupta, and Carol Espy-Wilson. On enhancing speech emotion recognition using generative adversarial networks. *arXiv preprint arXiv:1806.06626*, 2018.

[117] Saurabh Sahu, Vikramjit Mitra, Nadee Seneviratne, and Carol Espy-Wilson. Multi-modal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription. *Proc. Interspeech 2019*, pages 3302–3306, 2019.

[118] David Sander and Klaus Scherer. *Oxford companion to emotion and the affective sciences*. OUP Oxford, 2014.

[119] D.L. Schacter, D. T. Gilbert, and D. M. Wegner. *Psychology (2nd Edition)*. Worth, New York, 2011.

[120] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005.

[121] Harold Schlosberg. Three dimensions of emotion. *Psychological review*, 61(2):81, 1954.

[122] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[123] Björn Schuller. *Advances in Neural Networks: Computational and Theoretical Issues*, chapter Deep Learning Our Everyday Emotions, pages 339–346. Springer International Publishing, Cham, 2015. ISBN 978-3-319-18164-6.

[124] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Hidden markov model-based speech emotion recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 2, pages II–1. IEEE, 2003.

[125] Björn Schuller, Gerhard Rigoll, and Manfred Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 1, pages I–577. IEEE, 2004.

[126] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[127] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9-10):1062–1087, 2011.

[128] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic. Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456, 2012.

[129] Björn W Schuller. Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018.

[130] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, pages 2503–2511, 2015.

[131] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. Speaker normalisation for speech-based emotion detection. In *15th international conference on digital signal processing*, pages 611–614. IEEE, 2007.

[132] Yusuke Shinohara. Adversarial multi-task learning of deep neural networks for robust speech recognition. In *INTERSPEECH*. San Francisco, CA, USA, 2016.

[133] Nelson H Soken and Anne D Pick. Infants' perception of dynamic affective expressions: Do infants distinguish specific expressions? *Child development*, 70(6):1275–1282, 1999.

[134] Hagen Soltau, Hank Liao, and Haşim Sak. Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition. *Proc. Interspeech 2017*, pages 3707–3711, 2017.

[135] Huan Song, Megan Willi, Jayaraman J Thiagarajan, Visar Berisha, and Andreas Spanias. Triplet network with attention for speaker diarization. *arXiv preprint arXiv:1808.01535*, 2018.

[136] Paul E Spector and Suzy Fox. An emotion-centered model of voluntary work behavior: Some parallels between counterproductive work behavior and organizational citizenship behavior. *Human resource management review*, 12(2):269–292, 2002.

[137] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014. ISSN 1532-4435.

[138] Brian Stasak, Julien Epps, Nicholas Cummins, and Roland Goecke. An investigation of emotional speech in depression classification. In *Interspeech*, pages 485–489, 2016.

[139] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5688–5691, May 2011. doi: 10.1109/ICASSP.2011.5947651.

[140] Sining Sun, Binbin Zhang, Lei Xie, and Yanning Zhang. An unsupervised deep domain adaptation approach for robust speech recognition. *Neurocomputing*, 257:79–87, 2017.

[141] Marco Tagliasacchi, Beat Gfeller, Félix de Chaumont Quitry, and Dominik Roblek. Self-supervised audio representation learning for mobile devices. *arXiv preprint arXiv:1905.11796*, 2019.

[142] Toshiko Tanaka, Takao Yamamoto, and Masahiko Haruno. Brain response patterns to economic inequity predict present and future depression indices. *Nature Human Behaviour*, 1 (10):748, 2017.

[143] Dengke Tang, Junlin Zeng, and Ming Li. An end-to-end deep learning framework for speech emotion recognition of atypical individuals. In *Interspeech*, pages 162–166, 2018.

[144] Jianhua Tao and Tieniu Tan. Affective computing: A review. In *International Conference on Affective computing and intelligent interaction*, pages 981–995. Springer, 2005.

[145] Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. Combating human trafficking with multimodal deep models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1547–1556, 2017.

[146] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI Global, 2010.

[147] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5200–5204. IEEE, 2016.

