Improving the Generalizability of Speech Emotion Recognition: Methods for Handling Data and Label Variability

by

Biqiao Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2018

Doctoral Committee:

Associate Professor Emily Mower Provost, Chair
Assistant Professor Dmitry Berenson
Visiting Research Professor Georg Essl
Professor Rada Mihalcea

Biqiao Zhang

didizbq@umich.edu

ORCID iD: 0000-0003-1598-2660

To my mother, who has always supported me and been a role model to me.

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest appreciation to my advisor, Dr. Emily Mower Provost, and my co-advisor, Dr. Georg Essl. Emily introduced me to the field of speech emotion recognition and they both provided me with unparalleled support, guidance, and encouragement throughout my Ph.D. years. I have learned a great deal from them, both as a person and as a researcher.

I would like to extend my sincere thanks to my committee members, Dr. Rada Mihalcea and Dr. Dmitry Berenson, for their constructive comments and suggestions.

I must also thank my wonderful labmates Yelin Kim, Duc Le, June Shangguan, John Gideon, Zakaria Aldeneh, Soheil Khorram, Matthew Perez, and Mimansa Jaiswal, and my dear friends Yuqing Kong, Xintong Wang, Yue Gong, Shinuo Weng, Lu Tian, and Kaitlyn Mallett. They have made my five years in Ann Arbor an enjoyable experience. Special thanks to my collaborators on the successful or unsuccessful projects, Yuqing, Zakaria, and Soheil, for all the inspiring discussions and for tolerating me when I was being difficult. I am also grateful to Yelin and Duc for their invaluable help and advice. Lastly, I thank MIDAS, TRI, and NSF for their financial support.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Emotion is an essential component in our interaction with others. It transmits information that helps us interpret the content of what others say. Therefore, detecting emotion from speech is an important step towards enabling machine understanding of human behaviors and intentions. Researchers have demonstrated the potential of emotion recognition in areas such as interactive systems in smart homes and mobile devices, computer games, and computational medical assistants. However, emotion communication is variable: individuals may express emotion in a manner that is uniquely their own; different speech content and environments may shape how emotion is expressed and recorded; individuals may perceive emotional messages differently. Practically, this variability is reflected in both the audio-visual data and the labels used to create speech emotion recognition (SER) systems. SER systems must be robust and generalizable to handle the variability effectively.

The focus of this dissertation is on the development of speech emotion recognition systems that handle variability in emotion communications. We break the dissertation into three parts, according to the type of variability we address: (I) in the data, (II) in the labels, and (III) in both the data and the labels.

Part I: The first part of this dissertation focuses on handling variability present in data. We approximate variations in environmental properties and expression styles by corpus and gender of the speakers. We find that training on multiple corpora and controlling for the variability in gender and corpus using multi-task learning result in more generalizable models, compared to the traditional single-task models that do not take corpus and gender variability into account. Another source of variability present in the recordings used in SER is the phonetic modulation of acoustics. On the other hand, phonemes also provide information about the emotion expressed in speech content. We discover that we can make more accurate predictions of emotion by explicitly considering both roles of phonemes.

Part II: The second part of this dissertation addresses variability present in emotion labels, including the differences between emotion expression and perception, and the variations in emotion perception. We discover that it is beneficial to jointly model both the perception of others and how one perceives one's own expression, compared to focusing on either one. Further, we show that the variability in emotion perception is a modelable signal and can be captured using probability distributions that describe how groups of evaluators perceive emotional messages.

Part III: The last part of this dissertation presents methods that handle variability in both data and labels. We reduce the data variability due to non-emotional factors using deep metric learning and model the variability in emotion perception using soft labels. We propose a family of loss functions and show that by pairing examples that potentially vary in expression styles and lexical content and preserving the real-valued emotional similarity between them, we develop systems that generalize better across datasets and are more robust to over-training.

These works demonstrate the importance of considering data and label variability in the creation of robust and generalizable emotion recognition systems. We conclude this dissertation with the following future directions: (1) the development of real-time SER systems; (2) the personalization of general SER systems.

CHAPTER 1

Introduction

1.1 Problem Statement and Methods

Emotion is an essential component in our interactions with others. It transmits information that helps us interpret the meaning behind an individual's behavior. The goal of *speech emotion recognition* is to provide this information, distilling emotion from multimodal speech data. With Human-Computer Interaction (HCI) shifting from machine-centered to human-centered [210], emotion awareness is becoming a desired property for many applications. For example, in augmented driving, an in-car system could detect a driver's emotion and could provide warnings or additional assistance given observations of anger, stress, or fatigue [71, 76, 201]. In augmented homes (e.g., Siri, Alexa, or Google Assistant), emotion awareness could provide an enhanced understanding of a user's behavior and could help in facilitating more natural and human-like interactions. In call centers, systems that are emotion-aware could better understand a caller's state and reroute frustrated individuals to agents that can more directly meet their needs [35, 104, 120, 121].

Speech emotion recognition is challenging because of the variability brought by many factors. For example, the expression of emotion differs across individuals, the environmental properties (e.g., noise level) alter how emotion is expressed and recorded, the ground truth emotion labels are subject to the influence of human annotators (e.g., the speaker themselves or outside observers), the perception of the same emotional display varies across observers, and the signals that emotion recognition systems rely on are subject to modulations resulting from the lexical content of speech. Therefore, speech emotion recognition systems must be robust and generalizable to handle the variability effectively.

This dissertation addresses the variability mentioned above, in order to increase the generalizability and robustness of speech emotion recognition systems. We explain the sources of variability in Section 1.1.1 below. The different sources of variability may not be readily separable from each other. However, each of these sources influences either the data or the emotion labels used to create speech emotion recognition systems. Therefore, we group the methods proposed in this dissertation by the type of variability they aim to handle (i.e., variability in data, labels, or both) and briefly introduce them in Section 1.1.2 through 1.1.4.

1.1.1 Sources of Variability in Emotion Recognition

Individual Emotion Expression. Individuals express emotions in a way that is uniquely their own. This variability in individual expression styles pose challenges to speech emotion recognition systems. Most works in emotion recognition focus on general models that do not explicitly consider the variability in emotion expression [79, 105, 157]. It may be possible for such models to be robust and generalizable given data that are sufficient in terms of both quantity and variability in expression styles. However, emotion corpora are often small in size. As a result, general models may be unintentionally over-fitting to specific individuals in the training set, resulting in poor generalizability and poor robustness in cross-corpus tasks.

Speech Content. Speech content is another source of variability in emotion expression. Emotions modulate speech acoustics as well as language. The latter influences the sequences of phonemes that are expressed, which in turn further modulate the acoustics. Therefore, phonemes introduce an additional source of variability in speech signals, making it harder to distill emotion from acoustic cues. Yet, the emotion expressed through speech content is also reflected in phoneme sequences. Previous works in speech emotion recognition have considered either the acoustic or the lexical properties of phonemes, but not both together.

Environmental Properties. The environmental properties, such as recording devices, noise level, recording distance (e.g., near field, far field), vary across data collections. Most works on emotion recognition focus on increasing the performance of the systems trained and tested on different subsets of the same corpus. This approach allows researchers to concentrate on developing effective signal processing and machine learning methods. However, systems that are constructed from a single collection of data may not be robust and generalizable enough to work well in the wild.

Types of Emotion Labels. Emotion may be defined in multiple manners: recognition of a person's true *felt sense*, how that person perceives his or her own behavior (*self-report*), or how others interpret that person's behavior (*perceived emotion*). The selection of a definition fundamentally impacts system design, behavior, and performance. For example, emotion recognition systems that are designed to be "omniscient" (e.g., deception detection) may rely on felt sense. However, this style of emotion recognition is extremely challenging because individuals may intentionally mask their state, resulting in a large difference between measured behavior and label. Further, if systems attempt to detect emotions that users are deliberately concealing, users may think that the systems are overly intrusive. Other emotion recognition systems designed to enable natural interaction between agents and people (e.g., augmented cars and homes) mostly rely on self-report or perceived emotion. Research has identified differences between self-report and perceived emotion labels. However, emotion recognition systems have focused only on a single type of label traditionally, rather than leveraging the potentially complementary information conveyed by the separate strategies.

Emotion Perception. Individuals exhibit differences in emotion perception. Multiple evaluators assessing the emotional content of the same data may have different opinions. Variability in perceived emotion is one of the core problems for automatic emotion recognition, in part, because the existence of variability in perception does not indicate some evaluators are wrong. Past works often mitigated inter-rater variability by averaging the ratings of groups of evaluators, under the assumption that this amalgamation can remove perceptual "noise". However, inter-rater variability may provide information about the subtlety or clarity of an emotional display.

1.1.2 Methods for Handling Data Variability

Addressing Variability Related to Corpus and Gender. The data variability in speech emotion recognition is a mixed result of different factors, including individual expression styles and environmental properties. These factors may be interwoven in practice. For example, speaker demographics, recording conditions, and emotion elicitation method are usually tightly associated with emotion corpora. As a result, it may be hard to address a single source of variability while keeping other sources fixed. In this work, we approximate multiple sources of variability by considering the training corpus as the explicit factor. Besides, we also consider the influence of gender on expression styles. We hypothesize that a more accurate and generalizable system can be created by training on multiple corpora while explicitly addressing the factors causing the variability. We test this hypothesis in Chapter 4 using a multi-task learning approach, by defining the tasks according to the identity of training corpora and the gender of the speakers. We show that this approach leads to systems that generalize well across corpora.

Incorporating Phonetic Information. We investigate how we can improve the prediction of emotional valence (positive vs. negative) by jointly considering (1) the variability in speech signals introduced by phonemes, and (2) the potentially emotion-related speech content contained in phoneme sequences. We hypothesize that systems that take both aspects into account are more accurate than those that do not consider the impact of phonemes or only consider (1) or (2). We test this hypothesis in Chapter 5. We present a network that exploits both the acoustic and the lexical properties of phonetic information by multi-stage fusion. This network first captures how phonemes modulate acoustics by aligning the two modalities in time and fusing the input features. We then combine the resulting phoneme-dependent acoustic representation with the utterance-level representation of phoneme sequences, which contains knowledge of the lexical content. Our results on two emotional datasets show that this approach outperforms systems that do not consider the influence of phonetic information or only consider a single aspect of such influence on unseen speakers.

1.1.3 Methods for Handling Variability in Labels

Modeling Self-reported and Perceived Emotion. Past works have demonstrated the discrepancies between the labels provided by the speaker expressing the emotions him/herself and other observers. In this work, we hypothesize that self-report and perceived emotion labels provide complementary information that could be used to improve the performance of systems designed to recognize each type of label. Specifically, perceived emotion labels could act as stabilizers to reduce fluctuations caused by individual speaker differences, while self-report labels could function as stabilizers to control for inter-personal emotion expression differences. We evaluate this hypothesis in Chapter 6, where we show how the two types of labels could be jointly modeled to improve the overall recognition ability on unseen speakers. We experiment using both the audio and video modalities and show that a multi-task learning approach, combined with unsupervised feature learning, could enhance the performance of emotion recognition systems for both types of labels.

Modeling Variability in Emotion Perception. In this work, we investigate methods that can effectively capture and predict the variation that is present in a population of evaluators. We focus on dimensional descriptions of emotion, which characterize emotion as continuous values (explained in detail in Section 1.2.1). This characterization naturally captures variation in emotion perception, allowing us to retain rich information about the emotional content of a given expression. We hypothesize that: (1) it is possible to generate a distribution of emotion perceptions from a limited number of ordinal evaluations; (2) modeling short-time temporal patterns is important, because emotion is a dynamic process and the evaluations may be based on different regions of an utterances, and (3) modality influences the effectiveness of the models for predicting the distributions of emotion perception. We evaluate the hypotheses in Chapter 7. We propose a label processing method for generating emotion distributions and show that a dynamic model that uses multiple modalities leads to improved performance on unseen speakers, compared to unimodal or static models.

1.1.4 Methods for Handling Both Label and Data Variability

We investigate methods for addressing the variability in both emotion labels and acoustic recordings by preserving emotional similarity using deep metric learning (DML). DML seeks to learn representations that encode similarity between examples automatically. These approaches generally presuppose that data can be divided into discrete classes using hard labels. However, speech emotion recognition works with inherently subjective data, data for which it may not be possible to identify a single hard label. In this work, we propose methods for DML with soft labels, a problem that is under-explored in traditional DML approaches. We apply these methods to speech emotion recognition, a field in which the benefit of DML is not yet investigated. We design a family of loss functions, f-Similarity Preservation Loss (f-SPL), based on the dual form of f-divergence. The minimizer of f-SPL preserves the pairwise label similarities in the learned feature embeddings. We hypothesize that this approach can enhance the performance and robustness of speech emotion recognition systems. Specifically, we can reduce the risk of unintentionally over-training networks to capture signals that are specific to either certain speakers or lexical artifacts in the data by pairing examples that differ in these factors. We evaluate this hypothesis in Chapter 8. We demonstrate the efficacy of our proposed methods on the task of cross-corpus speech emotion recognition with dimensional emotion descriptors. We find that our methods significantly outperform a baseline SER system with the same structure but trained with only classification loss. We show that the presented techniques are more robust to over-training and can learn an embedding space in which the similarity between examples is meaningful.

1.2 Emotion Background

1.2.1 Descriptions of Emotion

There are two prevailing frameworks for describing emotion: the categorical (or discrete) view [38, 68–70, 114, 118, 122, 176, 196] and the dimensional view [101, 135, 144, 199]. There is an active debate regarding the appropriateness of each approach (see [39, 138] for a detailed discussion). Yet, both categorical and dimensional descriptions are widely used in the emotion recognition community. We briefly introduce both approaches.

Categorical descriptions of emotion posit that there exists a small number of "basic" emotions. In his seminal work in 1884, James identified fear, grief, love, and rage as basic emotions [70]. The work was later extended in [38, 68, 69, 114, 118, 122, 176, 196]. Ekman defined a basic emotion as one that is differentiable from other emotions across a set of properties, including, automatic appraisal, distinctive physiology, distinctive universals in antecedent events, distinctive universal signals, coherence among emotional response, presence in other primates, quick onset, brief duration, and unbidden occurrence [38]. Basic emotions define the basis of the human emotional space, with other more complex emotions described as combinations of these bases [38]. However, one challenge associated with basic emotions as a construct is the variable nature of the sets of emotion identified. For example, Watson proposed fear, love and rage [196]. Izard proposed anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, and surprise [68, 69]. Plutchik proposed acceptance, anger, anticipation, disgust, fear, joy, sadness, and surprise [122]. Panksepp proposed expectancy, fear, panic, and rage [118]. Tomkins argued for nine out of the ten basic emotions proposed by Izard (excepting guilt) [176]. Oatley and Johnson-Laird proposed anger, anxiety, disgust, happiness, and sadness [114]. Ekman proposed anger, disgust, fear, happiness, sadness, and surprise [38], which are the most widely used in the field of automatic emotion recognition.

Dimensional descriptions of emotion characterize emotion as points in a continuous space. These descriptions, initially introduced by Wundt in 1897, included three dimensions as the basis of emotion and feeling, which were pleasant-unpleasant, tension-relaxation, and excitement-calm [199]. Schlosberg proposed that emotion could be described on a circular surface with pleasantness-unpleasantness, attention-rejection and level of activation as axes [144]. The two dimensions proposed by Russell [135], valence and arousal (also called activation), are the most commonly used dimensions in emotion recognition. A third dimension, dominance (dominant-submissive), is often used to distinguish emotions that are fundamentally different but cannot be identified by valence and activation [101]. For example, both anger and fear have negative valence and high activation, but anger is a dominant emotion while fear is more passive.

Dimensional characterizations of emotion remove the contextualized component of emotion. Instead, this description focuses on the “core affect” of a display, a term coined by Russell as “a simple and non-reflective feeling” that integrates two dimensions, valence (pleasure-displeasure) and arousal (sleepy-activated), to describe common emotional states [136, 139]. The dimensional view of emotion mitigates known problems, such as the lack of basic emotion universality. For example, some of the common basic emotions, such as fear and anger, do not have exact equivalents in all languages [137]. In [136, 139], Russell further argued it is problematic to use concepts such as anger and fear as psychological primitives because they are object-directed and imply a cognitive structure. He suggested the categorical emotions are more like constellations rather than stars, because they are categorized by their prototypical examples, and different cultures identified them differently. The prototypes of emotions such as anger, happiness, sadness, and fear are the occasional co-occurrence of a set of events that fit the pattern defined by the culture.

1.2.2 Appraisal Theory

Psychologists have proposed theories on how emotion occurs. For example, in the earlier works, Watson claimed that emotion could be caused by events directly [195]; Cannon suggested that emotion occurs through physiological process [28].

Appraisal theory forwards the notion that emotion is not purely reflexive, but rather responses result from *appraisals* of perceived events or situations. Appraisals are evaluations of a stimulus with respect to an individual’s well-being. In this case, well-being refers to the satisfaction or obstruction of everything that an individual cares about, including needs, attachments, values, goals, and beliefs [103].

Researchers have specified appraisal criteria (referred to as “dimensions”) that are important for differentiating emotions. The most common dimensions are novelty (whether the change in the environment is expected), pleasantness (whether the event or environment is pleasant or unpleasant), goal significance (the importance of the event to the individual’s goal), agency (whether the individual is responsible for the event), legitimacy (whether the action of the individuals fits moral standards or social norms) [46, 141, 163, 171].

Appraisal theory is widely used in affective computing [95]. The Ortony Clore Collins (OCC) model [116] is one of the most commonly used Appraisal Theories in this domain. It posits that an individual’s emotions are the result of his or her valenced appraisal of the current situation with respect to events, actors, and objects. The OCC model clearly defines 22 emotion categories as conjunctions of situation appraisals [95] and provides tools that enable the prediction of an individual’s emotion given knowledge of his/her goals and perception of events [32].

1.2.3 Brunswik’s Functional Lens Model

Brunswik’s functional lens model provides an explanation for how emotion is produced, transmitted, and perceived [18]. It contains two entities: the individual who produces the message (“encoder”) and one or more individuals who perceive the message (“decoders”). The emotion communication process starts with the encoder producing a message that conveys his or her communicative goals, together with various paralinguistic properties (e.g., emotion, age, gender). This message is encoded in distal indicators, which are a set of cues that are expressed over multiple channels, such as voice, facial expressions, or gestures. These cues are then transmitted to the decoders and referred to as proximal percepts. Proximal percepts serve as signals to the observers who then perceive and interpret the affective information. Through the

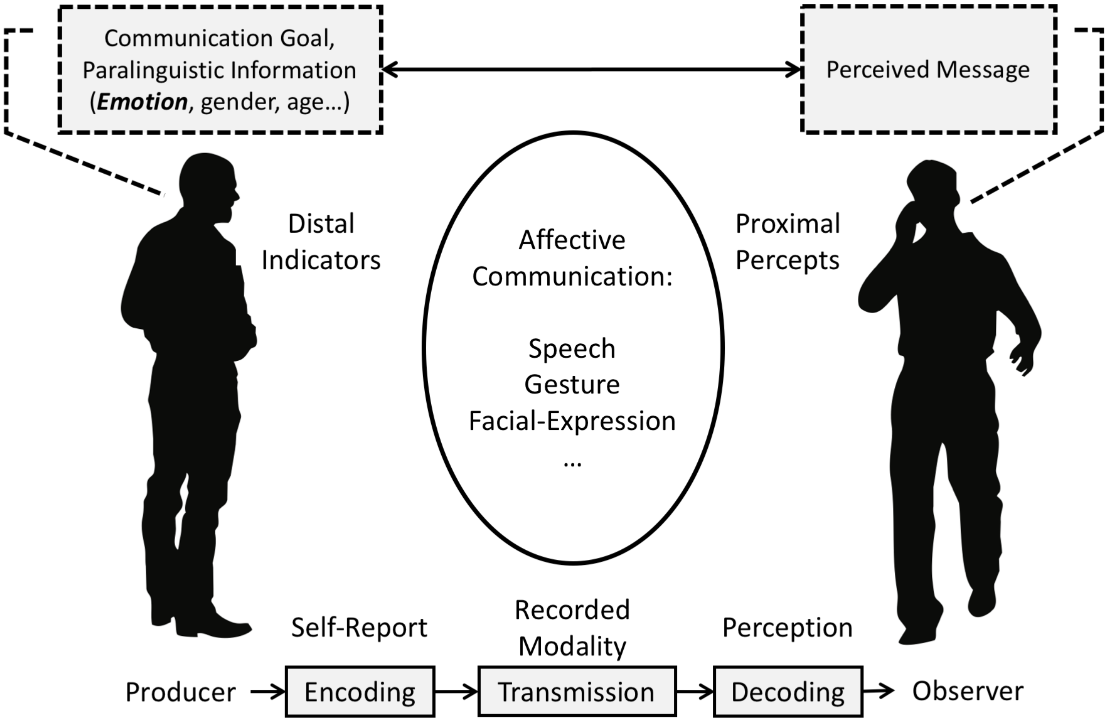


Figure 1.1: The Brunswik's Functional Lens Model

combination of multi-modal percepts, the decoders arrive at a higher-level judgment on the emotion of the encoder. Research has used adaptations of the Brunswik's functional lens model to explain the emotion communication process [11, 75, 142, 143].

The Brunswik's functional lens model both confirms the link between distal indicators and proximal percepts and highlights the fact that distal indicators produced by the encoder are not necessarily the same as the proximal percepts perceived by the decoders [142]. It provides a critical tool to understand the differences that exist between emotion expression and perception and between the perception of different individuals.

1.3 Contributions

Our works presents novel approaches to handle the data and label variability in speech emotion recognition and demonstrate that in doing so, we can create more robust and generalizable emotion recognition systems. The research contributions are as follows:

- We demonstrate the benefits of handling data variability with multi-task learning in cross-corpus speech emotion recognition in Chapter 4. We address multiple sources of variability, including individual expression styles and environmental properties, by considering two explicit factors: training corpus and gender. We explore effective ways to define and group the tasks according to these factors for different emotion dimensions.

- We address the variability caused by speech content by exploiting the acoustic and lexical properties of phonemes in a single model in Chapter 5. We explore how models utilizing the different functionality of phonemes perform given various types of lexical content (e.g., fixed, flexible given fixed scenarios) on valence prediction.

- We make the first attempt to jointly learn self-reported and perceived emotion in Chapter 6. We show the benefit of combining unsupervised feature learning and multi-task learning, with the tasks corresponding to the types of emotion labels.

- We quantify how groups of evaluators perceive emotional messages in Chapter 6. We show that the variability in emotion perception is a modelable signal and propose a label processing method for generating two-dimensional probability distribution from scarce ordinal labels. We make the first attempt to predict the 2D distributions of emotion perception for speech using a dynamic approach and show the advantage of considering both the acoustic and visual modalities.

- We propose a family of loss functions that aim to preserve real-valued label similarity, and a data sampling method for effectively implementing the loss functions in Chapter 8. We demonstrate the efficacy of the proposed methods on the task of cross-corpus speech emotion recognition. We show that these methods can be used to preserve emotional similarity and reduce the influence of non-emotional variability.

These approaches enable emotion recognition systems to handle different sources of variability, resulting in more generalizable and robust systems. The unification of these works will help in advancing the field of automatic speech emotion recognition. In addition, the loss functions that we propose for deep metric learning with soft labels can be applied to domains outside of speech emotion recognition.

1.4 Dissertation Outline

The dissertation is organized as follows. Chapter 2 provides an overview of related works. Chapter 3 introduces the datasets used in this dissertation. Part I (Chapter 4 and Chapter 5) discusses our work on handling data variability. Chapter 4 covers our work on addressing the variability caused by gender and training corpus, using multi-task learning. Chapter 5 introduces our work on handling variability in speech content by incorporating phoneme knowledge for the recognition of valence. Part II (Chapter 6 and Chapter 7) includes our work on handling label variability. Chapter 6 describes our work on jointly predicting self-reported and perceived emotions. Chapter 7 covers our work on capturing inter-rater variability by predicting emotion perception as distributions. Part III (Chapter 8) consists of our work on addressing the variability in both data and labels by preserving the real-valued emotional similarity between pairs of data using our proposed approach for deep metric learning with soft labels. Finally, Chapter 9 summarizes this dissertation and discusses potential future directions.

CHAPTER 2

Related Works

2.1 Observations of Emotion Variability

Emotion expression varies across individuals. This variability may be influenced by factors such as gender and culture. For example, Hall found that females can convey emotion through facial expression better than males [59]. Matsumoto et al. investigated the display rules of emotion on participants from five different countries and found that there were culturally-specific display rules [98].

In the Brunswik's functional lens model, emotion expression and speech content are encoded into multimodal cues together. Therefore, speech signals are modulated by both phonemes and emotions. Past work has investigated how emotions modulate the acoustics of individual phonemes. Leinonen et al. studied how the word "saara" was spoken in 10 different emotions by 12 speakers [88]. They found the phonemes contributed differently when expressing emotions. For example, there were specific intonations of "aa" associated with emotions of *astonished*, *scornful*, and *pleading*. Patel et al. studied short affect bursts using the "a" vowels. They analyzed physiological variations in phonation using the recordings produced by ten actors in five emotions [119]. They found that 11 acoustic parameters significantly differentiated the "a" vowels in the different emotion classes.

Emotion perception is also subject to the influence of individual, cultural, and gender differences. For example, the ability to decode emotion, often referred to as “emotion sensitivity”, differs across individuals [19, 132]. Past research has developed various methods for detecting the differences in emotion sensitivity, such as the Affective Sensitivity Test [27] and the Brief Affect Recognition Test [42], among other works [97, 117]. Ekman et al. conducted cross-cultural experiments on emotion perception from facial expressions [43]. They found that there were differences in the judgments of the absolute level of emotional intensity across cultures. The works of Elfenbein and Ambady found a within-group advantage for emotion perception [44, 45]. More specifically, they found that individuals can recognize the emotion of people from the same ethnic, national, or regional group more accurately than the emotion of people from different backgrounds. In addition, people may depend on different cues when perceiving emotions. Yuki, Maddux, and Masuda found that individuals from cultures that encourage emotional display rely more on the mouth region than the eyes region, while individuals from cultures that control for emotional display focus more heavily on the eyes, compared to the mouth [208]. Rotter and Rotter found that women are better at recognizing the emotion expression of both males and females in general [133].

The differences between emotion production and perception are demonstrated by perception experiments. Observers are not omniscient decoders of expressed emotion. For example, when actors are asked to portray a set of emotions, groups of observers correctly interpret the emotion only 50% to 80% of the time, where accuracy is defined in terms of correctly identifying the actors’ intentions [39, 142]. This demonstrates that emotion annotation is imperfect, even given clear emotional expression goals. In addition, emotion perception is influenced by the clarity of emotional displays. For example, individuals may intentionally suppress their emotion. Gross and Levenson conducted a study investigating the influence of emotion suppression on expressive behavior [57]. The participants of the study were asked to watch a short disgustinducing video in two conditions: (1) with no suppression, and (2) behaving "in such a way that a person watching you would not know you were feeling anything." They found that in (2), the expression behavior of the participants was reduced, compared to (1). Another example is the existence of micro-expressions, a fleeting facial expression, which reveals a true emotion that a person is trying to suppress [40]. They are hard to detect by the naked eye because of their brevity (they last less than 0.5 seconds) [200].

These observations have demonstrated that both emotion expression and emotion perception are variable and that there are differences between emotion expression and perception. They motivate our work on addressing the variability in speech emotion recognition.

2.2 Works Addressing Data Variability

2.2.1 Cross-Corpus Emotion Recognition

Emotion recognition systems rely on the audio-visual recordings of speech. Therefore, the systems are subject to the influences of recording equipment, noise level, among others. Past work has used cross-corpus evaluation for simulating this variability [87, 153, 160, 188].