[148] Shao-Yen Tseng, Sandeep Nallan Chakravarthula, Brian R Baucom, and Panayiotis G Georgiou. Couples behavior modeling and annotation using low-resource lstm language models. In *INTERSPEECH*, pages 898–902, San Francisco, CA, September 2016.

[149] Shao-Yen Tseng, Brian Baucom, and Panayiotis Georgiou. Unsupervised online multitask learning of behavioral sentence embeddings. *PeerJ Computer Science*, 5:e200, 2019.

[150] Ming Tu, Yun Tang, Jing Huang, Xiaodong He, and Bowen Zhou. Towards adversarial learning of speaker-invariant representation for speech emotion recognition. *arXiv preprint arXiv:1903.09606*, 2019.

[151] Verena Venek, Stefan Scherer, Louis-Philippe Morency, John Pestian, et al. Adolescent suicidal risk assessment in clinician-patient interaction. *IEEE Transactions on Affective Computing*, 8(2):204–215, 2017.

[152] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and vision computing*, 27(12):1743–1759, 2009.

[153] Thurid Vogt, Elisabeth André, and Johannes Wagner. *Affect and Emotion in Human-Computer Interaction: From Theory to Applications*, chapter Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation, pages 75–91. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[154] Qing Wang, Wei Rao, Sining Sun, Leib Xie, Eng Siong Chng, and Haizhou Li. Unsupervised domain adaptation via domain adversarial training for speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[155] Martin Wöllmer, Angeliki Metallinou, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling. In *Proc. INTERSPEECH 2010, Makuhari, Japan*, pages 2362–2365, 2010.

[156] Rui Xia and Yang Liu. Leveraging valence and activation information via multi-task learning for categorical emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[157] Wei Xia, James Gibson, Bo Xiao, Brian Baucom, and Panayiotis G Georgiou. A dynamic model for behavioral analysis of couple interactions using acoustic features. In *Sixteenth Annual Conference of the International Speech Communication Association*, Dresden, Germany, September 2015.

[158] B. Xiao, P. Georgiou, B. Baucom, and S. S. Narayanan. Head motion modeling for human behavior analysis in dyadic interaction. *IEEE Transactions on Multimedia*, 17(7):1107–1119, July 2015.

[159] Bo Xiao, Zac E. Imel, Panayiotis Georgiou, David Atkins, and Shrikanth S. Narayanan. "rate my therapist": Automated detection of empathy in drug and alcohol counseling via speech and language processing. *PLoS ONE*, dec 2015.

[160] Zong-Ben Xu and Fei-Long Cao. Simultaneous lp-approximation order for neural networks. *Neural Networks*, 18(7):914–923, 2005.

[161] AmirAli Bagher Zadeh. Cmu-multimodalsdk. Available at https://github.com/A2Zadeh/CMU-MultimodalSDK (accessed March 2019), Jan 2019.

[162] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2236–2246, 2018.

[163] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.

[164] Puming Zhan and Martin Westphal. Speaker normalization based on frequency warping. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1039–1042. IEEE, 1997.

[165] Biqiao Zhang, Emily Mower Provost, and Georg Essl. Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences. *IEEE Transactions on Affective Computing*, 10(1):85–99, 2017.

[166] WQ Zheng, JS Yu, and YX Zou. An experimental study of speech emotion recognition based on deep convolutional neural networks. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 827–831. IEEE, 2015.

# Appendices

# A Detailed Network Architecture and Training Parameters of Chapter 3

| Multi-Emotion Regression Network (ER) framework (Input: 84 * 100 ; Output:6) |
|---|
| Training details: Adam optimizer(lr = 1e-05), batch size 16, MSELoss |
| Conv1d(in_ch=84, out_ch=96, kernel size=10, stride=2, padding=0) ReLU |
| Conv1d(in_ch=96, out_ch=96, kernel size=5, stride=2, padding=0) ReLU |
| Conv1d(in_ch=96, out_ch=96, kernel size=5, stride=2, padding=0) ReLU |
| Conv1d(in_ch=96, out_ch=128, kernel size=3, stride=2, padding=0) ReLU |
| AdaptiveMaxPool1d(1) |
| Linear(in =128, out =128) ReLU |
| Linear(in =128, out =128) ReLU |
| Linear(in =128, out =6) |