Shami and Verhelst evaluated the generalizability of a segment-based speech emotion recognition method across two corpora, using three settings: within-corpus, cross-corpus, and integrated-corpus (i.e., merging corpora for training and testing) [160]. They found that the cross-corpus approach performed the worst, but integrated-corpus was more accurate than within-corpus. Lefter et al. found that cross-corpus performance was higher than within-corpus performance when the intra-corpus training set was limited and that integrating multiple corpora during training were beneficial [87]. These findings suggest that there are differences between corpora, but that common ground also exists.

Schuller et al. assessed the cross-corpus performance of emotion classification using four normalization methods (i.e., speaker-level, corpus-level, speaker-and-corpus-level, and no normalization) and found that speaker-level normalization performed the best [153]. They also found that cross-corpus performance could be improved by selecting datasets that have large distances between class centers, or selecting examples that are close to class centers [154]. Lefter et al. found that in cross-corpus evaluation, corpus-level normalization was better than normalizing based on the neutral examples of each corpus and that upsampling the sparse class had a positive effect [86]. Vlasenko et al. proposed a phoneme-based emotion classification system and achieved the state-of-the-art cross-corpus performance on two German emotion datasets [188].

Some works have focused on cross-corpus adaptation. Shah, Chakrabarti, and Spanias proposed two cross-corpus adaptation methods: (1) removing training instances that are classified incorrectly according to the development set in the test corpus; (2) penalizing the distance between the weights learned on the training corpus and the development set in the test corpus [158]. They found that both methods increased cross-corpus performance. Abdelwahab and Busso investigated two variants of Support Vector Machines (SVM) for domain adaptation: adaptive SVM and online SVM. They found that, for both methods, a significant performance gain could be achieved using only a small portion of the data from the target corpus for adaptation [1]. Similar findings were made in [93] using a domain adaptation method based on the idea of sharing priors between related classes of the source and the target corpora. Song et al. proposed transfer learning variants of two feature learning algorithms: Maximum Mean Discrepancy Embedding [169], and Non-negative Matrix Factorization [168]. They demonstrated the effectiveness of their proposed methods for cross-corpus evaluation on three speech emotion datasets.

Previous works have also investigated methods of enhancing the cross-corpus performance of speech emotion recognition using multiple training corpora. Schuller et al. proposed two methods: (1) merging multiple corpora for training; (2) training one classifier on each of the available training corpora, and fusing the results using majority vote [156]. They showed that both methods improved cross-corpus generalizability, although the preferred method varied across test corpora. In contrast, Lefter et al. found that for the recognition of negative interaction, training on two merged datasets produced a slightly lower performance than the best performance of training on each dataset separately [86]. Zhang et al. found that adding unlabeled data to merged multi-corpus training data increased the performance of cross-corpus emotion recognition [219]. However, the increase was only approximately 50% of the increase brought by adding labeled data.

These works have demonstrated the effectiveness of cross-corpus evaluation in measuring the generalizability of emotion recognition systems. In addition, they show that increasing the diversity by using multiple corpora for training is beneficial compared to training on a single collection of data. However, their approaches cannot take advantage of more diverse data and address variability across corpora at the same time. Besides, they did not consider variability caused by factors other than corpus.

2.2.2 Individual Variability in Expression

The expression of emotion is influenced by individual characteristics. One such factor is gender [59]. Most research in speech emotion recognition has focused on gender-independent models [79, 105, 157]. However, some works have analyzed gender variations in emotion recognition. Brendel et al. measured similarity between emotional corpora or sub-corpora of different genders using four similarity measures: recognition rate, correlation, groups of features, and feature-ranks [16]. They found that the data were less similar across genders than across corpora when using recognition rate and correlation as measures, yet the opposite holds when the latter two measures were used. This suggested that the differences between genders could be as large as the differences between corpora. Alghowinem et al. found that the best features for detecting depression from speech were different for females and males [5]. For example, log energy and shimmer were the most important for females, while loudness was the best feature for males. Vlasenko et al. applied context dependent vowel-level analysis based on gender-dependent features to emotion classification [189]. They showed that the system could detect high-arousal emotions accurately. Ververidis and Kotropoulos selected relevant features for each gender separately and trained gender-dependent classifiers [183]. Their results showed that the classification accuracy of gender-dependent classifiers was higher than that of a gender-independent classifier. Vogt and André combined gender detection and gender-dependent emotion recognition into a two-stage system [190]. They found that their system increased the emotion recognition rate by 2-4%, compared to gender-independent emotion recognition system. Similar observations were made in [159].

Research in fields related to speech emotion recognition, including the identification of affective facial expressions and body gestures, have investigated the effectiveness of considering individual differences using multi-task learning. Romera-Paredes et al. predicted pain level from facial expression and muscle activity from body gestures by applying multi-task learning in a transfer learning setting (MTL-TL), with subjects as tasks to account for idiosyncrasy [130]. They showed that their model outperformed models without MTL-TL. Another paper proposed a multilinear multitask learning method, and demonstrated its effectiveness on synthetic and real data, including recognizing the intensity of facial action units (AUs) associated with pain, with subjects and AUs as tasks in a tensor structure [129]. With the same task definition, Almaev, Martinez, and Valstar proposed an MTL-TL framework, and showed that it performed well even only with limited labeled data for the target tasks [7].

Shields et al. added a multi-task component to the Conditional Restricted Boltzmann Machines (CRM) [161]. They showed that jointly recognizing action, affect, and gender using their proposed model improved the performance of each task, compared to traditional CRBM and other baseline methods.

Related work has demonstrated the emotional variability caused by gender and the effectiveness of multi-task learning in related fields. However, most works in speech emotion recognition have not considered using multi-task learning approaches for addressing the variability in emotion expression across gender.

2.2.3 Influence of Phonemes

Previous works in speech emotion recognition have considered the phonetic modulation of acoustics using phoneme-level emotion modeling. Lee et al. presented one of the first investigations into the efficacy of phoneme-level modeling of speech signals for emotion classification [85]. They modeled the temporal behaviors of Mel-Frequency Cepstral Coefficients (MFCCs) using Hidden Markov Models (HMMs) in two settings: generic HMMs and HMMs based on five phonetic classes. They found that phoneme-class dependent models performed more accurately than the generic models and that emotion could be detected most accurately from vowels. Vlasenko et al. investigated phoneme-based classification of arousal, again using HMM models of MFCCs [188]. Their classifier showed a performance increase for cross-corpus evaluation, compared to the state-of-the-art. HMM-based phoneme-level emotion modeling was also used in [106, 107, 184–186]. Busso, Lee, and Narayanan classified emotion by investigating whether broad phoneme classes of emotional speech could be recognized using acoustic models of neutral speech [23]. They grouped phonemes into seven broad phoneme classes and trained a “neutral” HMM for each class using a neutral corpus. They then applied the models to emotion datasets and investigated how well the neutral reference models fit the emotional speech. They found that some phoneme classes were more heavily modulated by emotion than others. For example, vowels carried more emotional information than nasal sounds.

Other works have compared system performances when modeling emotion at different levels of granularity [125, 149, 187]. Ringeval and Chetouani compared emotion classification based on pseudo-phonetic segments (i.e., vowels and consonants) and voiced and unvoiced segments, using the k-nearest neighbors classifier [125]. They found that the vowel-based approach performed better than the voiced-based one. Schuller et al. compared emotion recognition accuracy at the word-level using phoneme-level models and word-level models [149]. They used HMM models of MFCCs for the phoneme-level model and Support Vector Machines (SVM) with 1,406 static acoustic features for the word-level model. They found that the word-level model performed better than the phoneme-level model. Similarly, Vlasenko et al. compared emotion recognition accuracy at sentence-level when using phoneme-, word-, and sentence-level model and found that sentence-level classification produced the best results [187]. These works suggest that longer segments are necessary for capturing emotion and that phoneme-level modeling may not be the best way to leverage phonetic knowledge. However, it is worth noting that the differences in features and classification methods during the comparison could also contribute to the performance difference.

Additional work has focused on generating emotion-salient acoustic features by leveraging phoneme information [14, 67, 115, 189]. Hyun, Kim, and Kwak evaluated acoustic features generally used in emotion recognition systems and categorized them into phoneme-dominant features and emotion-reflective features. They then used the former for phoneme detection, followed by extracting the latter from identified phonemes for emotion classification [67]. Bitouk, Verma, and Nenkova proposed to use statistics of MFCCs computed over stressed vowels, unstressed vowels, and consonants as utterance-level features. They showed that SVM emotion recognition systems using these features were more accurate than those using prosodic features or utterance-level MFCC features [14]. Vlasenko et al. extracted average first formant value from vowels as the acoustic feature and achieved a high accuracy on detecting high-arousal emotions [189]. Origlia, Cutugno, and Galatà proposed a feature extraction method, concentrating on the spectral content of syllabic nuclei and weighting features based on syllabic prominence [115]. Their features achieved comparable results to the state-of-the-art for continuous prediction of valence, activation, and dominance.

Some work has incorporated phonetic information in labels. Han et al. partitioned utterances into emotional and non-emotional segments, the latter consisting of silent regions, short pauses, transitions between phonemes, and unvoiced phonemes. They converted the utterance-level label to a label sequence of emotional state and *Null*, which corresponded to the emotional and non-emotional segments, respectively. They trained a Recurrent Neural Network with the Connectionist Temporal Classification loss using these label sequences [60]. Their method achieved the state-of-the-art performance on IEMOCAP.

Some recent works explored the potential of learning emotion-salient representations for speech content using phoneme sequences. In [54], Gamage, Sethu, and Ambikairajah proposed a representation called "bag-of-phoneme sequences" (BOP). They automatically recognized phonemes from utterances and identified all possible phoneme sequences of a fixed length. Each of these unique phoneme sequences is analogous to a word in the bag-of-words representation. Experimental results on the IEMOCAP dataset showed that their proposed relative-frequency-based lexical features extracted from the BOP representation achieved an Unweighted Average Recall (UAR) of 49% for the four-class emotion classification problem. An additional performance gain of 8.3% in UAR was achieved by fusing their lexical features with utterance-level acoustic features. In [53], Gamage, Sethu, and Ambikairajah proposed to use Bidirectional Long Short-Term Memory (BLSTM) networks for modeling variable-length phoneme sequences. They experimented using three possible representations: (1) one-hot vector, (2) phoneme compressed through embedding layer, and (3) Phone Log-Likelihood Ratio(PLLR). The best representation, PLLR, achieved a UAR of 56.4% for the four-class classification problem on IEMOCAP. Again, fusing the bottleneck layer of the BLSTM with the utterance-level acoustic features lead to increased performance (61.7%). Huang and Epps investigated the efficacy of using PLLR features for the recognition of valence and arousal in a multi-stage staircase regression system [64]. They found that PLLR features outperformed eGeMAPS acoustic features and that utterance-level PLLR features were more emotionally informative than frame-level PLLR features. Huang and Epps compared Bag-of-Audio-Word, Gaussian Mixture Model posteriors, bottleneck features, and PLLR features to acoustic features in continuous recognition of valence and arousal. They further proposed a set of phonetic-aware acoustic features that outperformed traditional acoustic features on three datasets [65]. Yenigalla et al. trained multi-channel Convolutional Neural Networks for acoustic features and phoneme, and showed that this approach outperformed models using only acoustic features in the literature [206].

Previous work has only considered a single aspect of phonemes: (1) phonemes as a source of modulation for speech signals, or (2) phoneme sequences as representations for speech content. Whether it is possible to leverage the dual roles of phonemes in a single model remains an open question. In addition, it is not yet clear how different lexical patterns (i.e., fixed, improvised with fixed targets, and spontaneous) will influence systems relying on phonetic information.

2.3 Works Addressing Label Variability

2.3.1 Self-Reported and Perceived Emotion

Most research in audio-visual emotion recognition focuses on using a single type of emotion label. Research in automatic recognition of both self-reported emotion and perceived emotion is relatively scarce. Truong, Neerincx, and Van Leeuwen conducted a series of experiments, comparing self-reported and perceived emotion using the TNO-GAMING corpus, which they collected. This corpus consists of audio-visual recordings of teams playing a multiplayer video game. The experimenters augmented the emotional experience by: (1) creating incentives, such as rewards for the best performing teams, (2) generating surprising events, such as sudden deaths, sudden appearances of threats, and (3) hampering the mouse and keyboard controls during the game. Afterward, each participant observed his/her recorded behavior, contextualized by the content of the game, and annotated his/her experienced emotion using categorical labels and semi-continuous dimensional labels of valence and arousal. These labels were augmented by perceived emotion labels under six different conditions: audio-only, video-only, audio and game context, video and game context, audio-visual, and audio-visual + game context. In the audio-visual + game context setting, the annotators have access to the same amount of material as in the self-report setting [178–180].

Truong, Neerincx, and Van Leeuwen measured the inter-observer agreement (agreement between multiple observers) and self-observer agreement (agreement between the outside observers and the individual him/herself) on the TNO-GAMING corpus using Krippendorff's α statistics [178]. They found that the inter-observer agreement is higher in the multi-modal evaluation setting, compared to the uni-modal settings. They then added self-report to the ratings of the observers and found that the agreement across individuals decreased, compared to only using the annotations of the observers. As a result, they constructed emotion recognition systems for the self-report and perceived emotion labels separately. They trained two sets of recognizers to detect levels of valence and activation, one with self-report labels and the other with perceived emotion labels using Support Vector Regression [180]. Their experiments suggested that it was easier to predict the perceived emotion labels, compared to self-report, for both valence and activation.

Truong, Van Leeuwen, and De Jong extended this work, conducting a more detailed comparison between self-report and perceived emotion and a comprehensive analysis on the performance of the systems when detecting the two types of emotion labels [179]. They observed that the emotions reported by the participants themselves tended to be more extreme than the average ratings of the outside observers. They predicted the valence and arousal ratings using both acoustic and lexical features and again found that perceived emotion could be predicted more accurately, compared to self-report. They argued that these results suggested that the emotions felt by the participants were not always perceivable by the observers. In addition, they conducted cross-label experiments: training on one type of label and testing on the other. They found that it the systems trained on the average of the perceived labels performed well on both perceived label and self-reported label, while systems trained using self-report labels had lower performance in comparison.

Other works have supported the mismatch between self-report and perceived emotion. Busso and Narayanan compared the self-report and perceived labels of the IEMOCAP dataset [25] across categorical and dimensional emotion descriptors [24]. They again found support for the mismatch between the self-reported and perceived emotion labels. Similar to [172, 178], they measured the agreement of labels obtained from different individuals using an entropy-based metric and the Kappa-statistic for evaluation agreement. They found that the level of agreement was significantly lower when the perceived emotion labels were augmented with the self-report labels. They also supported the finding that self-reported labels tend to have more extreme values for the dimensional descriptors, compared to the perceived emotion labels [24].

These works support the notion that there are differences between self-reported and perceived emotion labels [24, 178, 179] and that cross-training of different types of emotion labels may be possible [179]. However, it is not yet clear whether there is complementary information in different kinds of labels.

2.3.2 Inter-rater Variability

Emotion perception is subjective, which leads to inter-rater variability and uncertainty in emotion labels. Previous work has observed low inter-rater agreement in emotion datasets [162]. However, most works using categorical emotion labels arrive at a single hard emotion label by majority voting [12, 25, 148]. The majority works using dimensional descriptions of emotion either predict the mean of a group of evaluations [49, 58, 112, 115, 127, 198] or the mean weighted by rater-reliability [56], or restructure the emotion recognition problem as classification along each dimension [111, 153, 156, 216, 219]. While it is common practice for emotion datasets to collect multiple evaluations [25, 36, 126], none of these approaches models the variability in emotion perception captured by these evaluations.

One way to take this variability and uncertainty into account in classification tasks is to avoid the need to estimate single hard labels. For example, researchers have represented emotion information using probability distributions over emotion classes [4], confidence scores capturing the presence or absence of multiple emotion classes [105].

Research in handling emotion uncertainty has led to studies assessing the efficacy of augmenting speech emotion recognition training with soft labels. Research has shown that training with soft labels increases system performance in terms of standard classification measures [52] or an entropy-based measure that takes human confusion into account [172]. Lotfian and Busso proposed to consider the emotion perception of an utterance as a multidimensional Gaussian distribution over emotion classes [91]. They showed that systems trained using soft labels, calculated by taking the mean of the estimated Gaussian distribution, outperformed systems trained using hard labels.

Some works in music emotion recognition and cross-domain (song and speech) emotion recognition have forwarded the usage of soft labels to two-dimensional space and predicted emotion perception as probability distributions over valence and activation. There are two popular approaches: parametric (e.g., bivariate Gaussian and GMMs) [146, 192, 193] or non-parametric (discrete grid representation) [145, 202, 218]. In general, both approaches rely upon a large number of real-valued annotations.

Schmidt and Kim first proposed to model emotion perception from music as a probability distribution [146]. They assumed that the individual evaluations could be represented by a bivariate Gaussian. They formulated the task as a prediction of the Gaussian parameter associated with each short clip using several regression methods. They found that support vector regression (SVR) produced the best single-feature performance. However, the underlying assumption that the evaluations are guaranteed to follow a Gaussian distribution may not be valid, as noted in [192, 202]. Wang et al. proposed a generative model that learns two Gaussian mixture models (GMMs), one from acoustic features and the other from emotion labels [192, 193]. They predicted the emotional content of music as a probability distribution over the affective GMM components and summarized the prediction as a single Gaussian. However, while this approach used frame-level features directly, the utterance-level predictions were calculated by averaging the frame-level labels over the entirety of the utterance. The interactions across consecutive frames were not considered.

Another work of Schmidt and Kim represented emotion perception as a probability heatmap [145]. The evaluations were discretized into equally spaced grids. No assumption of the distribution of labels was made. They predicted the heatmaps over 1-second periods using Conditional Random Fields (CRF). The acoustic features were averaged over the 1-second window to reduce the frame-rate to that of the labels. Therefore, while CRF is context-dependent, there was information loss in the feature downsampling process. Yang and Chen generated a smooth probability density function from individual evaluations using Kernel Density Estimation (KDE) and then discretized the space [202]. The benefit is that the distribution is not biased by the position of the binning grids. They predicted the probability in each grid separately using SVR with utterance-level statistic features. Our previous work used a similar approach for predicting emotion perception across song and speech [218]. The difference is that we performed evaluator-dependent z-normalization to smooth the ordinal labels. This was valid because the evaluators were presented with relatively balanced data. However, both works used a static approach.

Works on predicting emotion perception as a distribution often rely on a large number of real-valued labels, which are not available in most popular emotion datasets. Besides, either feature downsampling or ignoring interactions across frames will result in information loss. The short-time temporal information is not fully exploited.

CHAPTER 3

Datasets

The number of publicly available emotion datas has continued to grow along with the popularity of the field. Early datasets in emotion recognition, such as the Berlin Emotional Speech-Database [20] and the Danish Emotional Speech Corpus [47], were recorded in laboratory environments with fixed lexical content and acted emotions [153]. The field has recognized the importance of modeling natural behaviors and have introduced new datasets that capture natural displays of affect, including human-robot interaction (e.g., FAU Aibo [12]), or recordings taken from public media (e.g., VAM [55]). Researchers have also focused on emotion induction as a technique to elicit emotional behaviors (e.g., SEMAINE [99, 100]). Recent acted datasets have included altered elicitation protocols to increase naturalness, for example, using improvisation (e.g. IEMOCAP [108]) or increasing the diversity of speakers' cultural backgrounds (e.g. eNTERFACE [96]).

In this dissertation, we use six emotion datasets in total, including the Berlin Emotional Speech-Database (EmoDB) [20], the eNTERFACE corpus [96], the Vera am Mittag German Audio-Visual Emotional Speech Database (VAM) [55], the AVEC (2011) corpus [151, 152], the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [108], and the MSP-IMPROV corpus [26]. In this chapter, we introduce these datasets and defer the explanation of the data processing methods used in each work to the corresponding chapters.

3.1 EmoDB

The Berlin Emotional Speech-Database consists of audio recordings of 10 German speakers reading lexically neutral sentences in seven emotions: anger, boredom, disgust, fear, joy, sadness and neutrality. The utterances were labeled using the target emotion. Human evaluations of perceived emotion and naturalness were also collected. Usually, utterances with a recognition rate higher than 80% and naturalness higher than 60% during human evaluation were kept. This results in 493 utterances. More details can be found in [20].

3.2 The eNTERFACE Corpus

The eNTERFACE Corpus consists of speech with fixed lexical content in six emotions: angry, happy, fearful, sad, disgust, and surprise. The emotions of speakers were elicited by short stories during the recording. The released dataset contains audio-visual recordings of 44 speakers from 14 different nations that were assessed as emotionally unambiguous by two experts. After excluding unsegmented data from a speaker (speaker number 6), this dataset contains 1,287 utterances from 43 speakers. See [96] for more details.

3.3 VAM

The Vera am Mittag German Audio-Visual Emotional Speech Database consists of spontaneous emotional speech from a German TV talk-show. We use the VAM-Audio portion of the corpus, which contains audio recordings from 47 speakers that were evaluated as "very good" or "good" by human evaluators in terms of the usability for emotion analysis. The recordings were provided as utterances, which are mostly complete sentences, but also include some exclamations, affect bursts, and incomplete sentences, because of the spontaneous nature of the data. This results in 947 utterances. The utterances were continuously labeled by human evaluators (17 for speaker 1-19, 6 for speaker 20-47) on valence, activation and dominance (weak vs. strong) [55].

3.4 AVEC (2011)

The AVEC (2011) corpus was created from the Solid-SAL partition of SEMAINE [99, 100]. It contains interactions between users and four emotionally stereotyped characters played by human operators. The combined training and development set of AVEC includes 63 sessions, where each session is the interaction between a user and a character. Each interaction was fully transcribed, and was annotated by at least two raters along the dimensions of valence, activation, expectation (expecting vs. being taken unaware) and power (weak vs. strong). Binary word-level labels are provided for each dimension. Additional information about the AVEC and SEMAINE datasets can be found in [99, 100, 151, 152].

3.5 IEMOCAP

The Interactive Emotional Dyadic Motion Capture Database consists of five sessions of dyadic interactions, each between a male and female actor. The emotional behaviors were elicited using scripted and improvised scenarios. The scripted scenarios in the five sessions have the same lexical content. The improvisation targets are shared across sessions, although the actual speech contents vary. The sessions were segmented into speaker turns (*utterances*). The scripted and improvised portions of IEMOCAP consist of 5,255 and 4,784 utterances, respectively. Manual transcriptions are provided for all the interactions.

IEMOCAP includes 12 hours of data across three modalities: audio, video, and motion-capture (referred to as “mocap”). The mocap recording was made over a single actor at a time. Consequently, only half of the data have matched audio-visual and mocap recordings (see [25] for details about the recording setup). This results in 5,042 utterances.

The data were evaluated at the utterance-level. The evaluations include both categorical and dimensional labels. Each utterance was annotated by at least three evaluators for categorical emotions. The categorical labels were chosen from the set of {angry, happy, neutral, sad, frustrated, excited, disgusted, fearful, surprised, other}. There was no limitation on the number of labels an evaluator could select for a given utterance. The valence, activation, and dominance levels of each utterance are evaluated by at least two annotators using a 5-point Likert scale. Six out of ten actors were asked to self-report the emotional content of their own recordings of the improvised scenarios. They used the same evaluation paradigm as the evaluators, described above. More details about IEMOCAP can be found in [25].

3.6 MSP-Improv

The MSP-Improv corpus is an audio-visual dyadic emotion corpus. MSP-Improv consists of six sessions, each including interactions between a male and a female actor. This results in 12 speakers in total. The emotional expressions of the speakers were elicited through carefully designed scenarios that include improvisations and target sentences with specific lexical content. Because of the recording paradigm of MSP-Improv, there are four types of recordings in the dataset: (1) the target sentences read by the actors; (2) the target sentences from the improvised scenes collected using emotionally evocative scenarios; (3) the speaker turns in the improvised scenes; (4) the natural spontaneous interactions during breaks between improvisations. The dataset includes over nine hours of data, segmented into 8,438 *utterances* (i.e., speaker turns or target sentences). The numbers of utterances corresponding to the four types of recordings are 620, 652, 4,381, and 2,785, respectively [26].

The emotional content of MSP-Improv was evaluated using crowdsourcing (Amazon Mechanical Turk). A scheme was designed to ensure the reliability of the labels by stopping evaluators when their inter-rater agreement with known "gold-standard" evaluations dropped [21, 22]. Each utterance was annotated by at least five evaluators using both dimensional and categorical rating paradigms. For the dimensional labels, the evaluators were required to access the valence, activation, dominance (dominant vs. submissive) and naturalness (acted vs. natural) of the utterances using a five-point Likert-scale. The categorical labels were selected from the set of {angry, happy, sad, neutral, other} [26].

Part I

Data Variability

CHAPTER 4

Addressing Variability in Corpus and Gender using Multi-task Learning

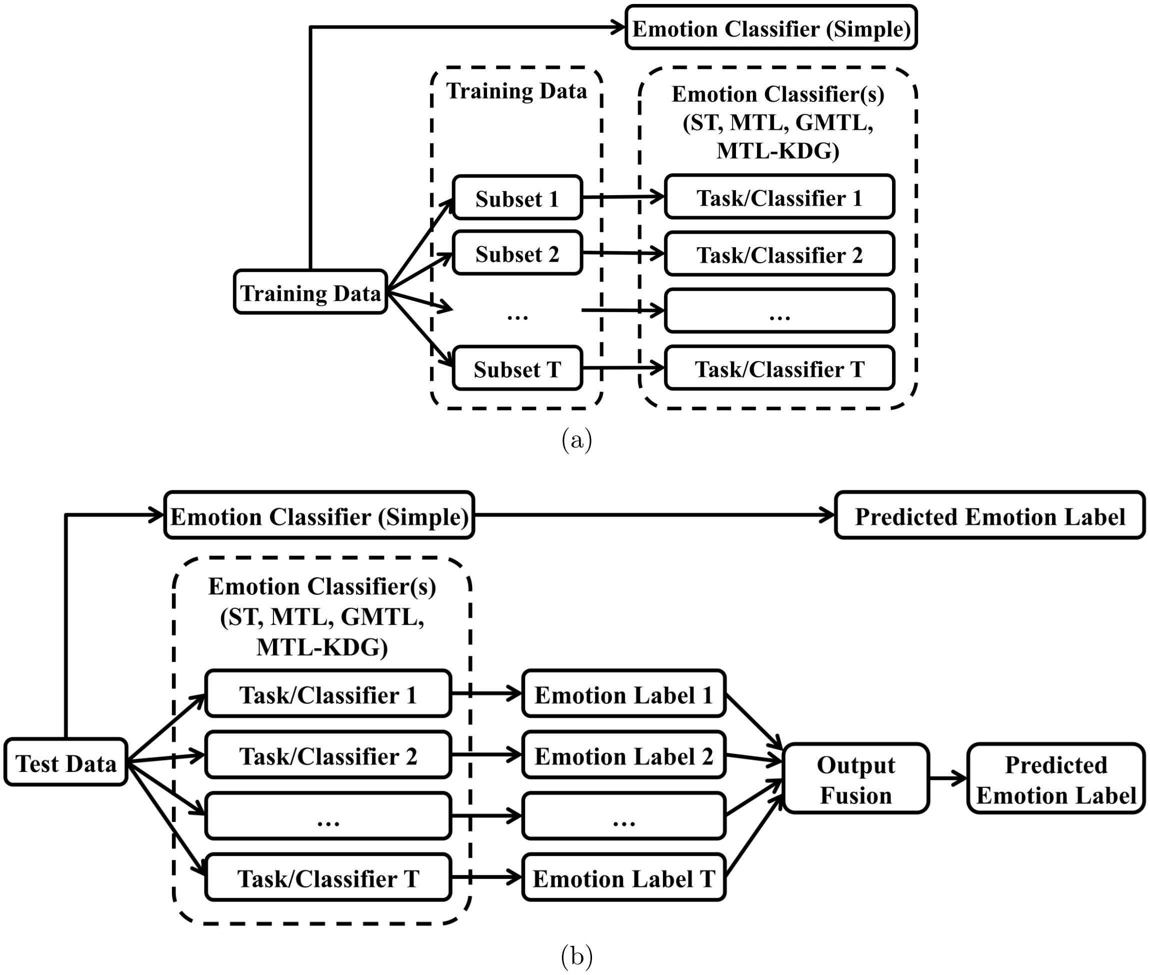
4.1 Introduction

One challenge that arises in the real use cases of emotion recognition systems is the presence of variations in emotion data that occur naturally in the wild, caused by factors including speaker characteristics, languages, lexical content, noise level and recording conditions. These factors may be closely related to the corpus that the data belong, resulting in difficulty to tease the influence of them apart, as mentioned in Chapter 1. As a result, researchers have approximated this challenge by performing cross-corpus analyses [87, 153, 154, 156, 169] and have demonstrated the efficacy of using multiple training corpora for enhancing cross-corpora robustness [153, 156]. However, it is not yet known how to best take advantage of the variability introduced by these training corpora.

There are additional sources of variability that emotion recognition systems need to handle, including gender. While most works in emotion recognition use gender-independent systems [79, 105, 157], previous studies have shown that gender-dependent models outperform those that are gender-independent [84, 183, 190]. This suggests that there exist similarities in emotion expression across genders, and that the performance of systems increases when controlling for the pervasive differences [84, 183, 190].

Most of the previous work in speech emotion recognition addresses the variations caused by corpus and gender differences in two ways: (a) increasing the variations in the training data, e.g., by merging multiple corpora during training [153, 156]; (b) controlling for particular sources of variation in the training data, such as training gender-dependent models [190] or training multiple corpus-specific classifiers and performing late fusion [156]. While multi-task learning has been demonstrated to be useful in affect recognition from visual input [7, 130, 161], its effectiveness on speech emotion recognition is under-explored. In this work, we investigate the influence of corpus and gender on emotion recognition by combining (a) and (b) using multi-task learning. We hypothesize that we will obtain a more accurate and generalizable emotion recognition system, compared to (a) and (b), by seeking common ground across different factors, while preserving the differences in the learned emotion patterns associated with a specific corpus, gender, or their combination. In this work, multi-task learning refers to jointly training multiple tasks, which contain non-overlapping sets of instances that share a same set of labels.

We explore five models to test our hypothesis: (1) a simple model, where we train a single classifier using all the data; (2) a separate-task (ST) model, where we train task-specific classifiers individually; (3) a multi-task learning (MTL) model, where all the tasks are considered related; (4) a group multi-task learning model (GMTL), where only the intra-group relatedness is assumed and the task grouping is learned with task-specific weights; (5) a multi-task learning with knowledge-driven grouping (MTL-KDG) model, where the group is pre-defined based on knowledge instead of learned as in (4). The first two models are our baselines because they have been shown to be useful for cross-corpus emotion recognition that considers corpora as tasks [153, 156]. Figure 4.1 illustrates the training and testing phase of the proposed



(a)

(b)

Figure 4.1: System diagram for the proposed classification framework, including the: (a) training phase and (b) testing phase. In the simple model, only one classifier is built using all the training data and only one label is generated. In all other models, either T classifiers are trained (ST) or one classifier with T classification tasks is trained (MTL, GMTL and MTL-KDG). T labels are output for each test case and are fused to determine the final label. ST: separate-task model, MTL: multi-task learning model, GMTL: group multi-task learning model, MTL-KDG: multi-task learning with knowledge-driven grouping.

methods.

We present a set of experiments to explore the influence of corpus and gender using four speech emotion datasets. The training data are separated into subsets according to corpus and/or gender, where each subset is treated as a task. We perform weighted majority voting to fuse the test labels output by the tasks, where the votes are weighted by a measure of confidence, as in [215].

We find that variations in corpus and gender all influence emotion recognition. In general, models using multi-task learning methods outperform models that treat the tasks as identical or independent. Data-driven grouping is comparable to knowledge-driven grouping. Defining tasks by gender is more beneficial than by corpus or both corpus and gender for valence, while the opposite holds for activation. On average, the system performance increases with the number of training corpora. The novelty of this work includes: (1) an analysis of the benefits of multi-task learning in cross-corpus emotion recognition, with tasks defined by corpus and/or gender; (2) an exploration of effective ways to define tasks for valence and activation; (3) an examination of the influence of sparsity on different feature spaces; (4) a comparison of knowledge-driven and data-driven task grouping.

4.2 Data

We select four datasets covering different types of emotion (acted and spontaneous) and languages (English and German) to investigate the cross-corpus generalizability of the proposed methods. We conduct experiments concentrating on the variability caused by training corpus and gender. We use EmoDB [20] and eNTERFACE [96] to represent acted emotion in German and English, respectively, and VAM [55] and AVEC (2011) [151, 152] to represent spontaneous emotion, again in German and English, respectively. These experiments use binary labels of valence (negative vs. positive) and activation (calm vs. excited). The meta information about the datasets can be found in Table 4.1. We describe our data processing procedures for each corpus below.

For each dataset, we convert the provided labels to the binary dimensional labels. Both EmoDB and eNTERFACE were annotated for categorical emotions only. We map the emotion categories to binary valence and activation labels (see Table 4.2), following [153, 156]. VAM was continuously labeled for valence and activation. We take the sign of the mean valence and mean activation of each utterance as the binary labels, following the process in [156]. We the training and development set of AVEC (2011). The dataset was labeled for binary valence and activation at the word level. We segment the recordings data into turns and generate the turn-level emotion labels from the word-level labels using majority vote.

We extract the "emo_large" feature set defined in openSMILE from each utterance, as in [156]. It consists of 6,669 features, generated from 57 acoustic frame-level low-level descriptors (LLDs) by calculating 39 statistics over the LLDs, ΔLLDs, and ΔΔLLDs. We apply speaker-dependent z-normalization to the utterance-level features.

4.3 Classification Models

Related work introduced in Section 2.2.1 and 2.2.2 has indicated that there exist both differences and similarities across corpora and genders for emotion recognition. In addition, multi-task learning methods have been demonstrated effective in visual affective computing. However, most works in speech emotion recognition either concentrated on increasing data variability (e.g. merging of multiple corpora as the training set), or focused on controlling for variability (e.g. separate classifiers for each available training corpus, gender-dependent classifiers). The design of classification approaches that leverage common ground across different corpora and genders while preserving the inherent differences is still under-explored.

		EmoDB	eNTERFACE	VAM	AVEC
Language Lexical Content Type		German fixed acted	English fixed acted	German natural spontaneous	English natural spontaneous
Speaker -level	# All	10	43	47	16
	# F	5	9	36	10
	# M	5	34	11	6
Utterance -level	# All	493	1,287	947	2,368
	# V(+)	142	427	72	1,534
	# V(−)	351	860	875	834
	# A(+)	246	857	445	1,280
	# A(−)	247	430	502	1,088
	# F	286	270	751	1,620
	# M	207	1,017	196	748

Table 4.1: Dataset details of EmoDB, eNTERFACE, VAM and AVEC. spon.: spontaneous; F: female; M: male; V: valence; A: activation

We present five classification models: the simple model, separate task (ST) model, multi-task learning (MTL) model [8, 9], group multi-task learning (GMTL) model [74], and multi-task learning with knowledge-driven grouping (MTL-KDG) model [8, 9]. We define a task as emotion recognition using data from a specific factor (e.g., a corpus), or a specific combination of two factors (e.g., a corpus-gender pair). The five classification models correspond to five different assumptions about the tasks. The simple model assumes that the tasks are identical, and merges data from all the

Emotion	Appearance	Valence	Activation
Anger	EmoDB, eNTERFACE	−	+
Happiness	EmoDB, eNTERFACE	+	+
Neutrality	EmoDB	+	−
Sadness	EmoDB, eNTERFACE	−	−
Fear	EmoDB, eNTERFACE	−	+
Disgust	EmoDB, eNTERFACE	−	−
Surprise	eNTERFACE	+	+
Boredom	EmoDB	−	−

Table 4.2: Mapping from categorical emotions to binary valence and activation.

tasks for training. The ST model sees the tasks as independent, and trains a separate classifier for each task. The simple and ST models are similar to the "pooling" and "voting" strategies in [156], respectively, if we consider each corpus as a task. Therefore, we use simple and ST as baselines in our experiments. The MTL model assumes that the tasks are related and share a common sparse feature representation. The GMTL model assumes that the tasks can be clustered into groups, and only intra-group information sharing is allowed. Finally, the MTL-KDG model assumes that information is shared within a group, but it predefines groups based on knowledge such as gender or corpus, instead of learning the groups from data.

We use linear Support Vector Machine (SVM) in the simple and ST models, as in previous cross-corpus emotion recognition works [153, 154, 156]. We adopt two types of regularization: L_2-regularization ([153, 154, 156]), and L_1-regularization, which assumes sparsity of the features. The MTL model and each group of tasks in the MTL-KDG model use the multi-task feature learning algorithm [8, 9]. The GMTL model uses the group multi-task learning algorithm [74].

4.3.1 Multi-task Feature Learning

The multi-task feature learning algorithm [8, 9] learns a common feature representation across tasks using the $L_{2,1}$-norm regularization, which enforces sparsity of the features across tasks. There are two settings of this algorithm: (a) feature learning (FL) and (b) feature selection (FS). The major difference between them is that in (a), the $L_{2,1}$-norm regularization is imposed on a transformed feature space, while in (b) the regularization is imposed directly on the original feature space.

The objective function of setting (a) is given by Eq. (4.1). It is assumed that the weight matrix, W, whose column vectors are the weights $\mathbf{w_t}$ of individual tasks, can be rewritten into $W = UA$, where $U^TU = I$ (identity matrix) and A is the weight matrix for a transformed feature space.

$$\min_{U,A} \sum_{t=1}^{T} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{a_t}, U^T \mathbf{x_{ti}} \rangle) + \gamma \|A\|_{2,1}^2 \tag{4.1}$$

Eq. (4.1) contains two terms: the loss term (first) and the regularization term (second). The loss term is the summation of the loss, $L(.)$, across T tasks. Here, m_t is the number of training instances in task t, $y_{ti} \in \{-1, 1\}$ is the label of the i-th instance in task t, $\mathbf{a_t}$ is t-th column of A, $\mathbf{x_{ti}}$ is the i-th training instance of task t, and $<>$ stands for inner product. The regularization term is the product of the regularization parameter γ and the squared $L_{2,1}$-norm of A. $L_{2,1}$-norm is defined as the L_1-norm of the vector produced by taking the L_2-norm of each row of A.

Setting (b) is a special case of (a). In (a), U and A are learned together from the data, while in (b), we force $U = I$. In this way, the "feature learning" in (a) reduces to the special case of "feature selection" in (b) [8, 9].

The problem given by Eq. (4.1) is non-convex. However, [8, 9] proved that it has an equivalent convex form that can be solved by iteratively minimizing over W (Eq. (4.2)) and a $d \times d$ matrix D, where d is the dimensionality of the input features. Specifically, we first initialize D to $\frac{I}{d}$, and then iteratively perform two steps:

- Fix D, solve the task-specific optimization by Eq. (4.3).
- Fix W, update D using Eq. (4.4) for setting (a) or Eq. (4.5) for setting (b). The ϵ in Eq. (4.4) is a perturbation parameter used to ensure the convergence of the problem. The $\mathbf{w^i}$ in Eq. (4.5) denotes the i-th row of W.

$$\min_{W} \sum_{t=1}^{T} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \sum_{t=1}^{T} \langle \mathbf{w_t}, D^{-1} \mathbf{w_t} \rangle \tag{4.2}$$

$$\mathbf{w_t} = \operatorname{argmin} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \langle \mathbf{w_t}, D^{-1} \mathbf{w_t} \rangle \tag{4.3}$$

$$D = \frac{(WW^T + \epsilon I)^{\frac{1}{2}}}{trace(WW^T + \epsilon I)^{\frac{1}{2}}} \tag{4.4}$$

$$D = Diag(\lambda), \text{where } \lambda_i = \frac{\|\mathbf{w^i}\|_2}{\|W\|_{2,1}} \tag{4.5}$$

Eq. (4.3) holds for any convex loss function. In this work, we choose the hinge loss (Eq. (4.6)) to match the linear SVM used in the simple model and ST model. Note that Eq. (4.3) with hinge loss is equivalent to linear SVM with a variable transformation trick.

$$L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) = \max(0, 1 - y_{ti}\langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) \tag{4.6}$$

4.3.2 Group Multi-task Learning

Group multi-task learning [74] assumes that the tasks belong to several groups that can be learned together with task-specific weights. Only the tasks that are grouped together share information. This method was built directly on the multi-task feature learning algorithm above. In [8], it was proved that the optimization problem given by Eq. (4.1) is equivalent to Eq. (4.7), where $\|W\|_{tr}^2 = trace(WW^T)$.

$$\min_W \sum_{t=1}^{T} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \|W\|_{tr}^2 \tag{4.7}$$

Analogous to Eq. (4.7), the objective function of group multi-task learning becomes Eq. (4.8) given the group assignments. Here, G is the number of groups. Q_g is a diagonal matrix with diagonal entries being the binary group assignment values for group g, and $\sum_g Q_g = I$. The optimal G is not known a priori and is treated as a hyper-parameter. When $G = T$, group multi-task learning is equivalent to solving each task individually, and when $G = 1$, it is the same as the multi-task feature learning.

$$\min_{W,Q} \sum_{t=1}^{T} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \sum_{g=1}^{G} \|WQ_g\|_{tr}^2 \tag{4.8}$$

Eq. (4.8) is a mixed integer programming problem. It can be solved by iteratively performing two steps:

- Fix Q, solve group-specific optimization given by Eq. (4.9). $W_g = WQ_g$, and q_{gt} is the tth diagonal entry of Q_g.
- Fix W and solve for Q. See details in [74].

$$\min_{W_g} \sum_{t:q_{gt}=1} \sum_{i=1}^{m_t} L(y_{ti}, \langle \mathbf{w_t}, \mathbf{x_{ti}} \rangle) + \gamma \|W_g\|_{tr}^2 \tag{4.9}$$

The second step is non-convex, and the solution could become stuck in a local optimum. We address this problem by training multiple times and fusing the labels.

All other models, except for the simple model, learn a different weight vector for each task. Therefore, there could be T predicted labels for a given test instance. Although it is common in the multi-task learning literature to assume knowledge about the tasks of the test data [8, 9, 74], we do not make this assumption. This is because: (1) it requires additional information about the test data, which may not be available in real applications; (2) the test data may not strictly belong to any of the tasks (e.g., test data is from an unseen corpus when using each training corpus as a task). In this work, we generate the final output label by weighted majority vote, where each task gives a vote to the label it outputs, weighted by the distance to the decision hyperplane. This method was demonstrated to outperform other output selection or fusion methods in [215].

4.4 Experimental Design

We designed a set of experiments using cross-corpus evaluation to investigate the influence of corpus and gender on on the binary (positive vs. negative) classification of valence and activation, using four speech emotion datasets. We hypothesize that we can achieve better performance by splitting data into tasks, and controlling for the degree of information sharing between the tasks using multi-task learning.

We conduct three sets of sub-experiments, focusing on the impact of corpus (c), of gender (g) . In experiments c and g, we train the simple, ST, and MTL models using corpora and genders as tasks, respectively. Experiment cg investigates the joint impact of corpus and gender, with each corpus-gender pair as a task. In cg, we also train GMTL and MTL-KDG in addition to simple, ST and MTL. We group the tasks by corpus and gender for MTL-KDG, denoted as MTL-GC and MTL-GG below. In each sub-experiment, we use one, two or three corpora for training, and test on each of the remaining corpora separately. Note that the simple model is the same across all sub-experiments when the same training corpora are used. When there is only one training corpus, all models in c are identical to the simple model and experiment cg is not performed.

In each sub-experiment, we compare the performance of different models to test the underlying assumptions. We compare the performance of ST and MTL across the sub-experiments, to investigate the three ways of defining the tasks (corpus, gender, or corpus-gender pair). In addition, we compare the performance on the same test corpus when using different numbers of corpora for training to investigate the impact of adding additional training corpus.

We tune the hyper-parameters by maximizing the leave-one-speaker-out (LOSO) cross-validation accuracy of the training set, in the ranges below:

- Regularization parameter γ (in simple, ST, MTL, GMTL and MTL-KDG):

$\{10^{-4}, 10^{-3}, ..., 10^{3}\}$. Note that γ is equivalent to the cost parameter C for the error term in linear SVM (simple and ST), where $C = 1/(2 \times \gamma)$.

- Permutation parameter ϵ (in FL setting of MTL, GMTL and MTL-KDG): $\{10^{-8}, 10^{-7}, ..., 10^{0}\}$.
- Number of Groups G (in GMTL): $\{1, 2, ..., T\}$.

We use 5-fold cross-validation, where the folds are divided at speaker-level for each task to avoid overfitting to known speakers. In the cross-validation process, we use average UAR of the tasks as the performance measure if the model contains more than one task, because the data are not evenly distributed across the tasks.

We solve the linear SVMs using Liblinear [50]. We use a fixed number of iterations as the stop criteria for multi-task learning, as in [2, 33]. For multi-task feature learning, we fix the number of iterations to 20, according to [8]. For group multi-task learning, we fix the outer-iteration to five as in the example code from the author of [74], and the group-specific inner-iteration to 20.

4.5 Results

We analyze the binary classification results of valence and activation on four speech emotion datasets. We compare the performance between: (1) different assumptions on feature sparsity, (2) different training-testing combinations, (3) different models while controlling for task definition (e.g., corpus as the task), (4) different task definitions while controlling for model, (5) different number of training corpora while controlling for model and task definition, and (6) cross-corpus and within-corpus.

We use a repeated measure model (denoted as RM) with mixed factors for the comparisons. We treat the test corpus (e.g., EmoDB) as the between-subject factor because there are multiple experiments run on each test corpus. Thus, the overall set of results has underlying dependencies. The within-subject factors (denoted as

WSF) include: version (e.g., L_1-regularization), model (e.g., ST), task definition (e.g., gender), and number of training corpora.

After fitting the results into an RM, we perform the repeated-measure ANOVA (denoted as RANOVA) for each dimension (i.e., valence and activation). If the WSF is significant, we perform the Tukey's honest significant difference test (denoted as Tukey test), which is a pairwise comparison between different values of the WSF using the model statistics of RANOVA.

4.5.1 Comparing Different Versions of the Models

We first investigate if a sparse representation on the original feature space can be found across corpora and genders for speech emotion data. We compare the UARs as a function of regularization (L_1 vs. L_2) for the single-task methods (simple and ST), and feature handling (FS vs. FL) for the multi-task methods (MTL, GMTL and MTL-KDG).

We use two RMs to compare: (1) L_1 vs. L_2 using all the experimental results of simple and ST, and (2) FS vs. FL using all the experimental results of MTL, GMTL and MTL-KDG. We use the version of the model as the WSF. For L_1 vs. L_2, the influence of regularization is significant for valence (RANOVA, F(1,84)=4.3, p=0.042), but not for activation. The Tukey test (Figure 4.2) shows that L_2 is significantly better than L_1 for valence (p = 0.042). For FS vs. FL, the influence of feature handling is significant for both valence (F(1,104) = 19.2, p = 2.8e-05) and activation (F(1,104) = 12.8, p = 5.3e-04). FL significantly outperforms FS for valence (Tukey test, p = 2.8e-05) and activation (p = 5.3e-04), as shown in Figure 4.2.

These results indicate that we cannot find a sparse feature representation on the original feature space that transfers well across corpora. The orthogonal projection decouples the original features by "collapsing" similar information onto the same dimension. Therefore, the sparse representation on the new feature space might be

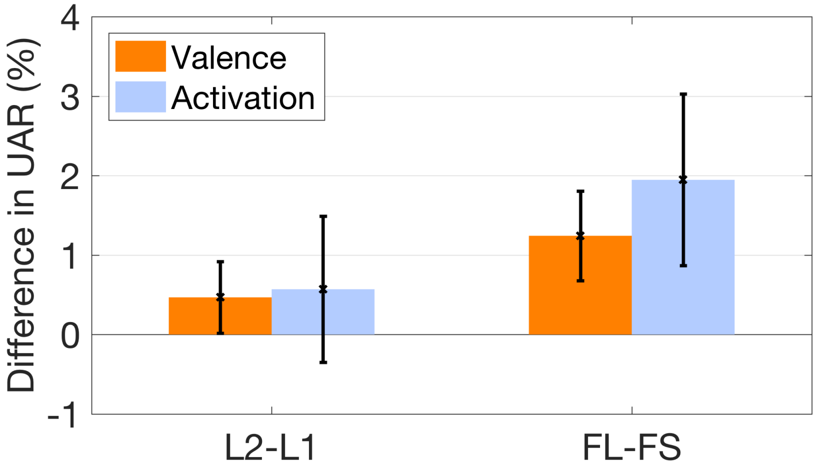


Figure 4.2: Left: the difference in UAR as a function of regularization (L_2-regularization minus L_1-regularization) in the simple and ST models. Right: the same difference as a function of feature handling (feature learning minus feature selection) in MTL, GMTL and MTL-KDG. The black lines represent the 95% confidence interval of the Tukey's honest significant difference test.

able to keep more emotion-related information. In addition, there are high variations in languages, lexical content, speakers and recording conditions within each individual task, and across tasks. As a result, the emotion-related patterns on the original feature space may be further masked. Therefore, we may need to learn a feature space where a sparse representation that generalizes well from the data.

For simplicity, we only present results of the L_2-regularization (simple and ST) and the feature learning setting (MTL, GMTL and MTL-KDG) for the rest of the analyses. Table 4.3 shows the UAR of simple, ST and MTL with a single training corpus (only experiment g has multiple tasks). Table 4.4 and 4.5 shows the UAR of all models when using multiple training corpora, with corpora, genders and corpus-gender pairs as tasks for valence and activation, respectively.

4.5.2 Different Corpus as Training Set

While we treat corpus as a single factor in this experiment, it includes variations in language, type of emotion, in addition to recording condition. We compare the

Test on	Train on	Valence Simple	Valence Task: ST	Valence Gender MTL	Activation Simple	Activation Task: ST	Activation Gender MTL
EmoDB	eNT	56.1	**61.0**	60.7	72.0	78.7	75.5
EmoDB	VAM	48.1	46.3	47.1	86.4	**87.8**	81.1
EmoDB	AVEC	52.3	52.6	52.5	51.5	58.6	54.0
eNT	EmoDB	49.6	48.0	48.9	63.2	65.5	63.7
eNT	VAM	49.3	47.1	47.9	66.2	66.7	**69.8**
eNT	AVEC	54.5	53.3	**56.0**	54.3	62.1	69.5
VAM	EmoDB	50.6	49.4	50.0	67.7	68.0	68.2
VAM	eNT	**59.3**	56.8	56.7	61.7	61.1	58.6
VAM	AVEC	51.4	56.0	53.4	53.4	64.2	**72.2**
AVEC	EmoDB	54.1	52.9	54.0	55.1	55.7	55.2
AVEC	eNT	53.5	53.7	**54.3**	55.9	57.2	56.1
AVEC	VAM	49.9	51.2	50.0	58.8	60.2	**60.7**
Total Avg.		52.4	52.4	52.6	62.2	65.5	65.4

Table 4.3: UAR (%) of valence and activation using a single training corpus. The best result for each test corpus and dimension is bolded. eNT: eNTERFACE.

cross-corpus UAR of classifiers trained on each dataset, and each combination of two datasets, to get some insights on the impact of language and type of emotion (i.e., acted and spontaneous).

When a single corpus is used for training, we find that the models trained on VAM achieve the highest UAR (bolded in Table 4.3) on EmoDB, eNTERFACE and AVEC for activation. We also find when testing on eNTERFACE, models trained on VAM outperform cross-corpus models from the literature (e.g., [219] achieved a maximal UAR of 63.9%, Table 4.6). This is surprising because eNTERFACE is in a different language and with a different type of emotion. There may be several reasons. Firstly, VAM contains recordings about personal and very emotional topics (e.g., paternity questions or affairs) [123], which makes the content more emotionally expressive. Secondly, VAM only contains the speakers evaluated as "very good" and "good", which also results in data that are more emotionally expressive. In addition,

Test	Train on				Simple	Task: Corpus		Task: Gender		Task: Corpus-Gender Pair				
on	EmoDB	eNT	VAM	AVEC		ST	MTL	ST	MTL	ST	MTL	GMTL	MTL-GC	MTL-GG
EmoDB		✓	✓		52.9	57.8	58.4	53.8	55.7	56.8	56.8	59.6	56.0	56.3
EmoDB		✓		✓	59.9	58.3	57.9	58.9	59.9	*60.8*	56.6	56.6	55.0	60.4
EmoDB			✓	✓	54.1	51.8	50.8	53.3	57.2	51.4	50.9	53.0	50.0	49.0
EmoDB		✓	✓	✓	57.0	58.1	58.8	59.5	57.1	57.1	57.9	59.1	56.3	59.8
eNT	✓		✓		49.7	48.4	49.9	50.4	51.3	46.3	48.3	47.4	47.5	48.1
eNT	✓			✓	54.2	50.5	49.9	52.9	54.0	48.4	50.5	50.5	54.0	50.5
eNT			✓	✓	49.0	52.7	52.4	50.8	*60.0*	49.1	56.5	56.7	55.4	55.8
eNT	✓		✓	✓	50.5	50.5	51.3	52.1	57.7	47.0	50.7	49.0	54.3	49.8
VAM	✓	✓			48.2	54.0	51.1	49.6	51.1	52.6	51.5	53.8	52.7	52.1
VAM	✓			✓	48.9	51.0	48.9	54.9	55.3	53.0	49.3	49.3	52.9	49.3
VAM		✓		✓	51.7	56.4	*58.1*	53.2	55.8	56.9	56.9	56.9	55.9	57.5
VAM	✓	✓		✓	51.6	54.2	52.0	53.5	57.3	51.0	52.2	52.4	54.5	52.2
AVEC	✓	✓			54.9	*55.9*	55.2	54.7	55.0	54.9	55.1	55.4	55.3	55.1
AVEC	✓		✓		51.0	53.4	53.9	53.1	54.6	52.9	52.2	53.0	54.0	53.8
AVEC		✓	✓		52.6	53.8	53.3	53.3	54.3	53.1	53.8	53.8	53.4	54.0
AVEC	✓	✓	✓		53.3	54.9	55.7	53.9	54.4	54.7	55.6	55.0	55.2	54.7
	Average				52.5	**53.9**	53.6	53.6 ∗	**55.7** ∗ †	52.9	53.4	**53.9**	**53.9**	53.7

Table 4.4: UAR (%) of valence using multiple training corpora. The best average performance in each experiment is bolded. The overall best performance is underlined. The ∗ (†) indicates that the difference in the mean UAR between the marked model and the simple (ST with same task definition) model is statistically significant when tested using the Tukey's honest significant difference test at 95% confidence level. V: valence; GC/GG: group tasks by corpus/gender; eNT: eNTERFACE.