Table 6.1: Network architecture of ER

| Single-Emotion Classification Network (EC) framework (Input: 84 * 100 ; Output: 2) | |
|---|---|
| Training details: Adam optimizer(lr = 1e-05), CrossEntropyLoss, batch size: 32; 64; 128 | |
| Conv1d(in_ch=84, out_ch=96, kernel size=10, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=96, kernel size=5, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=96, kernel size=5, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=128, kernel size=3, stride=2, padding=0) ReLU<br>AdaptiveMaxPool1d(1)<br>Linear(in =128, out =128) ReLU<br>Linear(in =128, out =128) ReLU | Pretrained |
| Linear(in =128, out =64) PReLU<br>Linear(in =64, out =64) PReLU<br>Linear(in =64, out =2) | Trainable |

Table 6.2: Network architecture of EC

| | |
|---|---|
| B-BP based context-dependent behavior recognition model (Input: seq_len*6; Output: 5) | |
| Training details: Adam optimizer(lr = 1e-04) + Polynomial learning rate decay,<br>Masked BCEWithLogitsLoss, batch size: 1 | |
| Emotion recognition framework | Pretrained |
| GRU(in_size =6, hidden_size = 128, num_layers=2)<br>Linear(in =128, out =64) ReLU<br>Linear(in =64, out =5) | Trainable |

Table 6.3: B-BP based context-dependent behavior recognition model framework

| | |
|---|---|
| E-BP based context-dependent behavior recognition model (Input: seq_len * 84 * 100 ; Output: 5) | |
| Training details: Adam optimizer(lr = 1e-05) + Polynomial learning rate decay,<br>Masked BCEWithLogitsLoss, batch size: 1, epochs=300 | |
| Conv1d(in_ch=84, out_ch=96, kernel size=10, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=96, kernel size=5, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=96, kernel size=5, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=128, kernel size=3, stride=2, padding=0) ReLU<br>AdaptiveMaxPool1d(1) | Partly pretrained<br>Partly trainable |
| GRU(in_size =128, hidden_size = 128, num_layers=2)<br>Linear(in =128, out =64) ReLU<br>Linear(in =64, out =5) | Trainable |

Table 6.4: E-BP based context-dependent behavior recognition model framework

| | |
|---|---|
| E-BP based reduced context-dependent behavior recognition model<br>(Input: 84 * seq_len ; Output: 5) | |
| Training details: Adam optimizer(lr = 1e-04) + Polynomial learning rate decay,<br>Masked BCEWithLogitsLoss, batch size: 48, epochs=350 | |
| Conv1d(in_ch=84, out_ch=96, kernel size=10, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=96, kernel size=5, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=96, kernel size=5, stride=2, padding=0) ReLU<br>Conv1d(in_ch=96, out_ch=128, kernel size=3, stride=2, padding=0) ReLU | Behavior primitive embedding<br>(Pretrained) |
| Conv1d(in_ch=128, out_ch=96, kernel size=3, stride=2, padding=0)<br>**AvgPool1d(kernel size=2, stride=2)** ReLU<br>Dropout(prob=0.4)<br>Conv1d(in_ch=96, out_ch=96, kernel size=3, stride=2, padding=0)<br>**AvgPool1d(kernel size=2, stride=2)** ReLU<br>Dropout(prob=0.4)<br>Conv1d(in_ch=96, out_ch=96, kernel size=3, stride=1, padding=0)<br>**AvgPool1d(kernel size=2, stride=2)** ReLU<br>Dropout(prob=0.4)<br>Conv1d(in_ch=96, out_ch=128, kernel size=3, stride=1, padding=0)<br>**AvgPool1d(kernel size=2, stride=2)** ReLU<br>Dropout(prob=0.5)<br>AdaptiveMaxPool1d(1)<br>Linear(in =128, out =128) ReLU<br>Linear(in =128, out =64) ReLU<br>Linear(in =64, out =5) | Trainable |

Table 6.5: E-BP based reduced context-dependent behavior recognition model framework. Those AvgPool1d layers are optional to adjust temporal receptive field size.