Test	Train on				Simple	Task: Corpus		Task: Gender		Task: Corpus-Gender Pair				
on	EmoDB	eNT	VAM	AVEC		ST	MTL	ST	MTL	ST	MTL	GMTL	MTL-GC	MTL-GG
EmoDB		✓	✓		74.7	86.2	87.6	84.2	86.0	*89.1*	88.9	87.6	81.7	88.0
EmoDB		✓		✓	56.8	70.8	73.0	58.2	62.9	77.1	75.3	75.3	56.8	70.4
EmoDB			✓	✓	74.7	83.0	79.1	76.9	70.0	85.8	75.7	75.7	70.4	75.7
EmoDB		✓	✓	✓	69.0	85.2	85.6	75.5	79.1	87.2	85.4	85.6	71.2	80.3
eNT	✓		✓		65.7	67.0	68.5	66.4	67.7	67.2	68.2	66.8	69.6	67.2
eNT	✓			✓	59.3	63.4	67.8	62.7	67.1	65.9	68.2	68.1	69.7	67.8
eNT			✓	✓	64.5	66.9	69.9	64.5	69.0	67.8	*70.5*	*70.5*	70.2	*70.5*
eNT	✓		✓	✓	62.8	66.9	69.2	64.0	68.0	67.5	68.3	68.3	70.2	68.6
VAM	✓	✓			62.1	68.2	68.0	65.7	65.2	67.8	67.2	64.6	67.1	67.7
VAM	✓			✓	63.8	68.1	73.1	68.3	74.5	69.3	*74.8*	74.2	73.8	74.6
VAM		✓		✓	60.1	61.0	65.9	62.1	69.6	65.5	68.6	68.6	72.1	73.4
VAM	✓	✓		✓	63.1	68.9	70.7	65.6	72.6	68.9	71.1	71.0	73.6	73.8
AVEC	✓	✓			56.2	56.1	55.7	58.3	56.7	56.8	56.7	55.7	56.4	57.2
AVEC	✓		✓		59.0	56.9	59.2	58.6	59.1	57.8	59.4	58.5	60.5	60.1
AVEC		✓	✓		58.8	59.3	59.9	59.7	60.1	60.2	60.6	61.2	60.6	*61.7*
AVEC	✓	✓	✓		58.8	57.2	58.5	59.1	59.1	58.5	58.2	58.5	60.7	59.5
	Avg. of Activation				63.1	67.8 ∗	**69.5** ∗ **†**	65.6 ∗	**67.9** ∗ **†**	69.5 ∗	**69.8** ∗	69.4 ∗	67.8 ∗	**69.8** ∗

Table 4.5: UAR (%) of activation using multiple training corpora. The best average performance in each experiment is bolded. The overall best performance is underlined. The ∗ (†) indicates that the difference in the mean UAR between the marked model and the simple (ST with same task definition) model is statistically significant when tested using the Tukey's honest significant difference test at 95% confidence level. A: activation; GC/GG: group tasks by corpus/gender; eNT: eNTERFACE.

the speakers in eNTERFACE are from 14 different nations. This may reduce the advantage of training on corpus with the same language, due to the presence of different accents in the testing data.

However, the models trained on VAM do not achieve the highest UAR for valence, even when the test corpus is of the same language (EmoDB) or same type of emotion (AVEC). This may be because it has very unbalanced data for valence, as shown in Table 4.1. These findings suggest that the cross-corpus performance is not only related to the connection in language or type of emotion between training and testing datasets, but is also influenced by other aspects, such as data distribution and quality.

We compare different combinations of training corpora for activation, where the data are more balanced compared to valence. We noticed that the highest UAR of each test corpus (italicized in Table 4.4 and 4.5) is achieved by training on two corpora. Interestingly, for EmoDB, VAM and AVEC, the best training combinations consist of the two corpora that have an aspect in common with the test corpus (i.e., a corpus with same language, and a corpus with the same type of emotion), but do not share common factors between them. The only exception is eNTERFACE, in which the performances of different training combinations are similar. This may be because: (1) we are able to combine knowledge related to language and type of emotion by training on corpora that each share a different common factor with the test corpus; (2) when the training corpora are more dissimilar in language and type of emotion, the common ground between them have higher possibility to be emotion-related. However, when the training corpora have the same language or type of emotion, we may be overfitting to this common factor. Therefore, the classifiers do not generalize well when the factor is different in the test corpus.

4.5.3 The Influence of Model

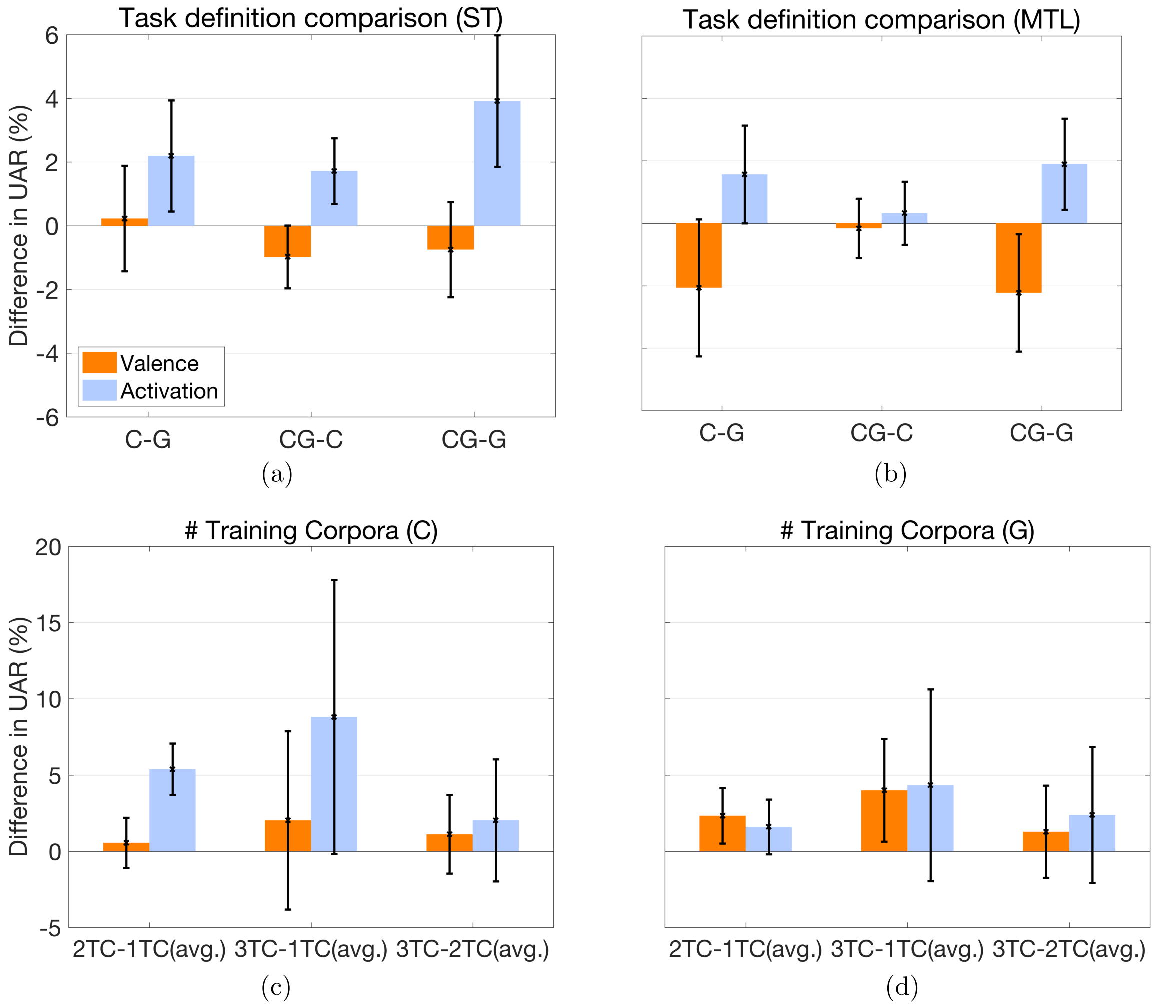
We hypothesize that the influence of model is significant, when task definition is controlled. Specifically, multi-task learning models are better than the simple model and the ST model. We test this hypothesis when corpus, gender or the corpus-gender pair is used to define tasks, respectively, using RMs with model as the WSF.

When corpus is used as the task, the influence of model is significant for activation (RANOVA, F(2,24) = 54.5, p = 1.2e-09), but not for valence. A pairwise comparison for activation shows that MTL significantly outperforms both simple and ST (Tukey test, p=1.0e-05 and 0.016, respectively). This supports the notion that different corpora should be treated as related tasks for the prediction of activation.

When gender is used as the task, we test the influence of model on the results from training on a single dataset (from Table 4.3) and on multiple datasets (from Table 4.4 and 4.5). This is to be consistent with other task definitions (i.e., corpus and corpus-gender pair), where only results using multiple training datasets can be compared. We find that the influence of model is significant for both valence (RANOVA, F(2,24) = 14.3, p = 8.2e-05) and activation (F(2,24) = 16.8, p = 2.8e-05) when training on multiple corpora, but not when training on a single corpus. This may be because we are not capturing the full range of gender variability with only one training corpus. In addition, splitting data by gender for a single corpus may result in insufficient training data per task. The Tukey tests show that when we use multiple training corpora, MTL significantly outperforms both simple and ST for valence (p = 0.0025 and 0.018, respectively) and activation (p = 0.0016 and 0.043, respectively). This reinforces the importance of separating data from different genders, yet still considering the relatedness between them.

When corpus-gender pairs are used as tasks, we find that the influence of model is significant for activation (RANOVA, F(5,60) = 24.3, p = 2.8e-13), but not for valence. The Tukey test for activation shows that all models that explicitly consider the variations in corpus and gender (ST, MTL, GMTL, MTL-GC and MTL-GG) significantly outperform the simple model (p=1.7e-06, 5.0e-04, 0.0015, 7.3e-04 and 1.3e-04 for ST, MTL, GMTL, MTL-GC and MTL-GG vs. simple, respectively). Interestingly, there is no significant difference between ST and other multi-task learning models. This may be because we fuse the results by weighted majority vote over all the tasks. Therefore, we are not only considering the differences in corpus and gender by training task-dependent classifiers, but also utilizing knowledge learned from all the tasks instead of just one. Comparing the multi-task learning methods, we can see that on average, GMTL and MTL-GC perform the best for valence, and MTL and MTL-GG perform the best for activation. However, the differences between the multi-task learning models are very small and not statistically significant, except for between MTL-GC and MTL-GG for activation ($p = 0.019$). We notice that grouping the tasks by corpus or gender generates the highest UAR on several classification tasks (e.g., MTL-GG for valence of VAM when using eNTERFACE and AVEC for training, and for activation of AVEC when using eNTERFACE and VAM for training, MTL-GC for activation of eNTERFACE and AVEC when training on three corpora), but their performances are not stable. For example, the UARs of MTL-GC on the activation of EmoDB are the lowest for all the training corpora combinations, compared to all other models except for the simple model. This may indicate that the closeness between the tasks may be related to the common factors between them, but that the relationship is not guaranteed.

These results support the notion that variations in training corpus, gender, and their interactions all modulate the data. It is beneficial to control for these sources of variation by defining tasks and allowing the tasks to share information using multi-task learning. Improvement in valence is harder to achieve, compared to activation, as found in [154].



(a)

(b)

(c)

(d)

Figure 4.3: Difference in UAR between different experimental conditions (e.g., C-G is the difference between defining the tasks by corpus and gender), along with the 95% confidence interval of the Tukey test, between: different task definitions for (a) ST and (b) MTL; different numbers of training corpora when (c) defining corpus as the task (using MTL) and (d) defining gender as the task (using MTL). Note that for (c) and (d), results with fewer training corpora are averaged across each corpus (1TC) or each combination of training corpora (2TC). MTL with only one training corpus in (c) is the same as simple. C: corpus as the task; G: gender as the task; CG: corpus-gender pairs as the tasks; TC: training corpora.

4.5.4 The Influence of Task Definition

We hypothesize that the way we define the tasks significantly influences the performance of a model. We test this hypothesis using RMs for ST and MTL, respectively, across experiment c, g and cg, with task definition as the WSF.

For ST, the effect of task definition (i.e., corpus, gender, corpus-gender) is significant for activation (RANOVA, F(2,24) = 19.7, p = 8.8e-06), but not for valence. The pairwise Tukey test for ST (Figure 4.3a) suggests that for activation, using either corpora or corpus-gender pairs as tasks is significantly better than using genders as tasks (p = 0.015 and 7.6e-04, respectively) and that the corpus-gender pairs significantly outperform corpora as tasks (p = 0.0021).

For MTL, the impact of task definition is significant for both valence (RANOVA, F(2,24) = 7.1, p = 0.0038) and activation (F(2,24) = 7.8, p = 0.0025). The pairwise comparison for MTL is shown in Figure 4.3b. For valence, gender is a significantly better task-separator than corpus-gender pair (Tukey test, p = 0.021), while for activation, the result is the opposite (p = 0.012). In addition, the advantage of gender over corpus is approaching significance for valence (p = 0.066), and the advantage of corpus over gender as the task is approaching significance for activation (p = 0.05).

The results indicate that defining tasks by gender is the best for valence, while defining a task as a corpus-gender pair is the most beneficial for activation. Interestingly, the benefits of using corpus-gender pairs as tasks in activation is consistent for ST and MTL, but the advantage of using gender as the task in valence only shows in MTL. This suggests that information sharing between genders is important for learning a more robust pattern associated with valence.

4.5.5 Number of Training Corpora

We hypothesize that the number of training corpora (denoted as TC) significantly influences the system performance, when both task definition and model are controlled. Specifically, we hypothesize that adding additional TC is helpful. We test this hypothesis by comparing the performance as the number of TC changes. The model is MTL and the task is either corpus or gender. We build RMs with the number of TC as the WSF for three settings: (a) 2TC vs. 1TC, (b) 3TC vs. 1TC, and (c) 3TC vs. 2TC. The challenge is that each TC size is associated with a different number of results. We compare by averaging over relevant subsets. For example, in the 3TC setting, where we are testing on VAM, the training corpora include EmoDB, eNTERFACE, and AVEC. We compare this result to the 2TC results, still with VAM as a testing corpus. In this case, we take the average performance of systems trained on EmoDB and eNTERFACE, EmoDB and AVEC, and eNTERFACE and AVEC. When comparing to 1TC, we calculate the average obtained by training systems on each of the training corpora, individually. We repeat this over all test corpora. The same comparison applies to 2TC vs. 1TC. Thus, in (a) there are 12 results for each dimension, in (b) and (c) there are four results for each dimension. The comparisons between different numbers of TC are shown in Figure 4.3c (corpus as task) and 4.3d (gender as task).

We find that when corpus is used to define the tasks, the influence of the number of TC is significant for activation (RANOVA for 2TC vs. 1TC, F(1,8) = 53.7, p = 8.2e-05), but not for valence. The Tukey test demonstrates that 2TC is significantly better than 1TC (p = 8.2e-05). The improvements of 3TC over 1TC and 2TC are not significant. However, there are only four results to be compared in these two tests.

When gender is used to define the tasks, the influence of the number of TC is significant for valence (RANOVA, F(1,8) = 8.7, p = 0.018 for 2TC vs. 1TC, F(1,3) = 14.3, p = 0.033 for 3TC vs. 1TC), but not for activation. Both 2TC and 3TC are significantly better than 1TC for valence (p=0.018 and 0.033, respectively). The performance gain of adding a third training corpus to a set already composed of two is not statistically significant.

These findings suggest that the addition of training corpora is helpful, especially given limited variability in the data (e.g., single training corpus). The results also support our earlier findings that gender is a better task-separator than corpus for valence, while corpus is a better task-separator than gender for activation.

4.5.6 Cross-corpus vs. Within-corpus

We present our best cross-corpus UAR and within-corpus LOSO UAR using the simple model in Table 4.6. We compare these results to both the benchmark within-corpus LOSO UAR from [150] and the state-of-the-art cross-validation UAR from the literature (see Table 4.6). We are not able to compare to [153, 154] because the UARs of the individual test datasets are not provided. Note that the number of instances in this work is off by 1 for EmoDB and VAM, and off by 10 for eNTERFACE, compared to [150, 156, 219]. We do not compare the results of EmoDB to [188] because the label matching method is different.

We find that the advantage of within-corpus classification is dominant for datasets with acted emotion (EmoDB and eNTERFACE). A possible explanation is that these acted datasets use fixed lexical content, making emotion recognition much easier. However, cross-corpus classification is effective for datasets with spontaneous emotion. The performance of cross-corpus classification is higher for VAM valence and for AVEC valence and activation. It is slightly lower for VAM activation. Direct comparison between our model and the literature is not possible due to the small differences in data described above and the differences in training datasets. We note that our models achieve comparable results to the state of the art.

4.6 Discussion

In this work, we explore the influence of corpus and gender in emotion recognition. We propose a multi-task learning approach to recognize emotion across corpora, with

Dimension	Setting	From	EmoDB	eNT	VAM	AVEC
Valence	Within	Our Model	84.5	83.4	53.2	53.6
		[150]	87.0	78.7	49.2	-
	Cross	Our Model	61.0	60.0	59.3	55.9
		Literature	-	58.4 [219]	58.6 [156]	-
Activation	Within	Our Model	95.9	84.0	76.1	56.5
		[150]	96.8	78.1	76.5	-
	Cross	Our Model	89.1	70.5	74.8	61.7
		Literature	-	63.9 [219]	71.9 [188]	-

Table 4.6: Within-corpus UAR ("Within") using the simple model and the best cross-corpus UAR ("Cross") from our models (%), and the within-corpus and cross-corpus UAR from literature. eNT: eNTERFACE.

data from multiple datasets as the training set. We present five different models: the simple model, the separate-task model, the multi-task learning model, the group multi-task learning model, and the multi-task learning model with knowledge-driven grouping. These models correspond to five assumptions about the relationship between the tasks: identical, independent, related, partially related and can be grouped based on data similarity, and partially related and can be grouped based on knowledge.

We find that a common sparse feature representation on a transformed feature space is more beneficial, compared to on the original feature space. We assume that this may be due to the higher dimensionality and larger variability in languages, types of emotion, lexical content, speakers and recording conditions.

The best cross-corpus performance with a single training corpus is not always achieved by training on a corpus that shares common language or type of emotion with the test corpus. This may indicate that the quality of the training corpus, in terms of cross-corpus generalizability, is not only related to its similarity with the test corpus, but is also influenced by factors such as class imbalance and the quality of the emotion content. This is inline with Schuller et al. [154]. They found that models trained on VAM produced the best cross-corpus performance on various testing datasets for activation, and that the supreme performance of VAM could be related to the large distance between the positive and negative classes. We also observed that training on corpora that each share common factors with the test corpus, but not with each other, improves activation recognition, in most cases.

My results support that variations in corpus and gender both influence emotion recognition. Overall, separating tasks by these factors and allowing for information sharing between tasks using multi-task learning methods is advantageous. When a single factor is considered, the best performances happen predominantly in cases where we treat the tasks as related, instead of identical or independent. When multiple factors are considered (i.e., corpus and gender), data-driven grouping is comparable to knowledge-driven grouping for corpus and gender.

Comparing different factors, we find that when using multiple datasets for training, separating data based on either corpus or gender, and training emotion classifiers with multi-task learning generates better results, compare to merging all the data together or training independent classifiers. This is inline with the findings in [16] that differences between genders can be as large as the differences between datasets. More specifically, we find that gender is a better task-separator for valence, compared to corpus or corpus-gender pair, while corpus and corpus-gender pair are better task-separators for activation, compared to gender.

The best cross-corpus performance in our experiments is better than or comparable to the within-corpus performance using the baseline method when the test corpus contains spontaneous emotion (VAM and AVEC). our findings may be influenced by the high degree of variability within the spontaneous dataset, which may have reduced the advantage of within-corpus testing.

4.7 Conclusion

In this work, we investigate methods of increasing the generalizability of speech emotion recognition systems, by controlling for two sources of variation: corpus and gender. These factors define the tasks. We use multi-task learning to enable the information sharing across tasks.

In general, defining the tasks by corpus and/or gender, and allowing for information sharing across tasks is beneficial. Defining tasks by corpus or both corpus and gender is better than by gender for activation predictions, while gender is the best task-separator for valence predictions. When multiple factors are used to define the tasks, data-driven grouping performs comparably to knowledge-driven grouping. On average, the system performance increases with the number of training corpora.

CHAPTER 5

Exploiting the Acoustic and Lexical Properties of Phonemes

5.1 Introduction

Emotions modulate acoustic signals both explicitly, through paralinguistic characteristics (e.g., the tone and tempo of speech), and implicitly, through the alteration of the content of speech. Therefore, speech content is a double-edged sword in emotion recognition: the variability it introduces makes it harder to distill emotion-related cues from the acoustic signals, yet the content itself is also reflective of emotion. In this work, we explicitly consider both roles of speech content and demonstrate that, in so doing, we are able to make more accurate predictions of emotion.

We present a speech emotion recognition system that: (1) considers acoustic variability in terms of both emotion and the underlying speech content, here defined as sequences of phonemes, and (2) directly leverages the connection between emotion and phoneme sequences. We investigate whether the additional phonetic information leads to improved performance. We concentrate on predicting valence (the positive vs. negative aspect of an emotional display [134, 136]) because it has been shown to be difficult given only acoustic signals [198]. We use soft labels, which incorporate the variability in emotion perception.

Previous research has investigated how phonemes modulate acoustics together with emotion by exploring phoneme-level emotion classification methods [23, 85, 188], or designing acoustic features [14, 65, 67, 140, 189] or labels incorporate phonetic knowledge [60]. The results of these studies showed that phonemes vary in how they are modulated by emotion and that features designed based on phonetic knowledge work well in emotion recognition. Recent works have shown that emotion can be predicted directly from sequences of phonemes without acoustic information, by modeling phoneme sequences like word sequences [54], using LSTM networks [53], or multi-channel CNN networks [206]. These works have also shown that combining utterance-level phonetic and acoustic representations brings further improvement. However, work that considers both the phonetic modulation of acoustics and the link between phoneme sequences and emotions is still missing. In addition, we do not yet know how models that exploit the acoustic and/or phonetic contributions of phonemes is influenced by elicitation method (i.e., fixed, improvised under targeted scene, spontaneous).

In this work, we seek to improve valence prediction by leveraging the dual-functionality of phonemes, using Convolutional Neural Networks with global pooling (*Conv-Pool*). We hypothesize that adding phonetic information at different stages has different effects and that we can exploit both the acoustic and the lexical contributions using a *multi-stage fusion* model that combines acoustic and phonetic information at both feature-level (*feature fusion, FF*) and utterance-level (*intermediate fusion, IF*). We investigate how models leveraging phonetic information at different stages are influenced by the emotion elicitation process (e.g., fixed script, improvisation, natural interaction) of the data. We test our hypothesis on the IEMOCAP dataset [25] and the MSP-Improv dataset [26].

Our results show that our multi-stage fusion model outperforms both FF and IF models, especially on data produced using improvisations and natural interactions.

We also find that both FF and IF are beneficial compared to unimodal models, and that IF outperforms FF. However, the advantage of modeling phoneme sequences independently, either in the unimodal phonetic model or IF, decreases as the lexical content becomes more spontaneous, indicating that this advantage may come from memorizing emotionally salient patterns in speech content. The novelty of this paper includes the presentation of: (1) a multi-stage fusion method that exploits the dual-functionality of phonemes and (2) an investigation into the influence of the type of lexical content on the performance of the models leveraging different functions of phonemes.

5.2 Data

5.2.1 Datasets and Transcriptions

We use two English dyadic emotion datasets in this work: IEMOCAP and MSP-Improv. We choose these two datasets because: (1) their sizes allow us to train neural networks; (2) they provide evaluations of valence; (3) they contain varying lexical patterns due to the use of different emotion elicitation methods, allowing us to conduct relevant analyses.

We conduct three experiments on IEMOCAP: over the entire dataset (*IEMOCAP-all*), on only the scripted utterances (*IEMOCAP-scripted*, 5,255), and on only the improvised utterances (*IEMOCAP-improv*, 4,784). We use the provided manual transcriptions for forced alignment (explained in Section 5.2.4) and exclude six utterances with no matching phonemes information.

For MSP-Improv, we focus on two types of data: speaker turns from improvised scenes (excluding the target sentences), and the natural interactions during preparation. Because the transcriptions are not provided, we use automatic transcriptions produced by the Microsoft Azure - Bing Speech API[1], provided by the creator of the dataset. We only use utterances that have transcriptions in our experiments. This decreased our data to 5,650 utterances, which we refer to as *MSP-I+N*. We choose to exclude target sentences and not to perform experiments for the improvised and natural partitions separately due to the limited size of the partitions.

5.2.2 Labels

We convert the 5-point ratings into three categories: negative, neutral, and positive, and generate fuzzy labels for each utterance as in [4, 30]. We represent each evaluation as a 3-dimensional one-hot vector by keeping 3 as “neutral” and merging 1-2 and 4-5 as “negative” and “positive”, respectively. We then use the mean over the evaluations for each utterance as the ground truth. For instance, given an utterance with three evaluations, 3, 4, and 5, we first convert the evaluations to [0, 1, 0], [0, 0, 1], and [0, 0, 1], respectively. After taking the mean, the ground truth label for this utterance is [0, 1/3, 2/3]. In this way, we form the problem of valence recognition as a three-way classification task.

5.2.3 Acoustic Features

We extract 40-dimensional log Mel-frequency Filterbank energy (MFB) using Kaldi [124]. The MFBs are computed over frames of 25ms, with a step size of 10ms, as in [3, 4, 212]. We perform speaker-dependent z-normalization at the frame-level.

5.2.4 Phonemes

For each utterance, we acquire the start and end time of each phoneme by using forced alignment between the audio and the manual (IEMOCAP) or automatic (MSP-Improv) transcriptions. We use Gentle [2], a Kaldi-based forced aligner. It identifies 39

[1] https://azure.microsoft.com/en-us/services/cognitive-services/speech/

[2] https://lowerquality.com/gentle/

unique phonemes from both IEMOCAP and MSP-Improv, with an additional "out of vocabulary" label for unrecognized sounds, resulting in a 40-dimensional one-hot vector for each phoneme. The phonetic representations are used in two different ways: (1) independently without repetition, and (2) repeated and with fixed step-size to be aligned with acoustic features. A more detailed description can be found in Section 5.3.1.

5.3 Methodology

5.3.1 Network Structures

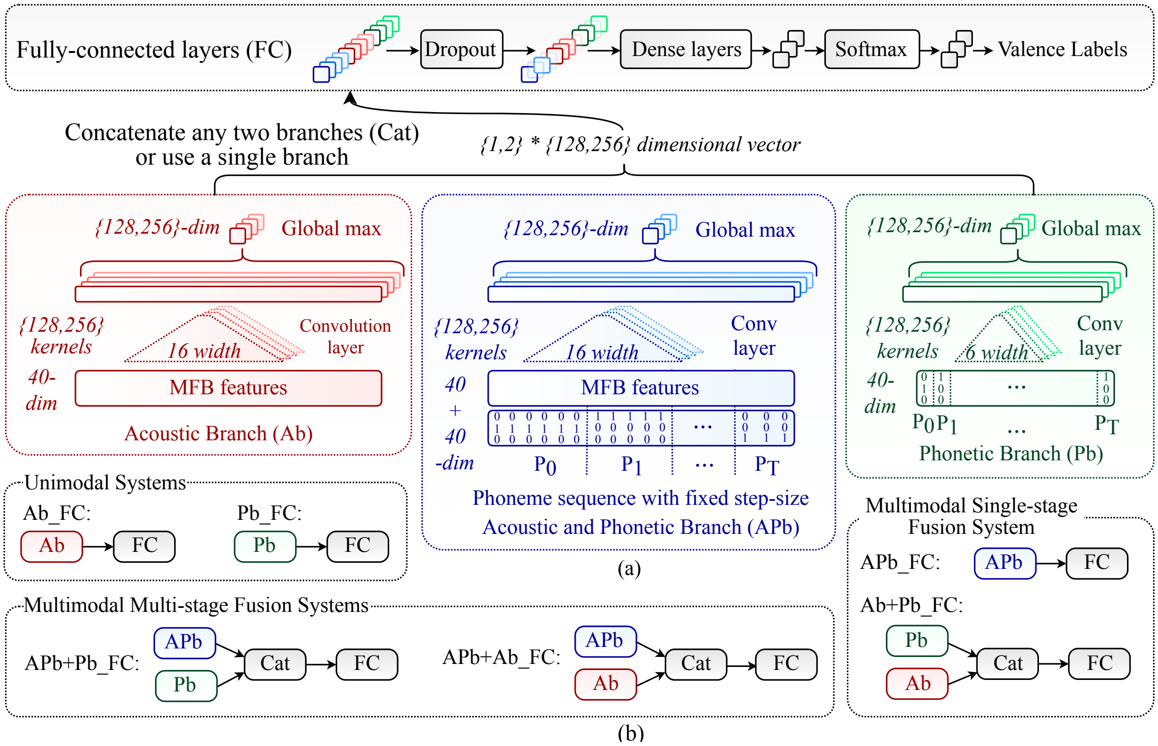
We design our models based on the temporal Convolutional Neural Networks with global pooling (*Conv-Pool*) structure. The Conv-Pool structure was first proposed by Yoon Kim [80] for sentence classification. Aldeneh et al. [3] applied this architecture to categorical emotion recognition using the acoustic modality and achieved the state-of-the-art performance.

Figure 5.1 shows the architectures of our networks. These networks consist of the following components (Figure 5.1(a)):

- A Conv-Pool sub-network (i.e., a 1D convolutional layer over time followed by a global max-pooling layer). The convolutional layer extracts a sequence of high-level features that specify emotionally salient regions of the input signal [3]. The global max-pooling layer summarizes the entire utterances for each dimension and generates a fixed length representation.

- A concatenation of the multiple utterance-level representations (if more than one, denoted as "+").

- An optional dropout layer, two fully-connected layers and a softmax layer (denoted as (*FC*)).



(a)

(b)

Figure 5.1: (a) A general network that illustrates all the components, including: the acoustic branch (*Ab*), the phonetic branch (*Pb*), the acoustic and phonetic branch (*APb*) that combines the features of the two modalities; the concatenation of the utterance-level representations (*Cat*), and a stack of dropout, fully-connected and softmax layers (*FC*). (b) Architectures for all models.

There are three Conv-Pool branches in our networks: the unimodal acoustic branch (*Ab*), the unimodal phonetic branch (*Pb*), and the feature-fusion branch (*APb*). *Ab* and *Pb* operate on variable-length MFB features and phoneme sequences. In *APb*, we aim to capture the phonetic modulations of acoustic features. We concatenate the phoneme label with the MFBs at each frame. For example, if a specific phoneme lasts 0.1 seconds, the same one-hot vector is repeated ten times and concatenated with the MFB features of the ten corresponding frames. For audio frames with no matching phoneme, a zero-vector is used instead. The number of input channels of the convolutional layer is 40, 40, and 80 for *Ab*, *Pb*, and *APb*, respectively. Feeding the output of a single branch to the *FC* sub-network results in three models

(Figure 5.1(b)): two unimodal models (i.e., *Ab_FC* and *Pb_FC*), and a multimodal single-stage feature-fusion model (*APb_FC*).

We concatenate the outputs from Ab and Pb for joint modeling in FC to captures the high-level interaction between the learned representations of acoustics and speech content. This results in our multimodal single-stage fusion model using intermediate-fusion, *Ab+Pb_FC* (Figure 5.1(b)).

We hypothesize feature fusion and intermediate fusion play different roles in the network. Feature fusion allows our network to capture how phonemes modulate acoustics. However, it may not be effective in linking speech content and emotional state, specifically, in extracting phoneme sequences that are informative identifiers of valence. This is because: (1) each single phoneme may be repeated several times in order to have the same step-size with the MFBs, resulting in insufficient temporal context for the phoneme sequences in the convolution layer; (2) the input phoneme sequences are much more sparse than the MFBs, resulting in representations dominated by acoustic information. On the other hand, intermediate fusion can more efficiently leverage the complementary emotionally salient information learned from audio and phoneme sequences. Because of the dual-functionality of phonemes, we propose to combine them into a multi-stage fusion model to exploit the advantages of both techniques. This network, *APb+Pb_FC* , concatenates the representation from *APb* with the phonetic branch Pb and jointly model them in *FC*. In addition, we explore another multi-stage fusion network, *APb+Ab_FC*, which concatenates *APb* and *Ab* for comparison. Both *APb+Pb_FC* and *APb+Ab_FC* are shown in Figure 5.1(b).

5.3.2 Hyper-parameters and Training Strategy

We use Rectified Linear Units (ReLU) as the nonlinear activation function in all layers, except for the output layer, where softmax is used. We select the layer size of the convolutional and fully-connected layers from {128,256} as in [212]. The layer size is kept consistent throughout each model. We fix the kernel width to 16 for MFB input, which is shown to perform well on both IEMOCAP and MSP-Improv in [3]. For the phoneme sequence input, we select a kernel width of 6, based on the average number of phonemes (6.38) per English word, according to the CMU pronunciation dictionary [3]. Besides, we incorporate an optional dropout layer after the global max-pooling to improve generalization of the networks. The dropout probability is selected from {0, 0.2, 0.5}, where 0 corresponds to no dropout, and 0.2 and 0.5 are from the suggested range in [170].

We implemented the models using PyTorch [4] version 0.2.0. The loss function is cross-entropy computed using the fuzzy ground truth labels. We weigh the three classes using $N/(3 * \sum_{j=1}^{N} gt_j^i)$ in the loss calculation, where N is the total number of training utterances, gt_j^i is the value at position i in the fuzzy ground truth label for utterance j. We train the models using a learning rate of 0.0001 with the Adam optimizer [81].

We use Unweighted Average Recall (UAR) as the performance measure due to unbalanced data [131]. When the ground truth labels have ties, we deem predictions for any of the tied positions as correct, as in [4]. For instance, when the ground truth is [0.5, 0.5, 0], prediction of either 0 or 1 are correct. As a result, the chance performance of making predictions uniformly at random is higher than 33.33%.

We use the leave-one-speaker-out evaluation setting for our experiment. Both IEMOCAP and MSP-Improv are organized by sessions. At each round, we left out data from a single speaker as the test set, and use data from the other speaker in the same session for validation. The data from remaining sessions are used for training. We run each experiment five times to reduce performance fluctuation. For each training-validation-testing combination, we select the number of training epoch

[3] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[4] http://pytorch.org

($\in [1, 30]$) by maximizing the validation UAR for each run separately and select the layer size and dropout probability by maximizing the validation UAR averaged over five runs. We report the test UAR corresponding to the selected hyper-parameters, averaged over speakers and runs. We set the batch size to 100, and zero-pad the features to the maximum length of each batch.

5.4 Results and Discussion

We present the UAR of all the models for the experiments on IEMOCAP-all, IEMOCAP-scripted, IEMOCAP-improv, and MSP-I+N in Table 5.1, together with the chance performance calculated by making predictions uniformly at random. For the results of each experiment, we first test if the influence of model is significant by using a repeated-measure ANOVA (RANOVA). We treat the per-speaker performance as the "subject" and model as the within-subject factor. We report the statistics in Table 5.1. We find that the influence of model is significant in all experiments when asserting significance at $p<0.05$, even with the lower bound correction. We compare pairs of models across experiments to understand the effect of each approach and the influence of the type of lexical content. We use Tukey's honest test based on the RANOVA model for these pairwise comparisons and assert significance at $p<0.05$.

5.4.1 Unimodal Results

The Conv-Pool structure has been shown to work well for acoustic emotion recognition [3, 212]. In our experiments, the *Ab_FC* model achieves UARs that are much higher than chance in all experiments.

The fluctuations in the results of *Pb_FC* across different types of lexical content are much higher than those of *Ab_FC*. We find that *Pb_FC* significantly outperforms *Ab_FC* on IEMOCAP-all ($p = 8.25e\text{-}3$) and IEMOCAP-scripted ($p = 5.64e\text{-}5$), while *Ab_FC* significantly outperforms *Pb_FC* on MSP-I+N ($p = 0.0444$).

Model	IEMOCAP	IEMOCAP	IEMOCAP	MSP
	-all	-scripted	-improv	-I+N
Chance	45.40	46.91	44.55	36.09
Ab_FC	64.04	61.18	65.00	51.84†
Pb_FC	69.18∗	**78.42**∗⋄	62.50	47.54
APb_FC	67.17∗	67.21∗	67.68∗†	53.98∗†
Ab+Pb_FC	73.33∗ † ⋄	75.09∗⋄	69.13∗†	54.99∗†
APb+Pb_FC	**73.79**∗ † ⋄	75.34∗⋄	**70.05**∗ † ⋄	**55.98**∗ † ⋄
APb+Ab_FC	67.09∗	65.54∗	67.44∗	54.34∗†
$F(5, 45/55)$	70.3	55.4	19.6	25.6
p_{LB}	1.52e-5	3.92e-5	1.66e-3	3.67e-4

Table 5.1: The UAR of all the models and the statistics of RANOVA (F and p_{LB}) for the influence of model. The best result in each experiment is bolded. $F(5, 45)$ and $F(5, 55)$ are for experiments on IEMOCAP and MSP-I+N, respectively. p_{LB} is the p-value with lower bound correction. $*$, †, and $\diamond$ represent that the marked model significantly outperforms *Ab_FC*, *Pb_FC*, and *APb_FC*, respectively, using Tukey's honest test and asserting significance at p<0.05.

It is clear that *Pb_FC* performs better than *Ab_FC* when all the data or a large portion of the data are scripted, while the opposite is true when there is less control on the lexical content of the data (i.e., improvisations and natural interactions). *Pb_FC* achieved the highest performance among all models on IEMOCAP-scripted. It is interesting to see that when emotion-related scripted data are repeated across training, validation, and testing sets, additional information from the acoustic modality brings more harm than good. This indicates that Conv-Pool with phoneme sequence can learn and memorize speech-content-related patterns that are strongly associated with emotion classes, but does not work as well as acoustics on unscripted/natural data.

5.4.2 Single-stage Fusion Results

We compare the feature-fusion model (*APb_FC*) with each of the unimodal models, respectively. We find that *APb_FC* achieves significant performance gain over

Ab_FC in all four experiments (p = 1.42e-3, 3.43e-4, 0.0224, and 0.0232 for IEMOCAP-all, IEMOCAP-scripted, IEMOCAP-improv, MSP-I+N, respectively). However, *APb_FC* only significantly outperforms *Pb_FC* on IEMOCAP-improv (p = 0.0363) and MSP-I+N (p = 4.86e-3), while shows significant performance loss on IEMOCAP-scripted (p = 9.84e-4). In addition, the performance of *APb_FC* is very stable across the different portion of IEMOCAP. These results support our hypothesis that in feature fusion, the phonetic information can work as a "guide" for learning emotion-salient acoustic representations, but cannot effectively capture the patterns in speech content that are related to emotion.

The intermediate-fusion model (*Ab+Pb_FC*), on the other hand, shows consistent improvement compared to both *Ab_FC* and *Pb_FC*, with the only exception of *Pb_FC* on IEMOCAP-scripted. The differences between *Ab+Pb_FC* and *Ab_FC* are significant in all experiments (p = 9.76e-6, 1.47e-4, 4.62e-4, and 4.38e-3 for IEMOCAP-all, IEMOCAP-scripted, IEMOCAP-improv, and MSP-I+N, respectively). *Ab+Pb_FC* also significantly outperforms *Pb_FC* for IEMOCAP-all (p = 7.16e-4), IEMOCAP-improv (p = 3.04e-3), and MSP-I+N (p = 6.05e-4). This indicates that there is complementary information from representations learned separately from the audio and phoneme modalities. This is in line with the results reported in [53, 54].

Comparing *APb_FC* with Ab+Pb_FC, we find that the advantage of intermediate-level fusion decreases with the flexibility of the lexical content. *Ab+Pb_FC* significantly outperforms *APb_FC* by on IEMOCAP-scripted (p = 8.76e-4) and IEMOCAP-all (p = 8.25e-6), but is only comparable to *APb_FC* on IEMOCAP-improv and MSP-I+N. This presents additional evidence that the memorization of patterns in phoneme sequences is most beneficial when the elicitation relies upon scripts. This suggests that there are multiple causes behind the improvements over the unimodal models, via feature-fusion and intermediate-fusion, and that we may achieve further performance gain by combining them using multi-stage fusion.

5.4.3 Multi-stage Fusion Results

We compare our proposed multi-stage fusion model, *APb+Pb_FC*, with each of the single-stage fusion models. This model aims to exploit the double functionality of phonemes. We find that*APb+Pb_FC* significantly outperforms *APb_FC* in all four experiments (p = 1.80e-6, 1.12e-3, 4.29e-3, and 0.0493 for IEMOCAP-all, IEMOCAP-scripted, IEMOCAP-improv, and MSP-I+N, respectively).*APb+Pb_FC* also shows consistent performance improvement over *Ab+Pb_FC*, and the advantage is larger on data with less control over the lexical content (i.e., IEMOCAP-improv and MSP-I+N). This result supports our hypothesis that the consideration of both the phonetic modulation of acoustics and the connection between phoneme sequences and emotions allows us to improve the performance of valence prediction.

We investigate the performance of another multi-stage fusion model, *APb+Ab_FC* , which merges the outputs of the feature fusion branch and the unimodal acoustic branch. We find that *APb+Ab_FC* is comparable to *APb_FC* in all experiments, and significantly outperformed by *Ab+Pb_FC* on IEMOCAP-all (p = 6.44e-6) and IEMOCAP-scripted (p = 5.19e-4). The fact that repeatedly adding the acoustic modality does not improve performance is in line with our hypothesis that the learned representation from fused acoustic and phonetic features is dominated by the audio modality.

We compare our best UAR with the state-of-the-art result using the same label processing, training-validation-testing folds, and evaluation method [4]. We find that*APb+Pb_FC* outperforms the intermediate-fusion of the acoustic and lexical modalities using outer-product in [4] by 4.4% in UAR on IEMOCAP-all. This further demonstrates the effectiveness of our method. We note, however, that we cannot attribute the performance gain completely to the use of phoneme sequences and multi-stage fusion. The differences in network structure (e.g., Conv-Pool vs. GRU, the use of dropout), hyper-parameters (e.g., layer size, kernel width), and training paradigm all have important influence on the final results.

5.5 Conclusions

In this work, we explore the impact of incorporating phonetic knowledge into acoustic valence recognition. We propose to repeatedly add phonetic features, at both feature-level and utterance-level, into a single temporal convolutional neural network. We show that this multi-stage fusion model outperforms all other models on IEMOCAP-all, IEMOCAP-improv, and MSP-I+N, even when the transcriptions are estimated using ASR systems (i.e., MSP). The gain over the most accurate network that fuses acoustic and phonetic information at a single stage is the greatest given improvised and natural interactions. This demonstrates efficacy of this approach given imperfect transcriptions and speech data that are collected without reliance upon a script. Finally, the proposed system outperforms the state-of-the-art approach from the literature.

Our results also show that the phonetic branch helps the network leverage the direct link between emotion and speech content contained in phoneme sequences. Feature fusion can capture the phonetic modulation of acoustics, but the resulting representation is dominated by the acoustic modality. The advantage of intermediate fusion over feature fusion decreases when the lexical content becomes more spontaneous. These findings support our assumption that feature fusion and intermediate fusion exploit acoustic and lexical contributions of phonemes, respectively. Future work will explore the feasibility of performing integrated phone recognition coupled with emotion recognition.

Part II

Label Variability

CHAPTER 6

Jointly Modeling Self-report and Perceived Emotion

6.1 Introduction

Expressions of emotion convey information about the underlying state of an individual. This information can be partially masked either intentionally or unintentionally, leading to variability in the labels associated with emotional displays. This variability in the label space is one of the main differences between emotion recognition and other machine learning tasks. As a result, emotion labeling experiments must clearly identify the purpose of a given set of labels: will the labels capture the *felt sense* of the individual who produces the emotion, will they instead capture how that person perceives his/her own emotional display (*self-report*), or will they instead capture how others perceive the display (*perceived label*)? Emotion recognition systems have focused only on a single type of label traditionally, rather than leveraging the potentially complementary information conveyed by the separate strategies. This work explores the impact of jointly modeling both an individual's self-report and the perceived labels of others.

Individuals differ in their ability to convey emotion. Therefore, the patterns of emotion expression can vary across individuals, resulting in difficulty in transferring models learned from a set of speakers to a new speaker when using self-reported emotion, as observed in [180]. The emotion labels provided by others can act as stabilizers to reduce fluctuations caused by individual differences. On the other hand, the varying patterns (e.g., intensity of cues) of emotion expression can result in different levels of difficulty for observers, as found in [73]. The emotion labels provided by the speakers themselves can work as a stabilizer to explain how a single individual expresses a range of emotions. Therefore, our motivating hypothesis is that controlling for the manner in which others perceive emotion and how one perceives one's own emotion will lead to improvement in both tasks. In addition, we hypothesize that we can get complementary improvement by better capturing the complexity inherent in the interactions between multimodal cues. We ask the following research questions: (1) can joint modeling lead to better performance across both types of labels; (2) can the same performance gain be achieved through complex feature learning; and (3) is the performance gain from joint modeling and complex feature learning additive?

We conduct an experiment on the IEMOCAP dataset [25] using a subset that contains both perceived and self-report labels. We construct the emotion recognition problem as binary one-against-rest classifications to account for the fact that an utterance can be labeled with multiple emotions. We use linear support vector machines (SVMs) as the baseline method. We propose a multi-task learning method that jointly models self-report and perceived emotion (each label type is a task). We also explore the influence of non-linear feature learning using deep belief networks (DBNs). Finally, we analyze the combined impact of both components.

The experimental results suggest that joint modeling is able to utilize the complementary knowledge presented in both self-report and perceived labels and that the combination of non-linear feature learning and joint modeling results in more effective emotion recognition systems. The novelty of this work includes: (1) the first attempt to jointly learn self-reported and perceived emotion; (2) an exploration of the influence of feature learning, using DBN feature pretraining, on multi-task learning.

6.2 Data

6.2.1 Labels

We experiment on the IEMOCAP dataset [25] using the categorical labels. We merge the classes of happiness and excitement as in [105]. The subset of IEMOCAP with matched audio and mocap data (see Chapter 3), which we refer to as the *original data*, contains 5,042 utterances.

A subset of the *original data* have self-reported emotion labels and matched audio and mocap data, as described in Chapter 3. This subset contains 1,184 utterances. The data in this subset have two labels: (1) perceived emotion and (2) self-reported emotion. The perceived emotion labels are a vector that describe the emotions perceived by the evaluators. We define the perceived emotion ground truth as any emotion label noted by at least two (out of three) evaluators. For example, the perceptual evaluations of three evaluators for $utterance_i$ may be distributed as [2 0 0 2 1 0 ... 0], where two evaluators noted anger, two noted sadness, and one noted frustration (evaluators were not restricted to the number of emotions selected). Therefore, $utterance_i$ would be associated with the perceived emotions of anger and sadness.

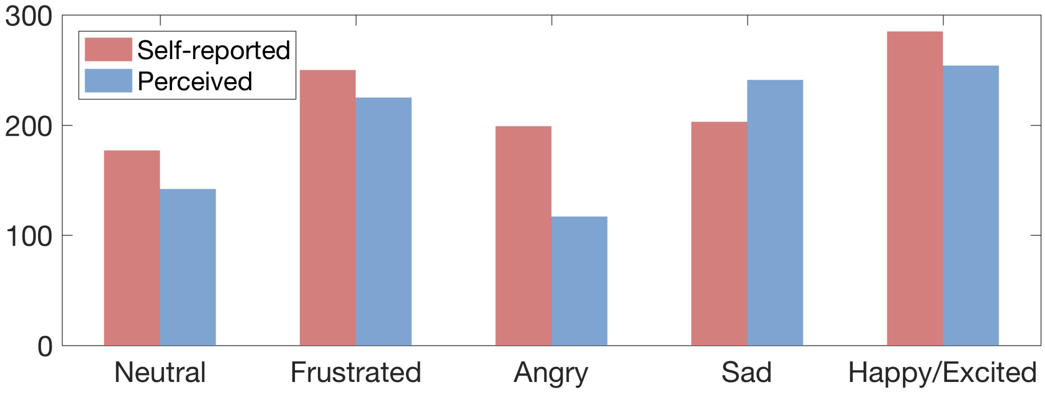


Figure 6.1: The number of utterances in each emotion.

The perceived ground truth label for each emotion is a binary vector that describes the presence of each label (e.g., for the example above the final label would be [1 0 0 1 0 ... 0]). The self-report label is also a binary vector that marks the presence or absence of a given label. We downsample the subset with self-evaluation labels to include only utterances with at least one perceived emotion and one self-reported emotion from the set of {angry, happy/excited, neutral, sad, and frustrated}. This results in 967 utterances. We refer to this data as the *self-evaluation subset*. On average, each utterance in the *self-evaluation subset* has 1.15 ± 0.39 self-reported emotion labels and 1.01 ± 0.11 perceived emotion labels. Figure 6.1 shows the distribution of self-reported emotion and perceived emotion. There are differences between the two distributions, notably for the class of anger. We compute the Hamming similarity between the two types of emotion, defined as the proportion of instances that have the same label in self-report and perceived emotions, given an emotion class. The similarity for neutral, frustrated, angry, sad, and happy/excited are 0.87, 0.83, 0.88, 0.91 and 0.90, respectively, and the average over all classes is 0.88.

We use both the *original* and *self-evaluation* sets of data. The *original* data are used for unsupervised feature learning (see details in Section 6.3.1). The *self-evaluation* data are used for supervised classification (see details in Section 6.3.2 - 6.3.3).

6.2.2 Features

6.2.2.1 Acoustic Features

We extract the Interspeech 2009 Emotion Challenge features [148] using openSMILE [48]. We use a relatively small feature set due to the limited size of the data. The feature set contains 16 frame-level Low-Level Descriptors (LLDs), including zero-crossing-rate, root mean square energy, pitch frequency, harmonics-to-noise ratio, and Mel-Frequency Cepstral Coefficients (MFCC) 1-12. Twelve statistics are applied to

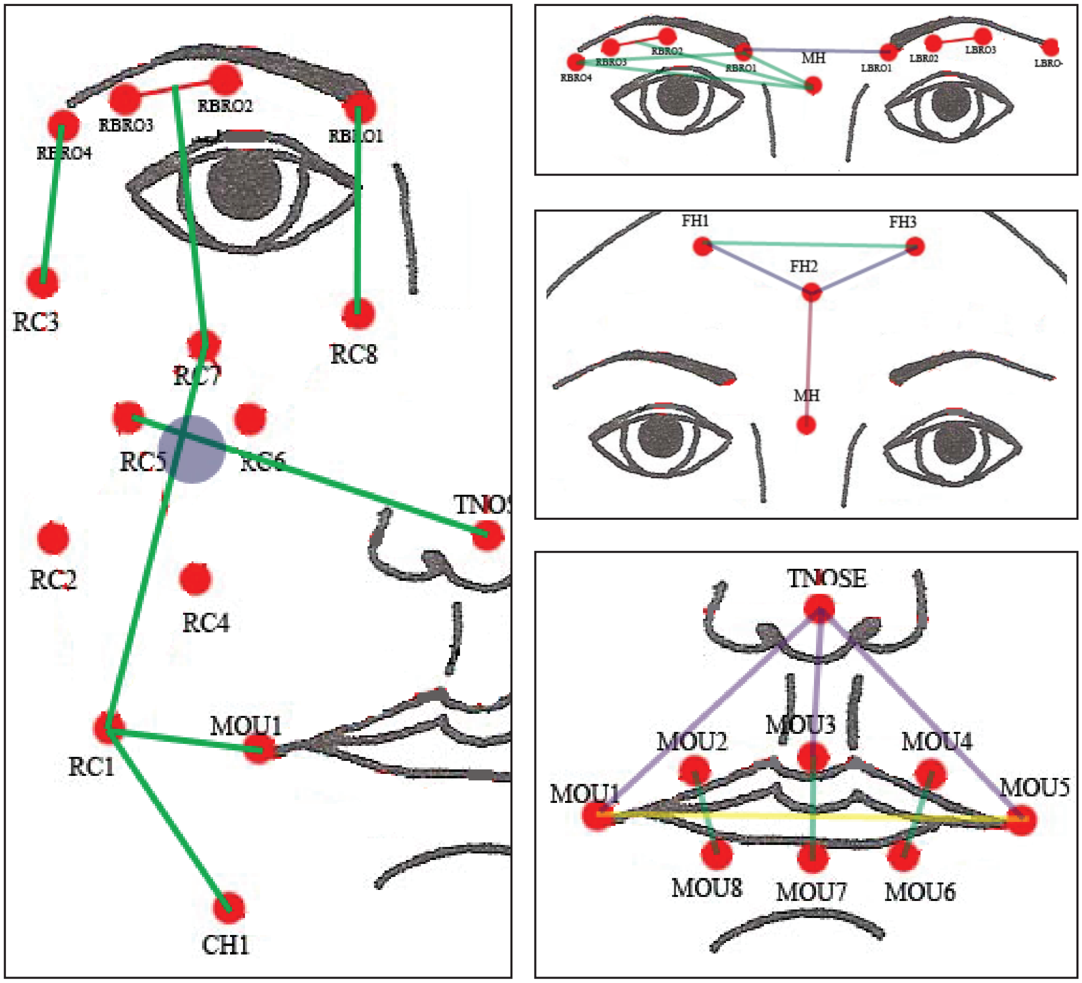


Figure 6.2: The positions of the markers and the distance features (only shown on right side of face). Image courtesy Mower, Mataric, and Narayanan [105] ©2011 IEEE, reprinted, with permission. Each marker position is represented in three-dimensional vector coordinates (x, y, z).

the frame-level LLDs and the first-order delta coefficient of the LLDs to generate the 384 utterance-level features. The statistics are: max, min, range, the position of the maximum and minimum value, arithmetic mean, standard deviation, the slope and onset of the linear approximation of the contours, quadratic error (between actual contour and the linear approximation), skewness and kurtosis.

6.2.2.2 Visual Features

We extract visual features using the 3D motion-capture markers. The mocap features are the Euclidean distances between the (x, y, z) coordinates of the markers. The positions of the markers and the distances calculated are shown in Figure 6.2. These features were introduced [105] and used in [78, 105]. They capture movements associated with emotional facial expressions. For example, the distance between TNOSE and MOU1/MOU5 changes as a function of smiles and frowns. We apply five statistics to the frame-level distance features, including mean, variance, quantile maximum, quantile minimum and quantile range. This results in 540 utterance-level features. We exclude missing data in the utterance-level calculations.

We perform speaker-dependent z-normalization on each feature. The normalization is applied separately for the *original* and *self-evaluated* data. In this way, the initial input for classification is identical for models that do and do not use feature learning, allowing for a more direct comparison.

6.3 Methodology

6.3.1 Feature Learning

We use the pretraining of deep belief networks (DBNs) [61] for feature learning. DBNs are formed by stacking Restricted Boltzmann Machines (RBMs) [164], which are undirected neural networks that only have inter-layer connections. RBMs learn the posterior probability of the output (often binary, referred to as "hidden units") given the inputs (binary or Gaussian, referred to as "visible units"). We select DBN feature learning because: (1) it can capture complex non-linear interactions between features; (2) its unsupervised nature makes the learned features task-independent; (3) it has been shown to be effective for reducing dimension and can outperform traditional feature selection methods, such as Information Gain and Principal Feature

Analysis [78].

We train three DBN models on the *original* data, one each for audio, mocap, and both modalities, using the implementation in [174]. We set the number of hidden layers to 3, as in [78]. We choose Gaussian-Bernoulli RBM (GBRBM) as the first layer, since it takes Gaussian visible units, and is suitable for the real-valued features. The second and third layers of the models are Bernoulli-Bernoulli RBMs (BBRBMs), where both the visible units and hidden units are binary. It is suggested in [13] that it is often more beneficial to have an over-complete first layer (i.e. number of hidden units > number of visible units), compared to an under-complete first layer (i.e. number of hidden units < number of visible units). In addition, previous work [83] found that networks that have the same number of hidden units for each layer generally outperform networks that have increasing or decreasing numbers of hidden units at each layer. We use these insights and set the number of hidden units in the first and second layer to be approximately 1.5 times over-complete of the original input features. We decreased the size of the final layer to be in line with prior work on this dataset [78]. The number of units for each layer are shown in Table 6.1. The number of units in the final layer is selected in cross-validation (Section 6.3.3). We fix the size of the mini-batches to 32, according to [13]. and set the learning rate to 0.004 for the GBRBM layer and 0.02 for other BBRBM layers based on empirical re-construction error.

Modality	Input	Layer 1	Layer 2	Layer 3
Audio	384	600	600	{100,200}
Mocap	540	800	800	{200,300}
Combined	924	1400	1400	{300,400,500}

Table 6.1: Number of units in each layer of the DBN models for audio, mocap and combined features.

6.3.2 Classification Models

We form the recognition of neutral, frustrated, angry, sad and happy/excited as five one-against-rest binary classification problems and train five separate models. This is because each utterance can be labeled with multiple emotions.

The main question we ask in this work is: can jointly modeling self-reported and perceived emotion lead to better performance for both types of emotions? Therefore, we propose two approaches: independent modeling (denoted as IM) and joint modeling (denoted as JM). The models are compared on the *self-evaluation* data. In IM, we train separate classifiers, one each for self-report and perceived emotion. We use linear Support Vector Machine (SVM) for the IM baseline. When training on the original features, we adopt L_1-regularization to serve as a "built-in" feature selection in addition to the commonly used L_2-regularization, since it can enforce sparsity of the features. We weight the cost of error in the positive class and negative class differently during training to deal with unbalanced data, as suggested in [182]. The per-class weight is calculated by the reciprocal of the proportion of that class in the training data.

In JM, we model self-report and perceived emotion in a single classifier using multi-task learning, with each emotion type as a task. We use the multi-task feature learning (MTFL) algorithm of [8, 9]. This method is based on the hypothesis that task 1 through T share a common feature representation. Therefore, the weight vectors w_1 through w_T for the tasks can be jointly learned. The weight matrix W, defined as $[w_1, w_2, ..., w_T]$, can be rewritten as UA, where U is an orthogonal matrix for feature transformation, and $A = [a_1, a_2, ..., a_T]$ is the weight matrix on the new space.

$$\min_{U,A} \sum_{t=1}^{T} \sum_{i=1}^{m} c_{t(y_{ti})} \max(0, 1 - y_{ti} \langle a_t, U^T x_i \rangle) + \gamma \|A\|_{2,1}^2 \tag{6.1}$$

Equation (6.1) shows the objective function of MTFL used in this work. Here, m is the number of training instances, $y_{ti} \in \{-1, 1\}$ is the label of the i-th training instance in task t, x_i is the i-th training instance, $<>$ stands for inner product, $c_{t(y_{ti})}$ is the cost for error in task t for the class y_{ti} belongs to, and γ is the regularization parameter. We use hinge loss ($\max(0, 1 - y_{ti}\langle a_t, U^T x_i\rangle)$) to match the linear SVM. MTFL encourages sparsity of the transformed features and couples the tasks by regularizing on A using the $L_{2,1}$-regularizer. In the general case, U and A are both learned from the data. However, if we force $U = I$, the regularization would be directly imposed on W, in which case the "feature learning" problem reduces to a "feature selection" problem [8, 9]. The convex variants of Equation (6.1) can be solved by iteratively performing a supervised task-specific step and an unsupervised task-independent step. The former step becomes solving linear SVM with a variable transformation process when hinge loss is used [215]. More details about the algorithm can be found in [8, 9]. In this work, we use both the general setting and the special case where $U = I$ as the multi-task equivalent of L_2 and L_1-regularization when training on the original features. Liblinear is used to solve the linear SVMs [50].

We ask an additional question in this work: can IM and JM be improved by operating on a feature space learned through DBN? We investigate whether the advantages of JM are diminished given a non-linear feature preprocessing step. We also train a single task linear SVM model and multi-task MTFL model on the DBN pretrained features. When feature pretraining is applied, we use only the L_2 regularization for linear SVM and the general case of U for MTFL. The reason we are not selecting the input features by enforcing sparsity is that the DBN feature learning has already played the part of dimensionality reduction.

6.3.3 Cross-Validation and Model Selection

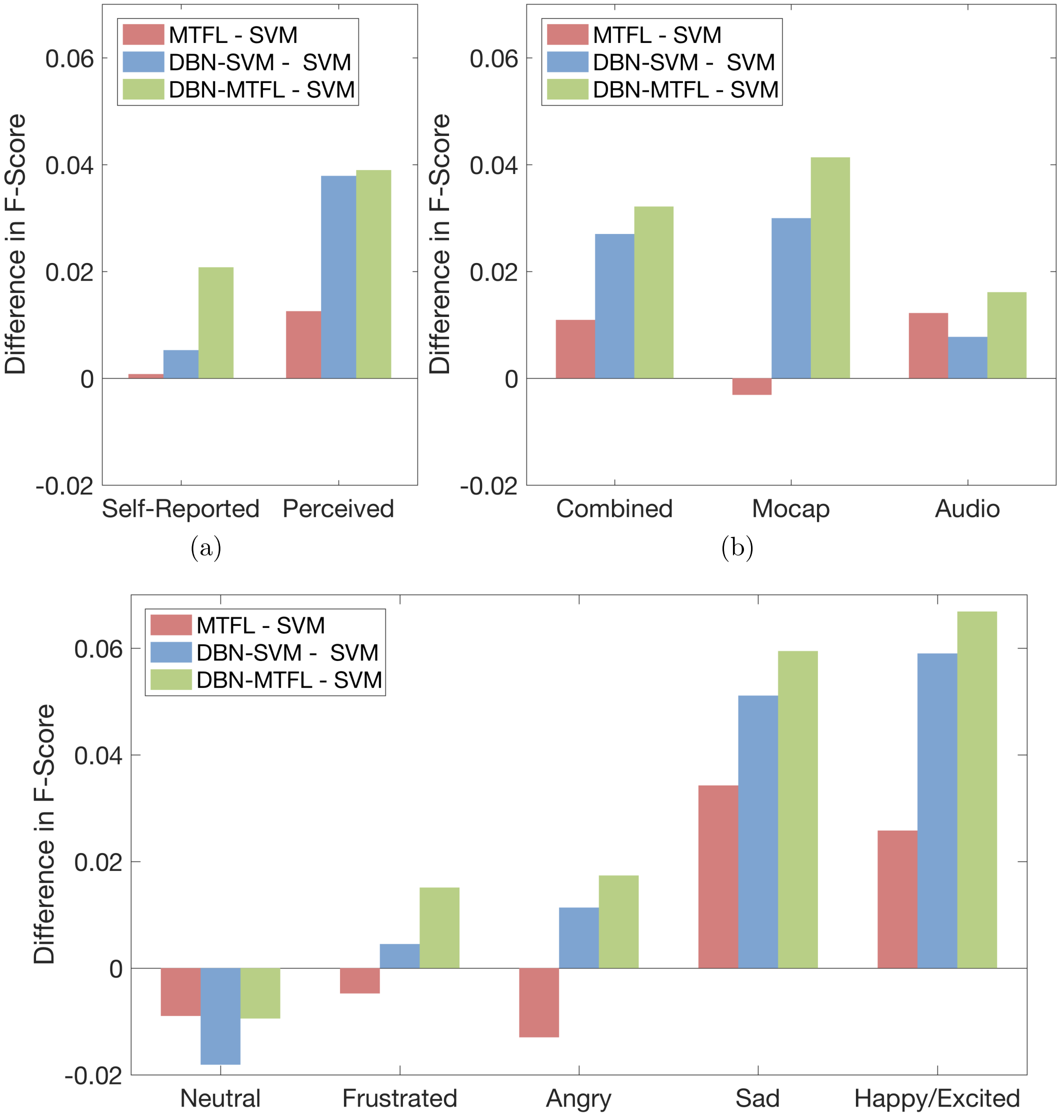
It is important to make the reader aware that we use F-score as a performance measure, rather than the common metric of unweighted recall, to account for the multi-label binary classification problem. The F-score is defined as the harmonic mean of the precision and recall of the positive class (i.e. the presence of the emotion), as in [166]. We report the leave-one-speaker-out cross-validation F-score for each model. At each round, data from one speaker is left out as the test set, while data from other speakers are used for training. In the DBN pretraining, the data of the test speaker are also excluded.

We compare four different settings: modeling self-reported emotion and perceived emotion individually or jointly, on the original features or on the DBN pretrained features. This leads to four models: original-SVM (SVM), original-MTFL (MTFL), DBN-SVM and DBN-MTFL, with SVM being the baseline. There are at least two versions for each model to be selected from: L_1 vs. L_2-regularization for SVM, learned U vs. $U = I$ for MTFL and different number of final hidden units (input to classifiers) for DBN-SVM and DBN-MTFL. We select the version and the hyper-parameters by optimizing the cross-validation F-score on the training set only, where cross-validation is also performed in a leave-one-training-speaker-out way. The range of the regularization parameter γ (in all models) is $\{10^{-4}, 10^{-3}, ..., 10^{5}\}$ and the range of the permutation parameter ϵ (in MTFL and DBN-MTFL) is $\{10^{-8}, 10^{-7}, ..., 10^{-1}\}$. Note that γ is equivalent to the cost of error C for linear SVM, and $C = 1/(2 \times \gamma)$.

6.4 Results and Discussion

6.4.1 Performance of Classification Models

We compare the performance of the SVM, MTFL, DBN-SVM, and DBN-MTFL on the task of recognizing perceived and self-reported emotion labels. On average, all models outperformed the baseline SVM model, in the order SVM < MTFL < DBN-SVM < DBN-MTFL (Table 6.2). MTFL performs better than SVM in the majority of the cases, excepting the prediction of the self-reported emotion label given the uni-



(a) (b)

(c)

Figure 6.3: Average differences in F-score between the baseline SVM and MTFL, DBN-SVM, and DBN-MTFL, respectively. The average is taken for (a) self-reported and perceive emotion, across modality and emotion classes; (b) combined modality, mocap and audio, across types of emotion and emotion classes; (c) each emotion class, across types of emotion and modalities.

Model	Combined		Mocap		Audio		Average
	SR	Perceived	SR	Perceived	SR	Perceived	
SVM	0.574	0.534	0.532	0.475	0.477	0.492	0.514
MTFL	0.579	0.551	0.513	0.487	0.493	0.500	0.521
DBN-SVM	0.578	0.584	0.533	0.533	0.487	0.497	0.535
DBN-MTFL	0.588	0.585	0.555	0.534	0.502	0.499	0.544

Table 6.2: Average F-score of SVM, MTFL, DBN-SVM and DBN-MTFL. The best results in each combination of modality and emotion type is underlined. SR: self-report; Combined: both modalities.

modal mocap data. The improvement in performance from SVM to MTFL suggests that jointly predicting self-reported and perceived emotion is beneficial with respect to performance on both tasks. We find that the DBN feature learning increases the performance for both SVM and the MTFL (DBN-SVM and DBN-MTFL, Table 6.2). The DBN-MTFL model produces the highest accuracy overall (exception: perceived emotion from unimodal acoustic features). This suggests that the individual benefits of the non-linear feature learning and the joint modeling are additive.

The statistical significance are assessed using repeated measure ANOVA on the F-score of each emotion. Model and modality are treated as the two within-subject factors. We find that the influence of model and the interaction between modality and model are significant for perceived emotion ($p = 0.011$ and 0.005, respectively), but not for self-reported emotion. We compare the difference in F-score between each pair of models over the 5 emotions × 3 modalities for perceived emotion using paired t-test. We find that DBN-SVM is significantly better than SVM and MTFL ($p = 0.006$ and 0.020, $t = 3.26$ and 2.64, respectively), and DBN-MTFL is significantly better than SVM and MTFL ($p = 0.004$ and 0.010, $t = 3.44$ and 2.99, respectively).

We compare the performance of the models in Figure 6.3, assessing the influence of self-reported vs. perceived label (Figure 6.3a), modality (Figure 6.3b) and emotion class (Figure 6.3c). In each figure, we treat SVM as the baseline model and assess the change in F-score as a function of the model types (MTFL, DBN-SVM, DBN-

MTFL). We find that the overall performance gain is higher for the perceived labels, and that the advantage of non-linear feature learning is more obvious for perceived labels, compared to the self-reported labels. In the self-reported emotion problem, joint modeling and DBN feature learning, by themselves, show little improvement. However, the combined influence of the two approaches is greater than simple addition.

Of the two unimodal systems (mocap and audio), the mocap modality benefits more from the DBN feature learning. In fact, only using MTFL harms the performance for mocap, yet the combination of feature learning and MTFL leads to a large gain in average performance. On the contrary, the performance gain for joint modeling is higher for audio than mocap, and the addition of DBN feature learning introduces a relatively smaller gain.

The increase in performance is the highest for the emotions of sadness and happiness/excitement. The system performs worse, across all model types, for neutrality compared to the baseline SVM system. In addition, MTFL has lower performance, compared to baseline SVM, for the emotions of frustration and anger. The emotion-specific results mirror the trend in similarity from Section 6.2. The self-reported and perceived emotion labels for the classes of sadness and happiness/excitement are the most similar, compared to those of neutrality, frustration, and anger. This may suggest joint modeling on the original feature space is most effective when the discrepancies between self-evaluation and perception are smaller. However, the performance gain of DBN-MTFL over DBN-SVM is quite consistent in all the five emotions, indicating feature learning increases the robustness of joint modeling.

For each utterance, there are three different situations for prediction: (1) both self-report and perceived label are correct, (2) one label is correct, the other is incorrect, and (3) both labels are incorrect. We present the three situations (averaged over emotion class and modality) for SVM, MTFL, DBN-SVM and DBN-MTFL in Figure

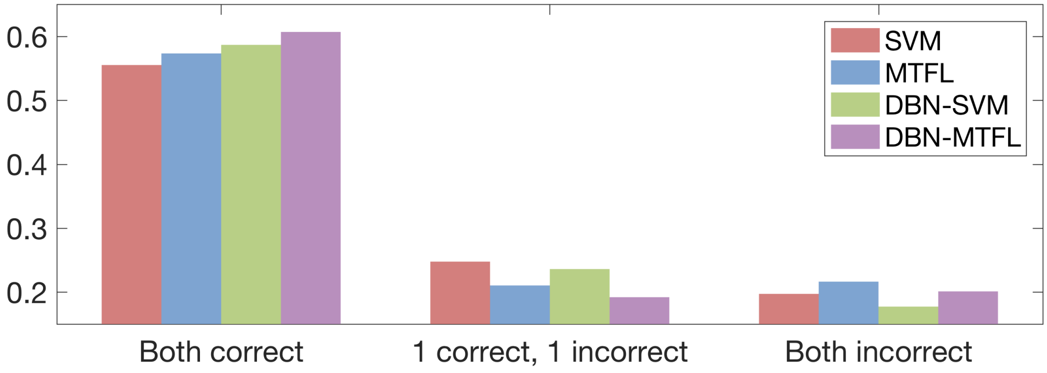


Figure 6.4: Percentage of utterances that have self-report and perceived labels correct/incorrect.

6.4. We find that the proportion of the co-occurrence of success mirrors the overall performance of the models, namely SVM < MTFL < DBN-SVM < DBN-MTFL. Interestingly, DBN feature learning decreases both (2) and (3) (SVM vs. DBN-SVM, MTFL vs. DBN-MTFL), yet joint modeling only contributes to reducing (2), not (3) (SVM vs. MTFL, DBN-SVM vs. DBN-MTFL). The fact that joint modeling increases the co-occurrence of both success and error gives support to our hypothesis that joint modeling leverages the knowledge carried in both labeling methods.

6.4.2 Prototypical vs. Non-Prototypical Emotion

The performance of emotion recognition systems is often assessed as a function of subtlety, described in terms of prototypicality. Prototypicality is defined as complete agreement between evaluators, while non-prototypicality is defined as only majority vote agreement. Previous works have found that it is harder to classify utterances with non-prototypical emotions, compared to utterances with prototypical emotions [78, 107]. In this study, we compare the performance of the proposed techniques as a function of the prototypicality over the perceived emotion label only.

In the *self-evaluation* subset, 52% of the data are prototypical. The Hamming similarity (averaged over five emotions) between two types of labels are 0.92 and 0.83 for prototypical and non-prototypical data, respectively. This suggests a larger discrepancy between self-report and perception for non-prototypical data.

Table 6.3 shows the average performance of each model on prototypical and non-prototypical data for self-reported emotion and perceived emotion. Similar to previous works, we also find that results on prototypical data consistently outperform results on non-prototypical data for both self-reported and perceived emotion. DBN-MTFL achieved the highest average performance, on both prototypical data and non-prototypical data. We compare DBN-MTFL with the baseline SVM on the bottom line. We find that the performance gain for non-prototypical data is higher than for prototypical data, especially for perceived emotion.

We assess the performance change of the models over the prototypical data and non-prototypical data, again using repeated measure ANOVA. The influence of model and the interaction between modality and model are significant on the non-prototypical data for perceived emotion (p = 0.023 and 0.004, respectively), but not on prototypical data or for self-reported emotion. Comparing the models in a pairwise manner for the perceived emotion of non-prototypical data using paired t-test, we find that MTFL is significantly better than SVM (p = 0.016, t = 2.75), DBN-SVM is significantly better than both SVM and MTFL (p = 0.013 and 0.043, t = 2.85 and 2.23, respectively), and DBN-MTFL is significantly better than both SVM and MTFL (p = 0.008 and 0.029, t = 3.11 and 2.43, respectively).

6.4.3 Mixed vs. Clear Emotion

The definition of prototypicality used in Section 6.4.2 does not extend well to self-reported labels because they are derived from a single evaluator (the actor). Instead, we describe subtlety in self-reported labels in terms of the number of labels provided by the actor. We use the term “mixed” when the actor describes his/her data with multiple labels and “clear” when only one label is provided.

In the *self-evaluation* subset, 14% of the utterances are mixed. The average Ham-

Model	Prototypical		Non-Prototypical	
	Self-report	Perceived	Self-report	Perceived
SVM	0.541	0.511	0.490	0.452
MTFL	0.538	0.517	0.494	0.470
DBN-SVM	0.542	0.535	0.498	0.504
DBN-MTFL	0.559	0.536	0.512	0.505
Performance Gain	0.017	0.025	0.022	0.053

Table 6.3: Average F-score of self-reported and perceived emotion on prototypical and non-prototypical utterances. The best results in each column are underlined. The performance gain of feature pretraining + joint modeling is shown in the bottom line.

Model	Mixed		Clear	
	Self-report	Perceived	Self-report	Perceived
SVM	0.618	0.404	0.492	0.505
MTFL	0.603	0.438	0.497	0.514
DBN-SVM	0.564	0.447	0.510	0.540
DBN-MTFL	0.582	0.428	0.525	0.543
Performance Gain	-0.036	0.024	0.033	0.038

Table 6.4: Average F-score of self-reported and perceived emotion on utterances with mixed self-reported emotion (>1 labels) and clear self-reported emotion (=1 labels). The best results in each column are underlined. The performance gain of feature pretraining + joint modeling is shown in the bottom line.

ming similarity between self-report and perceived labels for mixed and clear emotions is 0.75 and 0.90, respectively. This suggests that when an individual notes variability in his/her performance, it is more likely that the self-report and perceived emotion will disagree. However, this does not automatically lead to a designation of non-prototypicality; only 54% of mixed emotions are non-prototypical. This highlights a difference between the perception of variability for self and for other evaluators.

Table 6.4 shows the average performance of each model on mixed and clear data for self-reported and perceived emotion in. The bottom line lists the difference in performance between DBN-MTFL and the baseline SVM. We find the largest performance gain for perceived emotion labels from the clear subset. Joint modeling and DBN pretraining actually have negative influence on the self-reported emotion of mixed data. Interestingly, the performance of perceived emotion on the mixed data is lower than on the clear data. This indicates that when emotion expression is considered subtle by the producer, it is indeed harder for both the classifier (Table 6.4) and the human evaluator (Hamming similarity result) to accurately recognize it, although this subtlety itself may not be fully captured by variation in evaluation.

The repeated measure ANOVA shows that for data with mixed emotion, the influence of model on self-reported emotion is significant (p = 0.009). Pairwise comparison using paired t-test indicates that SVM is significantly better than DBN-SVM and DBN-MTFL (p = 0.002 and 0.015, t = 3.72 and 2.77, respectively), and so is MTFL (p = 0.004 and 0.024, t = 3.44 and 2.54, respectively). This suggests that non-linear feature learning has a negative effect in this case. For data with clear emotion, the influence of model is significant for both self-reported and perceived emotion (p = 0.038 and 0.019, respectively), and the combined influence of model and modality is significant for perceived emotion (p = 0.020). Pairwise model comparison shows that DBN-MTFL is significantly better than SVM (p = 0.002, t = 3.76), MTFL (p = 0.002, t = 3.74) and DBN-SVM (p = 0.021, t = 2.59) for self-reported emotion. For perceived emotion, DBN-SVM is significantly better than both SVM and MTFL (p = 0.011 and 0.022, t = 2.94 and 2.57, respectively), and DBN-MTFL is significantly better than both SVM and MTFL (p = 0.005 and 0.007, t = 3.29 and 3.16, respectively).

6.5 Conclusion

In this work, we explore the impact of jointly predicting self-reported emotion and perceived emotion in addition to non-linear feature learning. We hypothesize that joint modeling using multi-task learning leads to performance increases for both kind of labels, and the performance gain of joint modeling and DBN feature learning is complementary. We experiment on IEMOCAP using a multi-label classification paradigm to test this hypothesis.

The results show that overall, DBN feature learning and joint modeling together produce the highest performance, suggesting the individual benefits of the two approaches are additive. The performance gain is higher for the perceived labels, compared to the self-reported labels, yet we notice that the combined influence of the two approaches is greater than simple addition for self-reported labels. We find that while DBN feature learning does not show preference over different kinds of error, joint modeling increases the co-occurrence of both success and error, and decreases the mismatch of correct and incorrect predictions for self-report and perceived labels. Our findings suggest that joint modeling is able to leverage the potentially complementary information conveyed by both individual labeling strategies, and combining non-linear feature learning with joint modeling leads to more effective emotion recognition systems.

The Brunswik Lens model discusses how communicative cues produced by an encoder (distal indicators), are altered by transmission (proximal percepts), and interpreted by a decoder. The experiment results suggest that when an individual produces an emotional message that he/she believes to be clear, there is benefit to capturing variability in the distal indicators (feature learning) and variability due to transmission (multi-task learning). Interestingly, when an individual does not believe his/her emotion to be clear, this approach is ineffective. This suggests that additional research is needed to understand how to automatically interpret ambiguous emotional expressions.

CHAPTER 7

Modeling Distribution of Emotion Perception

7.1 Introduction

Emotions are not perceived uniformly across individuals. In emotion recognition experiments, inter-rater variability is often mitigated by averaging the ratings of groups of evaluators, under the assumption that this amalgamation can remove perceptual "noise". However, inter-rater variability contains signal, in addition to noise. It provides information about the subtlety or clarity of the an emotional display. In this work, we investigate methods that can effectively capture and predict the variation that is present in a population of evaluators.

We focus on dimensional descriptions of emotion, which characterize emotion in terms of continuous values. Compared to categorical emotion descriptions (e.g., anger, happiness, and sadness), these dimensions are less dependent on context or language [136]. In addition, this characterization naturally captures variation in emotion perception [202], allowing us to retain rich information about the emotional content of a given expression. This information could provide insight on the clarity or subtlety of the emotion expression, and could be used as high-level features for tasks such as the prediction of complex emotions, and the inference of appropriate response in emotion-aware agents. In this work, we use the two most commonly accepted dimensions: valence and activation [135].

Previous work in music emotion recognition has explored methods for generating and predicting distributions of emotion perception [145, 146, 192, 193, 202, 218]. However, they often require a large number of real-valued evaluations, focus on a single modality (i.e., audio), and do not fully exploit short-term temporal information. In speech emotion recognition, there has been work incorporating inter-rater consistency into systems using categorical labels [172, 203]. However, most work using dimensional labels focuses on predicting either the mean evaluation across multiple evaluators [49, 58, 112, 115, 127, 198] or the classes categorized from the mean evaluation [111, 153]. Works that seek to provide emotion variation as a usable and modelable signal for speech are still missing.

In this work, we present a new approach that generates probability distributions on the valence-activation space and captures the variability of emotion perception from speech, using a limited number of ordinal evaluations. We demonstrate how these two-dimensional distributions can be predicted using frame-level audio-visual features. We then ask the following two research questions: (1) can we predict a probability distribution more accurately by modeling local temporal patterns; and (2) can combining audio and video modalities result in better performance compared to when a single modality is used?

We conduct experiments on the MSP-IMPROV dataset [26]. We upsample the evaluations for each utterance by repeatedly performing random subsampling and averaging. The resulting set of evaluations is used to calculate ground-truth probability distributions. We use convolutional neural networks (CNNs) to predict these distributions by leveraging regional temporal patterns for both unimodal and multimodal input. We compare the proposed CNN approach with support vector regression (SVR), the state-of-the-art approach [202, 218], to answer the first question, and compare the performance of different modalities and fusion methods to answer the second.

My experimental results suggest that modeling local temporal patterns is beneficial with respect to both Total Variation and Jensen-Shannon Divergence compared to SVR with utterance-level statistical features. Combining audio and video modalities at the feature-level outperforms approaches that either use a single modality or combine the modalities at the decision-level. The proposed CNN model predominantly improves the prediction of valence. The novelty of this work includes: (1) a label processing method for generating two-dimensional probability distribution from scarce ordinal labels; (2) the first attempt to predict two-dimensional probability distributions of emotion perception for speech using a dynamic approach; (3) an exploration of the influence of modality on predicting the distribution of emotion perception.

7.2 Data

We experiment on MSP-Improv [26]. We choose MSP-Improv because: (1) the available modalities and size of the dataset allow us to train multimodal models and (2) the crowdsourcing evaluation method and the number of evaluations capture variations in emotion perception.

7.2.1 The Dimensional Evaluations

In this work, we focus only on the dimensional labels of valence and activation. We rescale the evaluations to [-1, 1] from [1, 5]. We show the distribution of the number of evaluations per utterance in Figure 7.1a. The majority of the utterances were annotated by less than ten evaluators, yet a portion of the dataset has approximately 30 evaluations. Figure 7.1b illustrates the distribution of the mean evaluations of each utterance on the valence-activation space. The dataset is relatively balanced along the valence dimension but skewed towards positive for the activation dimension.

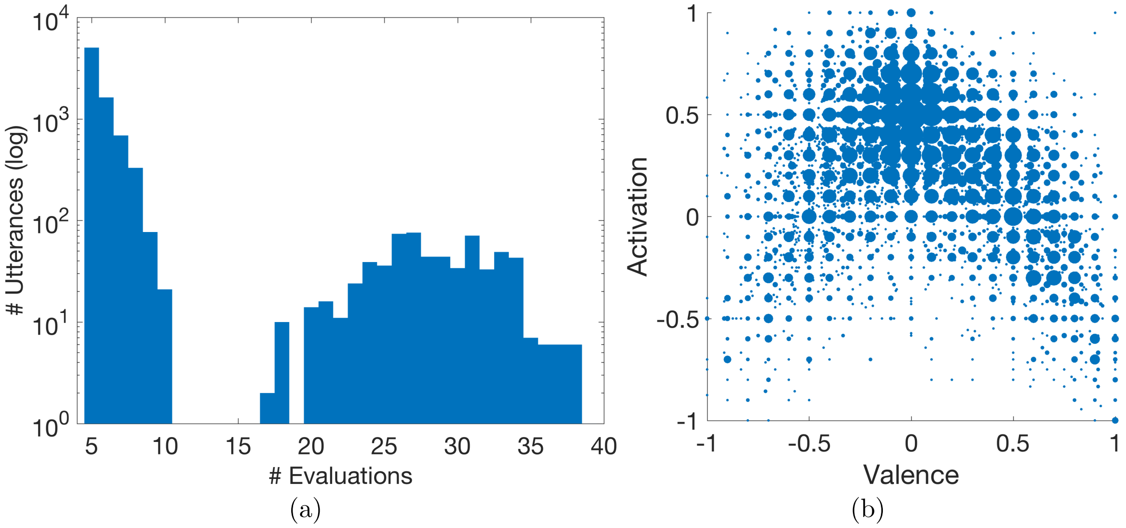


Figure 7.1: Dataset details about MSP-Improv: (A) number of evaluations per utterance (in log scale); (B) average valence-activation per utterance (size of dot proportional to the number of utterances). Note that the direction of activation in (B) is upside-down to be consistent with original MSP-Improv labels: -1 corresponds to high-activation while 1 corresponds to low-activation.

7.2.2 Feature Extraction

7.2.2.1 Acoustic Features

We use 40 log Mel-frequency filterbank features (MFBs) for the audio modality, as in [3]. We first trim the silence at the beginning and end of each utterance and then extract the MFBs from each frame with a window size of 25ms and a step size of 10ms using Kaldi [124].

7.2.2.2 Visual Features

We use the intensity of facial action units (AUs) for the video modality. The AUs, which are the contraction or relaxation of single or multiple facial muscles, stem from the Facial Action Coding System (FACS) proposed by Ekman and Friesen [41]. Using FACS, common facial expressions can be deconstructed into the specific Action Units (AU) that produced the expression. We choose to use AUs to represent the video

AU	Description	AU	Description
1	Inner Brow Raiser	2	Outer Brow Raiser
4	Brow Lowerer	5	Upper Lid Raiser
6	Cheek Raiser	7	Lid Tightener
9	Nose Wrinkler	10	Upper Lip Raiser
12	Lip Corner Puller	14	Dimpler
15	Lip Corner Depressor	17	Chin Raiser
20	Lip Stretcher	23	Lip Tightener
25	Lips Part	26	Jaw Drop
45	Blink		

Table 7.1: Action Unit Features Used in Experiments.

modality because of the close relationship between facial expression and emotion.

We extract the intensity of 17 AUs (Table 7.1) using OpenFace [10], which provides a intra-class correlation coefficient of approximately 0.6 on the test set of the 2015 Facial Expression Recognition and Analysis challenge [181]. We use the "static" prediction model, which relies on a single frame to estimate the intensity of the AUs at each time step. This is because some videos have a limited dynamic range, thus using the dynamic model that attempts to perform pose calibration may be harmful [10].

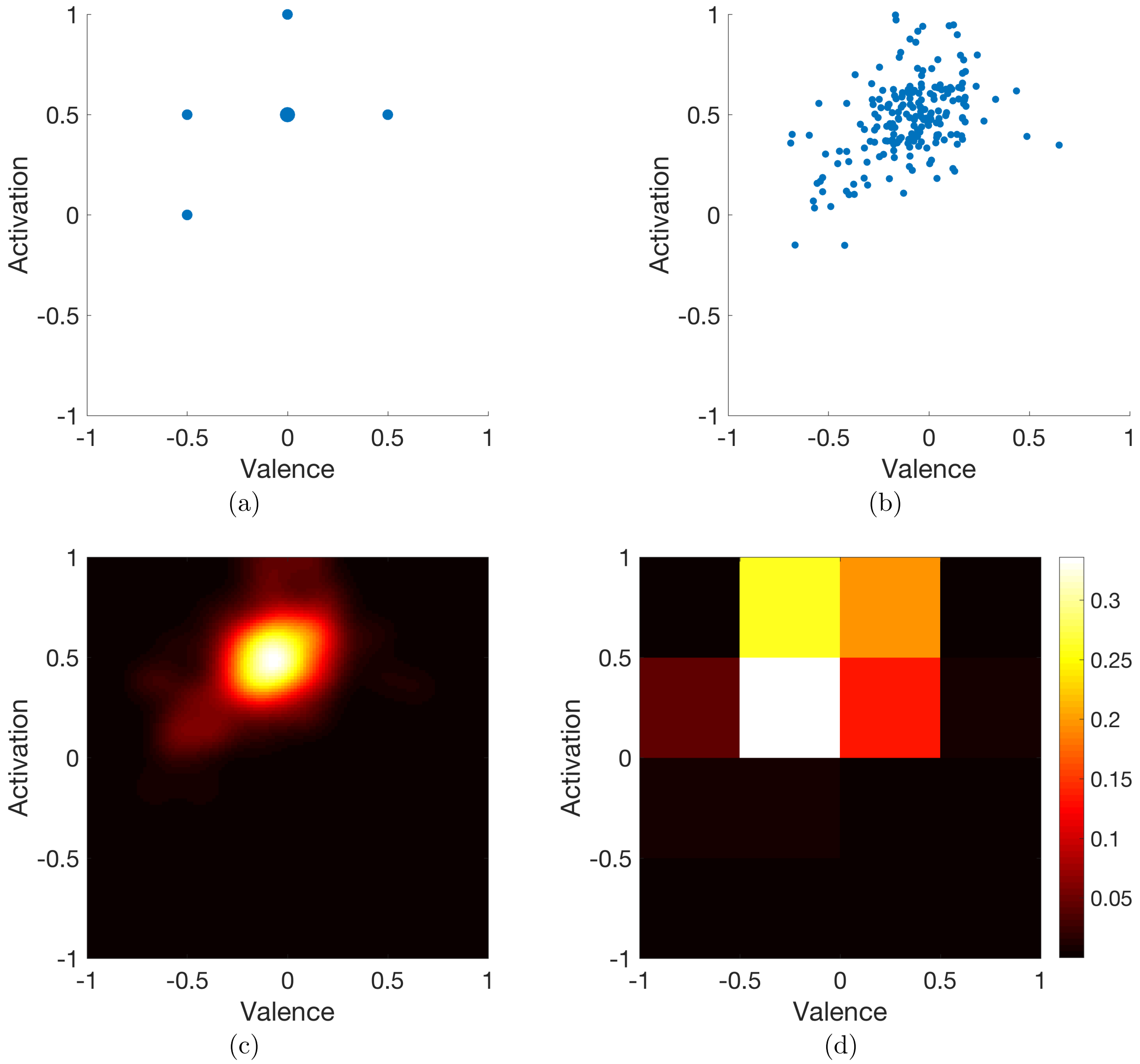
7.3 Methodology

7.3.1 Label Processing

Dimensional annotations are often collected using evaluations over m-point Likert-scales [25, 26, 218]. Previous work has approximated the distribution over evaluations using Kernel Density Estimation (KDE) either from original continuous labels [202] or after applying evaluator-dependent z-normalization [218]. KDE assigns a two-dimensional Gaussian "energy" to each evaluation. The probability density of any point in the valence-activation space can be calculated by summing over the "energy" emitted by all the evaluations. In this work, we adopt the same method because of its ability to generate smooth probability density distributions. However, there are a few challenges that we need to address first:

- The majority of the utterances have less than 10 evaluations, which may not be sufficient to conduct KDE.
- The dimensional labels are ordinal instead of continuous. Therefore, we cannot apply KDE directly.
- It is not guaranteed that each evaluator was given utterances with balanced emotional content. As a result, we cannot use evaluator-dependent z-normalization as in [49, 198, 218].

We argue that the mean of any subset of evaluations of each utterance can be considered a potential ground-truth label of that utterance, inspired by the fact that researchers often use the mean of evaluations as the ground truth, and that the number of evaluations varies within and across datasets. Therefore, for a given utterance, we randomly subsample from one to N evaluations, where N is the total number of evaluations for that utterance, and use the mean as a new annotation. We repeat the process 200 times for each utterance. We add random noise to each generated annotation to avoid the same value being repeated multiple times. The random noise follows a uniform distribution centered at zero, with the width and height corresponding to the standard deviation of the valence and activation for the given utterance, respectively. The generated annotations share similar statistical properties with the original evaluations. On a -1 to 1 scale, the mean absolute difference across all utterances between the mean of original labels and the mean of generated labels are 0.011 and 0.015 for valence and activation, respectively. The correlations between the per-utterance standard deviation of the original labels and generated labels across all utterances are 0.96 for valence and 0.95 for activation. We show an example of the original labels and the corresponding generated labels in Figure 7.2a-7.2b, respectively.



(a)

(b)

(c)

(d)

Figure 7.2: The process of generating the two-dimensional discrete probability distributions: (A) individual evaluations (size of dot proportional to the number of evaluations); (B) annotation cloud generated by averaging subsample of evaluations and adding random noise; (C) probability density distribution calculated by KDE; (D) discretized probability distribution at 4 × 4 resolution.

We then perform KDE using the approach from [15]. Since predicting a continuous function is both challenging and unnecessary, we transform the density function to a discrete probability distribution by creating equally spaced partitions along both valence and activation. Note that we create partitions from a smoothed density distribution instead of from individual labels directly, because the latter approach highly depends on the position of the partitions and can lead to biasing. For example, if the annotations are far apart and we want to predict at a higher resolution, we may end up with grids with high probability surrounding grids with zero probability. We use the mean of the density values within a grid to represent this grid. The values of all the grids are then normalized to sum to one. We show the density function from KDE and the discrete probability distribution in Figure 7.2c-7.2d, respectively.

7.3.2 Models

We ask two main research questions: (1) can we better predict the distribution of emotion perception by focusing on salient local regions and jointly optimizing the predictions across the grids; (2) can we understand the influence of modality?

We answer the first question by comparing a static regression approach from [202] and an approach that takes regional temporal patterns into account. We answer the second question by building four models for both approaches: two unimodal models (audio modality and video modality), a model combining the two modalities at decision-level by averaging (denoted as combined-late), and a model combining the two modalities at feature-level (denoted as combined-early).

Past research has found that ν-Support Vector Regression (SVR) with a Gaussian kernel is effective for predicting the discrete probability distribution of emotion [202, 218]. We use this approach with the same implementation (LibSVM [29]) as the baseline. SVR takes in static utterance-level features and predicts the probability of each grid separately. Because the regressors are optimized individually and the predictions are not bounded, we truncate negative values to zero and normalize the estimations over all grids to sum to one, as in [202, 218]. We concatenate the acoustic and visual features for the combined-early model.

We choose convolutional neural networks (CNNs) as my approach. CNNs have been used in affective computing to learn emotionally salient features from audio [66, 94, 177] and video [51, 205]. Recent works have explored the efficacy of CNNs for modeling temporal patterns. Mao et al. extracted emotion-salient features for speech emotion recognition using CNNs [94]. They first used a sparse auto-encoder to learn filters at different scales from unlabeled speech signals and convolved the spectrogram segments with learned filters to form a series of feature maps. They performed mean-pooling and stacked the feature maps into a feature vector. After that, they used the feature vector as the input to a fully-connected layer to learn emotion-salient features before feeding the learned features into a Support Vector Machine (SVM) classifier. Aldeneh and Mower Provost used a CNN with a convolutional layer, a global max-pooling layer, and several dense layers to identify emotionally salient local patterns and classify emotion from temporal low-level acoustic features [3]. They obtained comparable results to the state of the art utterance-level statistic features and SVMs. Khorram et al. used dilated CNN and downsampling-upsampling CNNs for predicting time-continuous valence and activation labels [77]. Their methods outperformed BLSTMs and were 46 times faster. These works support that CNNs can be used to model temporal patterns.

We reframe the problem as classification using soft labels, rather than as regression. This leverages the fact that the output layer is usually a softmax, the output of which can be interpreted as the probability of each class.

We design the unimodal (i.e., audio or video) CNNs similar to [3], with an 1D-convolutional layer, a global max-pooling layer, several dense layers and a softmax layer (Figure 7.3). The 1D-convolutional layer takes in variable-length input and

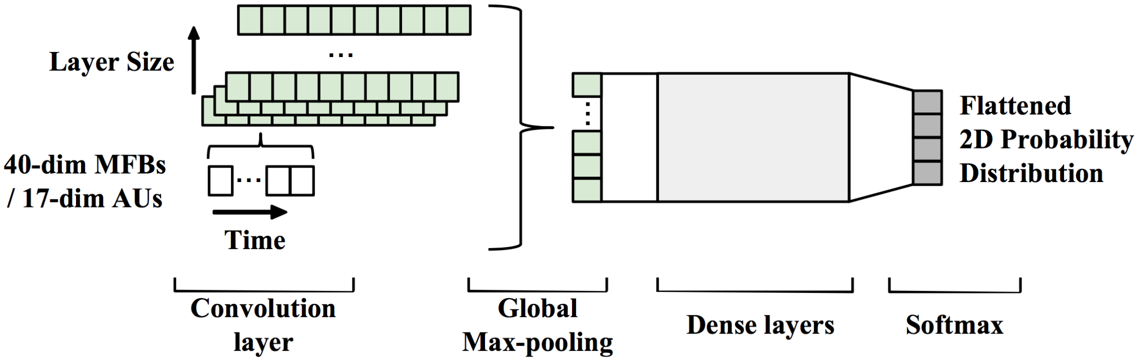


Figure 7.3: The structure of the unimodal CNN.

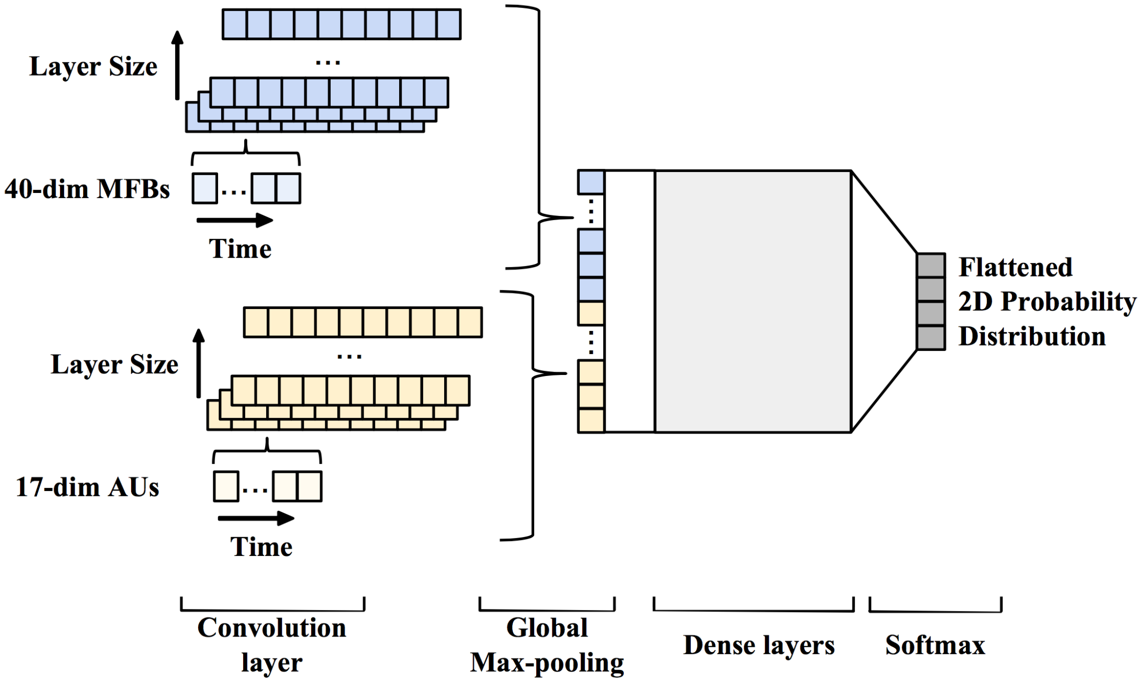


Figure 7.4: The structure of the multimodal CNN with the combined-early approach.

learns a sequence of feature representations by sliding N_F filters of length L_F through time. Each filter takes L_F consecutive frames and outputs an activation. By learning the filters, we are finding emotion-salient local temporal patterns. The global max-pooling layer identifies the highest activation of each filter over time and produces a feature vector of length N_F. This step allows us to focus on the most informative portion of an utterance and minimize the influence of padding and frames with invalid features. The interactions between the N_F features are further learned by applying several dense layers. Finally, we use a softmax layer to output the probability of emotion perception in each grid.

For the multimodal combined-early model, we build a separate 1D-convolutional layer and corresponding global max-pooling layer for audio and video, respectively. We concatenate the output of the two global max-pooling layers before feeding the features into dense layers. In this way, we allow for the difference in frame-rate of audio and video input, while still being able to explore the complex non-linear relationships between features across modalities. The multimodal CNN is shown in Figure 7.4.

We use the Rectified Linear Unit (ReLU) [209] as the activation function for the convolutional and dense layers, and cross-entropy as the loss function. We apply L_2-regularization (0.0001) on the learned weights of the convolutional layers. The filter length of the convolutional layer (L_F), layer size (N_F), and the number of dense layers are treated as hyper-parameters.

7.4 Experimental Settings

7.4.1 Feature Preparation

While the CNN models use frame-level features directly, the SVR models use a static approach with utterance-level features. We calculate 11 statistics over the frame-level MFBs and AUs and their first-order delta coefficient to generate the 880 and 374 utterance-level features for audio and video, respectively. The statistics include mean, standard deviation, max, position of the max frame, min, position of the min frame, range, interquartile range, mean absolute deviation, skewness, and kurtosis. For the video modality, the statistics calculation is applied only to frames with successfully extracted features (>98%). We also extract the state of the art Interspeech 2013 acoustic feature set [155] (6,373 utterance-level features, denoted as "IS13") to use in the SVR models for comparison.

We perform speaker-dependent z-normalization on all features before they are input into models. We normalize the features at the frame-level for CNN models and at the utterance-level for SVR models. Similarly, we exclude frames with unsuccessful AU extraction when z-normalizing the frame-level AUs. We replace these frames with zeros after normalization. We do not interpolate between frames that are successfully extracted because the unsuccessful extractions are usually a consecutive sequence of frames, and interpolation may introduce noise.

7.4.2 Performance Evaluation and Validation

We conduct experiments at two grid resolutions: 2 × 2 and 4 × 4. We use two metrics to evaluate the performance of the models: Total Variation (TV) and Jensen-Shannon divergence (JS). Both metrics can measure the difference between two discrete probability distributions. The metrics are calculated per utterance, and the lower TV and JS, the better the performance.

The value of Total Variation ranges in $[0, 1]$. Given two probability distribution X and Y over N states, The total variation between them is defined as

$$TV(X,Y) = \frac{1}{2}\sum_{i=1}^{N}|X_i - Y_i|, \tag{7.1}$$

Jensen-Shannon divergence is extended from the Kullback-Leibler divergence (denoted as KL). It is defined as

$$JS(X,Y) = \frac{1}{2}KL(X,M) + \frac{1}{2}KL(Y,M), \tag{7.2}$$

$$\text{where } M = \frac{1}{2}(X+Y) \text{ and } KL(X,Y) = \sum_{i=1}^{N} X_i \log \frac{X_i}{Y_i}.$$

We use JS instead of KL because: (1) JS is symmetric while KL is not, and (2) the value of JS ranges in [0, 1] when using log_2. Note that we replace zeros with 1e-8 when calculating the KL step in JS.

We train the models using a leave-one-speaker-out approach. At each round, a speaker is left out as the test set, while the other speaker in the same session is used as the validation set. The remaining ten speakers are used for training. We calculate the mean TV and JS for each test speaker and report the value averaged across all rounds as the performance of the models.

We use TV as the main validation metric because of its robustness to zeros. We select the hyper-parameters according to the validation TV. For the SVR models, the ranges are C (cost of error) $\in \{10^{-3}, 10^{-2}, ..., 10^{1}\}$, γ (kernel width) $\in \{10^{-5}, 10^{-4}, ..., 10^{-1}\}$ and ν (lower bound on the proportion of support vectors) $\in$ {0.5, 0.6, 0.7, 0.8}. For the CNN models, the ranges of the filter length, layer size and the number of dense layers are shown in Table 7.2. Note that the number of filters in the convolutional layer (after concatenation in the case of combined-early) and the

Modality	Input	Filter length	Layer size (before concatenation for combined-early)	# Dense Layers
Audio	40	{8,16}	{128,256}	{1,2,3}
Video	17	{2,4}	{64,128,256}	{1,2,3}
Combined	40; 17	{8,16}; {2,4}	{64,128,256}	{1,2,3}

Table 7.2: Range of Hyper-parameters in CNNs.

size of the dense layers are kept the same. We use a training strategy of learning rate decay after N epochs. We randomly initialize the weights and start training with a learning rate of 0.001. We maintain the learning rate for 10 epochs, and select the one with the best validation TV to continue training. After that, we restore the previous weights and halve the learning rate when there is no improvement in validation TV after an epoch. We stop training when we reach the minimum learning rate or have five consecutive epochs with no improvement in validation performance, whichever comes first.

7.5 Results and Discussion

7.5.1 Performance Comparison

Table 7.3 shows the performance of the SVR and CNN models at two resolutions for different modalities. In addition to the baseline, we provide chance performance of: (1) a uniform distribution across grids (denoted as Uniform), and (2) the mean distribution of the training set (denoted as MTrain). To answer the two research questions, we compare the performance between: (1) the SVR and CNN models when controlling for modality, and (2) the different modalities and combinations when controlling for the model.

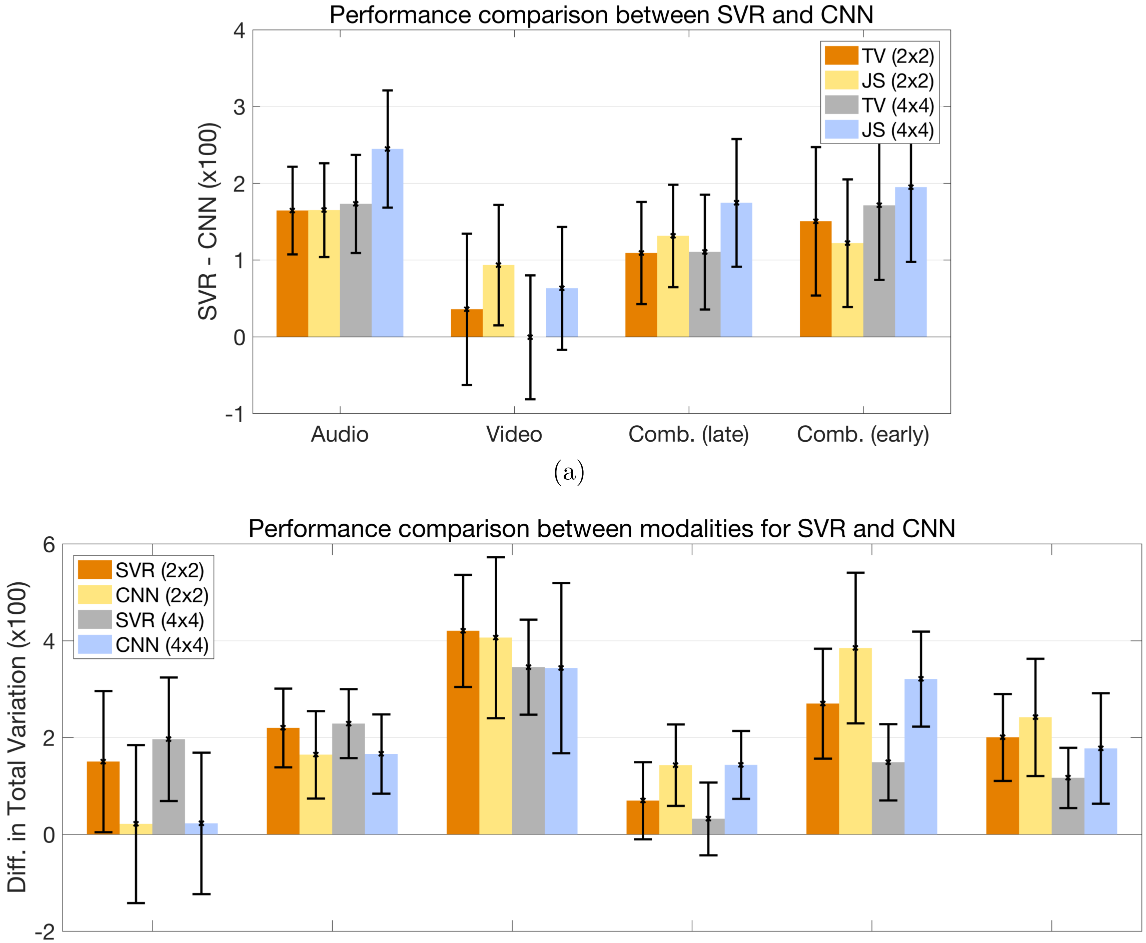
We show the performance difference between the SVR models with utterance-level MFB and/or AU features and CNN models with the corresponding frame-level features in Figure 7.5a, along with the 95% confidence interval of paired t-test. We see significant performance improvement when using CNN for both resolution and evaluation metrics (in the order of TV (2×2), JS (2×2), TV (4×4), and JS (4×4)), for audio ($p = 5.6\text{e-}5$, 9.7e-5, 9.4e-5, and 2.1e-5, respectively), combined-late ($p = 0.0041$, 0.0012, 0.0077, and 7.4e-4, respectively), and combined-early ($p = 0.0057$, 0.0081, 0.0026, and 0.0010, respectively). We also compare the CNN models with SVR with

Modality	Model	Features	2×2		4×4	
			TV	JS	TV	JS
Chance	Uniform	-	.531	.303	.680	.481
	Mtrain	-	.475	.260	.596	.391
Audio	SVR	IS13	.390	.204	.528	.329
	SVR	MFB	.399	.213	.536	.340
	CNN	MFB	.383	.196	.519	.316
Video	SVR	AU	.384	.200	.516	.318
	CNN	AU	.381	.191	.516	.312
Combined-late	SVR	IS13+AU	.373	.185	.510	.307
	SVR	MFB+AU	.377	.189	.513	.311
	CNN	MFB+AU	.366	.176	.502	.293
Combined-early	SVR	IS13+AU	.362	.181	.507	.304
	SVR	MFB+AU	.357	.178	.501	.300
	CNN	MFB+AU	**.342**	**.166**	**.484**	**.281**

Table 7.3: The performance of SVR and CNN models at two resolutions for each modality and combination. The best performance for each metric-resolution combination is bolded. The chance performances are also provided.

the state of the art IS13 feature set for models using audio input. The performance improvement of CNN is significant for audio (p = 0.0011, 0.013, 0.046, and 0.0077 for TV (2×2), JS (2×2), TV (4×4), and JS (4×4), respectively), combined-late (p = 0.0081, 0.049, and 0.0041 for JS (2×2), TV (4×4), and JS (4×4), respectively) and combined-early (p = 2.7e-4, 0.0014, 6.3e-4, 3.9e-4 for TV (2×2), JS (2×2), TV (4×4), and JS (4×4), respectively). This indicates that focusing on salient local regions and jointly optimizing for all the grids together is beneficial for audio input and multimodal input with either decision-level or feature-level fusion. Audio input benefits the most from the CNN architecture. Of the two multimodal systems, the performance gain from feature-level fusion is higher than decision-level fusion. However, there is no significant difference between the performance of CNN and SVR for video input, except when using JS as metric at 2×2 resolution (p = 0.024). This may be because that while the close relationship between AUs and emotion ensured the relevance of the features, the small dimensionality of the input and the high-level nature of the AUs limit the learning ability of the CNN. In addition, the errors propagated from AU estimation may have larger influence on the dynamic and more complex CNN models.

We perform the repeated-measure ANOVA (denoted as RANOVA) to compare different modalities for SVR and CNN. As the compound symmetry assumption may not be satisfied, we evaluate significance of the influence of modality based on the p-value with Lower bound adjustment (p_{LB}). If p_{LB} is smaller than 0.05, we perform the Tukey's honest significant difference test (denoted as Tukey test) for pairwise comparison using the model statistics of RANOVA. For simplicity, we only compare TV because it is used as the validation metric. We find that the influence of modality is significant for both SVR ($F(3, 33)$ = 48.4 and 47.9, p_{LB} = 2.4e-5 and 2.5e-5 for 2×2 and 4×4, respectively) and CNN ($F(3, 33)$ = 35.1 and 31.5, p_{LB} = 1.0e-4 and 1.6e-4 for 2×2 and 4×4, respectively). We show the pairwise comparison with the 95% confidence interval of the Tukey test in Figure 7.5b. For both SVR and CNN, combined-early significantly outperforms both unimodal models and combined-late. This suggests that both models have the ability to learn the interaction between audio and video when we perform fusion at the feature level. While decision-level fusion also brings improvement, this improvement is not always significant (e.g., combined-late vs. video for SVR). For unimodal inputs, video performs significantly better than audio for SVR while the performance of audio and video modalities are comparable when using CNN. This again supports that the CNN architecture may not be ideal for the high-level AU features and that there may be a larger information loss in the video modality.



(a)

(b)

Figure 7.5: Performance difference at 2 × 2 and 4 × 4 resolutions, between: (a) SVR and CNN (in both total variation and JS-divergence), along with 95% confidence intervals of paired t-test; (b) different modalities for SVR and CNN (direction of subtraction shown as x-labels), along with 95% confidence intervals of Tukey's honest test.

Modality	Model	2×2		4×4	
		V	A	V	A
Chance	Uniform	.362	.356	.362	.356
	MTrain	.363	.271	.363	.271
Audio	SVR	.322	.182	.324	.189
	CNN	.300	.182	.298	.183
Video	SVR	.267	.216	.272	.216
	CNN	.258	.221	.257	.224
Combined-late	SVR	.285	.190	.291	.196
	CNN	.269	.191	.267	.193
Combined-early	SVR	.264	**.176**	.272	.184
	CNN	**.248**	**.176**	**.254**	**.173**

Table 7.4: TV of SVR and CNN models (using matching utterance-level and frame-level features) along valence and activation, calculated from the prediction at 2 × 2 and 4 × 4 resolutions. The best performance for each dimension-resolution combination is bolded. V: Valence; A: Activation.

7.5.2 Analysis along Valence and Activation

We assess the ability of the models to predict valence and activation. We marginalize by summing over the predictions along valence or activation to generate distributions of resolution 2×1 for negative and positive activation or 1×2 for negative and positive valence, and compare to the ground truth distribution processed in the same way. We show the resulting TV in Table 7.4, along with the chance of Uniform and MTrain for reference. We find that the audio modality is better at predicting activation, while the video modality is better at predicting valence. This is in line with previous findings [49]. We see multimodal improvement in the feature-level fusion setting (combined-early). In addition, when we compress the output to a two-state probability distribution along valence or activation, the predictions for 2×2 and 4×4 resolution have similar performance.

We illustrate the drop in TV of CNN compared to SVR along valence and activation dimensions in Figure 7.6, together with the 95% confidence interval of paired

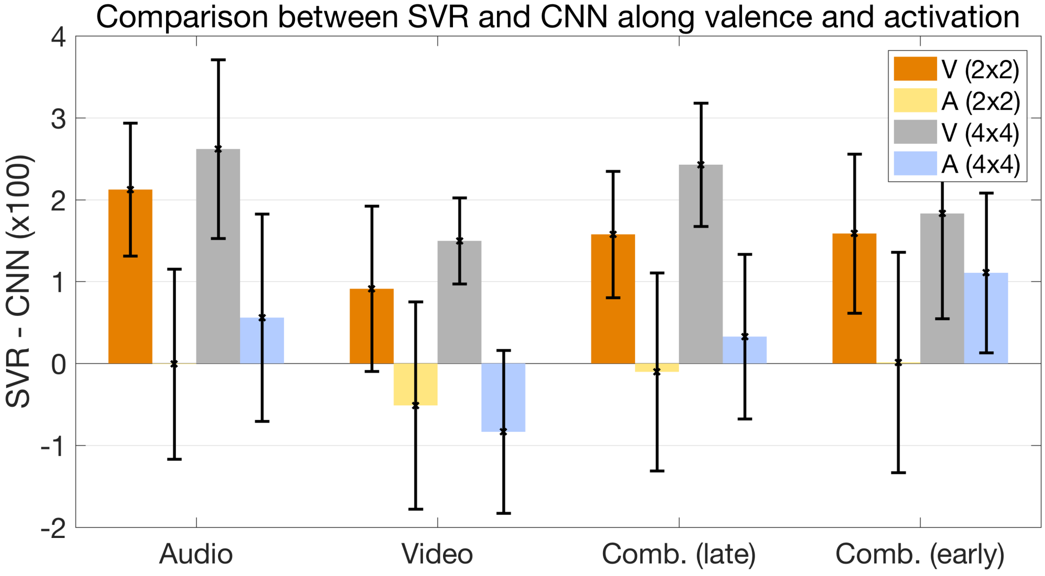


Figure 7.6: Performance difference (in TV) between SVR and CNN along valence and activation dimensions, calculated from predictions at 2 × 2 and 4 × 4 resolutions, with 95% confidence intervals of paired t-test. The best performance for each type-resolution combination is bolded.

t-test. While most works using dynamic approaches witness higher performance improvement in activation [49, 127, 198], we find that the performance gain of the CNN predominantly comes from valence, regardless of modality. More specifically, the difference between CNN and SVR is significant for valence for audio (p = 1.3e-4 and 2.6e-4 for 2×2 and 4×4, respectively), video (p = 6.1e-5 for 4×4), combined-late (p = 8.9e-4 and 2.0e-5 for 2×2 and 4×4, respectively), and combined-early (p = 0.0042 and 0.0093 for 2×2 and 4×4, respectively). The only significant result for activation comes from combined-early with 4×4 resolution (p = 0.030). These results indicate that identifying salient local patterns using CNN brings more benefit in predicting valence, compared to activation. This might be related to the observation in Section 7.2 that the data is more balanced along valence compared to activation.

Modality	Model	2×2 T	 I	 N	4×4 T	 I	 N
Chance	Uniform	.530	.530	.529	.695	.676	.680
	MTrain	.489	.478	.456	.613	.595	.585
Audio	SVR	.410	.415	.366	.541	.549	.508
	CNN	.381 ∗	.411	.337 ∗	.533	.539 ∗	.481 ∗
Video	SVR	.372	.402	.358	.502	.535	.488
	CNN	.367	.405	.346	.519	.533	.486
Combined-late	SVR	.377	.395	.347	.509	.530	.483
	CNN	.357 ∗	.394	.326 ∗	.511	.522 ∗	.467
Combined-early	SVR	.342	.377	.328	.495	.519	.469
	CNN	**.315** ∗	**.370**	**.313**	**.479**	**.507**	**.451** ∗

Table 7.5: TV of SVR and CNN models (using matching utterance-level and frame-level features) for each type of recordings. T: Target sentences; I: Improvised turns; N: Natural interactions.

7.5.3 Analysis of Different Types of Recordings

We investigate the performance difference between SVR and CNN for different types of recordings. As mentioned in Section 7.2, MSP-IMPROV consists of four types of recordings: read target sentences, target sentences from improvised scenes, other speaker turns from improvised scenes, and natural interaction during the breaks. We combine the first two types in this analysis, because of the lack of the read target sentences for five of the speakers.

We present the TV of prediction for different types of recordings in Table 7.5, together with the chance performances. We find that in general, all the models are the best at predicting the emotion perceived from natural interactions, followed by target sentences. The emotion perceived from improvised scenes is the hardest to predict. This matches the classification accuracy of different types of recording using categorical labels reported in [26]. This might be because that the improvised scenes contain a wider range of emotion than the natural interactions since they were designed to enable the speakers to express a variety of emotions. The standard deviation of the mean evaluation of each utterance is higher in improvised scenes, suggesting a larger difference in the emotional content across utterances.

We mark the significant improvement (paired t-test, p<0.05) of CNN compared to SVR using "∗" in Table 7.5. We find that the performance gain of CNN is not consistent across different resolutions. For example, CNN significantly outperforms SVR for the target sentences for audio ($p = 0.0028$), combined-late ($p = 0.011$), and combined-early ($p = 0.029$) in the 2×2 case, but not in the 4×4 case. On the other hand, CNN significantly outperforms SVR for the improvised scenes for audio ($p = 0.010$) and combined-late ($p = 0.011$) in the 4×4 case, but not in the 2×2 case. The most consistent improvement is observed in natural interaction with audio and multimodal inputs. The performance difference between CNN and SVR is significant for audio at both resolutions ($p = 0.0093$ for 2×2, $p = 0.018$ for 4×4), combined-late at 2×2 ($p = 0.011$) and combined-early at 4×4 ($p = 0.048$), and is approaching significance for combined-early at 2×2 ($p = 0.052$).

7.6 Conclusion

In this work, we proposed a label processing method to generate two-dimensional discrete probability distributions on the valence-activation space from a limited number of ordinal labels. We showed that this method can preserve the mean evaluation of the original labels and that the correlation between the standard deviations of the original labels and up-sampled labels is high. Further, we explored the impact of modeling approaches (i.e., static SVR with individual optimization for each grid vs. dynamic CNN with joint optimization for all grids) and modalities on predicting the probability distribution of emotion perception. We hypothesized that using CNN models with a focus on salient local temporal patterns leads to a performance gain. In addition, combining audio and video modalities results in better performance compared to using each individual modality.

My results show that the CNN models significantly outperform the SVR models when using the audio modality and combined audio and video modalities, supporting the effectiveness of modeling locally salient patterns and jointly predicting the distribution over all grids. CNN and SVR are comparable when the video modality is used. This indicates that the potential of CNN may not be fully explored when using a limited number of high-level AUs as inputs. In addition, the errors from AU estimation may have larger influence on the dynamic and more complex CNN models. We find that using both audio and video modalities is better than using either individually and that feature-level fusion is more beneficial than decision-level fusion. Analyses along different dimensions show that the audio modality is better at predicting activation while video modality is more advantageous at predicting valence, and that we can obtain improvement over the joined valence-activation space with feature-level fusion. This is in line with previous findings. In addition, we find that the performance gain brought by CNN mainly comes from the valence dimension. We see a consistent performance improvement over natural interactions when using CNN models, compared to SVR models.

Part III

Data and Label Variability

CHAPTER 8

Preserving Emotional Similarity using Deep Metric Learning with Soft Labels

8.1 Introduction

Speech emotion recognition is complicated by the complex co-modulations that are present in the speech signal (e.g., lexical information and speaker identity). Networks may be unintentionally over-trained to capture signals that are specific to certain speakers or lexical artifacts in the data, resulting in poor generalizability and poor robustness in cross-corpus tasks.

A possible approach for addressing this problem is Deep Metric Learning (DML). DML aims to use deep neural networks to project input data to a learned space, in which the similarity between examples can be directly measured [92]. DML has been successfully applied to many visual understanding tasks, such as face verification, image classification, and person re-identification [63, 147, 173, 207]. For speech emotion recognition, DML can be used to generate an embedding space in which distances between examples correspond to the label relationships. This provides a mechanism to reduce the influence of factors other than emotion.

However, DML approaches generally presuppose that data can be divided into classes using hard labels. This is not ideal for speech emotion recognition, because the variability in emotion expression and the variability in emotion perception lead to datasets with uncertain labels. Previous work in SER has demonstrated the efficacy of using soft labels given uncertainty [52, 105, 172]. Works in other field have also demonstrated that soft labels may be preferable to hard labels in some cases: they provide more information for each training example [62] and are more robust against label noise [175].

The additional information contained in soft labels is not fully exploited in traditional DML approaches. Motivated by this, we propose a family of loss functions, the f-Similarity Preservation Loss (f-SPL), based on the dual form of f-divergence. f-SPL is designed for DML with soft labels, here defined as real-valued labels that are distributed along one or multiple dimensions. f-SPL aims to preserve the label similarities in the learned feature space and can be applied to tasks that require either continuous or discrete (e.g., a class index) test output. Further, we introduce a pair sampling method for the efficient implementation of f-SPL in neural networks.

We evaluate our methods on cross-corpus speech emotion recognition (SER). We form the problem as binary classification of soft-labeled valence (positive vs. negative) and activation (calm vs. excited) [135]. We combine the proposed loss with classification loss in the training of DNN classifiers. Our baseline is the same classifier trained with classification loss only. The results show that our multi-task framework with the added f-SPL statistically significantly increases system performance in the majority of the experiments and is more robust to over-training than the baseline system.

8.2 Related Works

8.2.1 Deep Metric Learning

Deep metric learning approaches predominantly focus on hard labels [92]. These approaches often rely on loss functions that aim to pull data from the same class closer while pushing data from different classes farther apart. Some works use contrastive loss for pairs of examples through Siamese networks [17, 31], identifying "positive pairs" of examples from the same class and "negative pairs" of examples from different classes. This loss then aims to learn a space where the distance between a positive pair is less than a margin τ_+ while the distance between a negative pair is larger than a margin τ_- , where $0 \leq \tau_+ < \tau_-$.

Some works have proposed loss calculations over triplets, defined as sets of three examples: an anchor, a positive example from the anchor's class, and a negative example from a different class [197]. DNNs with triplet loss [63, 147] aim to learn an embedding space where the distance between the anchor and the positive example is at least smaller than the distance between that anchor and the negative example by a margin τ.

Some works have extended the triplet loss, by considering all positive and negative pairs within a batch [167], using multiple negative examples in each set [165], or using the cluster center rather than a single example as the anchor [89, 90]. Yang et al. proposed a loss function designed for image sentiment analysis [204], based on the relationships between neighboring sentiment classes on the Mikels' emotion wheel [102]. They added "related" examples, defined as examples from a different class than the anchor but on the same half of the emotion wheel, to triplets. Denoting the distance between anchor and the positive example as anchor-positive, their approach aimed to find a space where anchor-positive is at least smaller than anchor-related by τ_1, and anchor-related is at least smaller than anchor-negative by τ_2. The distance is scaled by class similarity, implemented using a factor proportional to the class distance on the emotion wheel.

Two recent works have used DML for regression. Wang, Wan, and Yuan combined metric learning for kernel regression with DNN for crowdedness regression [194] . Doumanoglou et al. proposed a loss function via Siamese network for pose estimation [37]. They compared the distance between labels (d_l) and embeddings (d_f) given pairs of data. Their approach aim to minimize $d_f - d_l$. However, the theoretical soundness was not verified since $d_f - d_l$ can be negative.

In this work, we propose a family of loss functions for DML with real-valued labels and provide theoretical justifications. We experiment on classification tasks with soft labels. However, the application of the loss functions could also be extended to other tasks, including regression.

8.2.2 f-Divergence

f-divergence is a family of non-symmetric measures of difference between two distributions, based on the family of convex functions f [6]. These measures are widely used in the learning literature. Common members of the f-divergence family include Kullback-–Leibler (KL) divergence and total variation distance. Nguyen, Wainwright, and Jordan proposed a duality technique of f-divergence [110], which plays a key role in mutual information estimation [109], the design of a type of generative adversarial networks, f-GANs [113], and the design of information elicitation mechanisms and co-training algorithms [82]. We use the dual formulation of f-divergence to derive our f-Similarity Preservation Loss.

8.3 f-Similarity Preservation Loss (f-SPL)

Our goal is to learn an embedding space on which the similarity between examples equals to the label similarity. In Section 8.3.1, we define a family of loss functions,

f-SPL, based on the dual form of f-divergence. Then in Section 8.3.2, we mathematically prove that we can achieve our goal by minimizing f-SPL. Finally in Section 8.3.3, we explain how f-SPL can be implemented in a multi-task framework.

f-divergence	$f(t)$ [113]	f-SPG(S, C)	f-SPL(S, C)
KL divergence	$t \log t$	$C * (1 + \log S) - S$	$S - C \log(S) - C + C \log(C)$
Reverse KL	$-\log t$	$C * (-\frac{1}{S}) - (\log S - 1)$	$\log(S) + \frac{C}{S} - \log(C) - 1$
Pearson χ^2	$(t-1)^2$	$C * 2(S-1) - (S^2 - 1)$	$(C-S)^2$
Squared Hellinger	$(\sqrt{t} - 1)^2$	$C * (1 - \sqrt{\frac{1}{S}}) - (\sqrt{S} - 1)$	$\sqrt{S} + C\sqrt{\frac{1}{S}} - 2\sqrt{C}$
Jensen-Shannon (JS) Divergence	$-(t+1) \log \frac{t+1}{2}$ $+t \log t$	$C * \log \frac{2S}{1+S} + log(\frac{2}{1+S})$	$(C+1)\log(1+S) - C\log(2S)$ $-(C+1)\log(1+C) + C\log(2C)$

Table 8.1: Reference for common f-divergences, their corresponding convex functions f, f-SPG(S, C), and f-SPL(S, C).

8.3.1 Definition of f-SPL

We denote data and soft labels as $x_1, x_2, ... \in A_X$ and $y_1, y_2, ... \in A_Y$, respectively. The function $C : A_Y \times A_Y \mapsto [0, 2]$ measures label similarity. A feature learning function (i.e., a neural network) $g \in G$, maps inputs from A_X to a new space A_G and $S : A_G \times A_G \mapsto [0, 2]$ measures the similarity on A_G. We seek to find a function, $F(S(g), C)$, that optimizes over g. The optimal solution of F, g^*, satisfies $S(g^*(x_i), g^*(x_j)) = C(y_i, y_j)$ for every $i \neq j$, i.e., the similarity between the examples on the learned space is the same as the similarity between their labels.

We use the dual form of f-divergence to construct $F(S(g), C)$. We name the resulting functions f-Similarity Preservation Gain (f-SPG). We then modify f-SPG to a family of loss functions, f-SPL, such that: (1) f-SPL is always non-negative and (2) maximizing f-SPG is equivalent to minimizing f-SPL.

f-SPG Given a convex function f, a feature learning function $g \in G$, and a pair of examples $p = [(x, y), (x', y')]$, we define f-SPG based on the dual formulation of f-divergence (Section 8.3.2, Lemma 2) as:

$$
\begin{aligned}
f\text{-SPG}(p; g) :=& f\text{-SPG}(S_p(g), C_p) \\
:=& C_p * \partial f(S_p(g)) - f^\star(\partial f(S_p(g))),
\end{aligned}
$$

where $C_p := C(y, y'),\ \ S_p(g) := S(g(x), g(x'))$, ∂f is the subdifferential of f, and $f^\star$ is the convex conjugate of f (formally defined in Section 8.3.2).

Given a set of pairs $I = \{[(x, y), (x', y')], ...\}$, we define the total f-SPG as the sum of the individual f-SPG:

$$f\text{-SPG}(I; g) := \sum_{p \in I} f\text{-SPG}(p; g).$$

Fixing the set I, we seek g that maximizes f-SPG$(I; g)$. When the convex function f is differentiable and ∂f is invertible, and the set, I, satisfies a *balance condition*,

$$\sum_{p \in I} (C_p - 1) = 0,$$

our main theorem (Theorem 3) in Section 8.3.2 will show that: (1) the maximizer of f-SPG, g^*, preserves the pairwise similarity, that is, for every $p = [(x, y), (x', y')] \in I$, $C(y, y') = S(g^*(x), g^*(x'))$; (2) the maximum of f-SPG represents the amount of information contained in the pairs.

f-SPL We convert f-SPG to a loss function, f-SPL, so that it can be used as a component of neural network training. To do this, we identify the maximal point of f-SPG at which the label similarity is equal to the feature similarity, f-SPG(C_p, C_p), and subtract from it f-SPG$(S_p(g), C_p)$:

$$\begin{aligned} &f\text{-SPL}(p;g) := f\text{-SPG}(C_p, C_p) - f\text{-SPG}(S_p(g), C_p) \\ &\text{and} f\text{-SPL}(I;g) := \sum_{p\in I} f\text{-SPL}(p;g). \end{aligned}$$

As a result, f-SPL has the following properties: (1) f-SPL is always non-negative; (2) minimizing f-SPL$(I;g)$ over g is equivalent to maximizing f-SPG$(I;g)$ over g. Table 8.1 shows five special cases of f-SPL based on the convex functions corresponding to common f-divergence measures.

8.3.2 Theoretical Justifications

We will show the feature learning function, g^*, that minimizes f-SPL$(I;g)$ to zero and maximizes f-SPG$(I;g)$ to the amount of information contained in the set I, also preserves the pairwise similarity of I.

To give the theoretical justification, we first give the formal definition of f-divergence and its dual form.

f-divergence [6, 34] Given set Σ and the set of all possible distributions over Σ, Δ_Σ, f-divergence $D_f : \Delta_\Sigma \times \Delta_\Sigma \mapsto \mathbb{R}$ is a non-symmetric measure of the difference between two distributions, $\mathbf{p}, \mathbf{q} \in \Delta_\Sigma$, and is defined as

$$D_f(\mathbf{p}, \mathbf{q}) = \sum_{\sigma\in\Sigma} \mathbf{p}(\sigma) f\left(\frac{\mathbf{q}(\sigma)}{\mathbf{p}(\sigma)}\right)$$

where $f : \mathbb{R} \mapsto \mathbb{R}$ is a convex function and $f(1) = 0$. Two common special cases of f-divergence are KL divergence and total variation distance.

Definition 1 (Fenchel Duality [128]). *Given any function $f : \mathbb{R} \mapsto \mathbb{R}$, we define its* *convex conjugate $f^\star$ as a function that also maps $\mathbb{R}$ to $\mathbb{R}$ such that*

$$f^\star(x) = \sup_t tx - f(t).$$

Lemma 2 (Dual form of f-divergence [109, 110])**.**

$$\begin{aligned} D_f(\mathbf{p}, \mathbf{q}) &\geq \sup_{u \in \Sigma} \mathbb{E}_{\mathbf{p}} u - \mathbb{E}_{\mathbf{q}} f^\star(u) \\ &= \sup_{u \in \mathcal{G}} \sum_\sigma u(\sigma)\mathbf{p}(\sigma) - \sum_\sigma f^\star(u(\sigma))\mathbf{q}(\sigma) \end{aligned}$$

where $\mathcal{G}$ is a set of functions that map Σ to $\mathbb{R}$. The equality holds if and only if $u(\sigma) = u^*(\sigma) \in \partial f\left(\frac{\mathbf{p}(\sigma)}{\mathbf{q}(\sigma)}\right)$, i.e., the subdifferential of f on value $\frac{\mathbf{p}(\sigma)}{\mathbf{q}(\sigma)}$.

We define $\mathbf{C}$ and $\mathbf{1}$ as distributions over the pairs in I such that $\mathbf{C}(p) = \frac{C_p}{\sum_{p \in I} C_p}$ and $\mathbf{1}(p) = \frac{1}{|I|}$ for all $p \in I$. $D_f(\mathbf{C}, \mathbf{1})$ measures the amount of information contained in the chosen pairs: when all chosen pairs are neither very similar nor very dissimilar, $\mathbf{C}$ is close to $\mathbf{1}$, which implies a small amount of information contained in the pairs; when all chosen pairs are either very similar or very dissimilar, $\mathbf{C}$ is far away from $\mathbf{1}$, which implies a large amount of information contained in the pairs.

We use the above lemma to show our main theorem:

Theorem 3. *Given a convex function f, a balanced set I, when f is differentiable and ∂f is invertible, for every minimizer g^* of f-SPL$(I; g)$, for every $[(x, x'), (y, y')] \in I$,*

$$S(g^*(x), g^*(x')) = C(y, y').$$

g^* minimizes f-SPL$(I; g)$ to zero and maximizes f-SPG$(I; g)$ to $D_f(\mathbf{C}, \mathbf{1})$.

Proof. The balance condition implies that

$$\sum_{p \in I} C_p = |I|$$

Thus, by dividing $|I|$, we can rewrite f-SPG$(I; g)$ as

$$\sum_{p \in I} \partial f(S_p(g))) * \mathbf{C}(p) - f^\star(\partial f(S_p(g))) * \mathbf{1}(p).$$

Based on Lemma 2, for every maximizer g^* of f-SPG$(I; g)$/minimizer of f-SPL$(I; g)$, we have

$$\partial f(S_p(g^*)) = \partial f\left(\frac{\mathbf{C}(p)}{\mathbf{1}(p)}\right) = \partial f(C_p)$$

for every $p \in I$ and the maximum of f-SPG is $D_f(\mathbf{C}, \mathbf{1})$, which also implies the minimum of f-SPL is zero.

Therefore, when f is differentiable and ∂f is invertible, g^* preserves the pairwise similarity of the pairs in I.

□

8.3.3 Multi-Task Framework

In this work, we use a multi-task framework that jointly reduces classification loss and f-SPL, as shown in Figure 8.1. The first block of neural network layers corresponds to g and the second block of layers is denoted as ω. Previous work has demonstrated the efficacy of using DML loss with hard labels within multi-task frameworks [90, 204]. We hypothesize that DML loss will also enhance classification

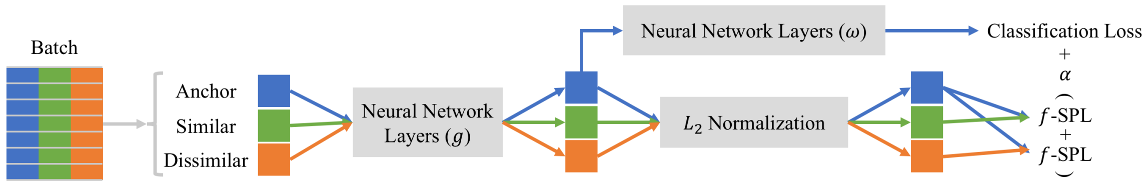


Figure 8.1: The proposed multi-task framework for training. The inputs are batches of triplets. The loss is the combination of the classification loss calculated on the anchor a only and the f-SPL between a and a similar example s plus a and a dissimilar example d (s.t., $C(y_a, y_s) + C(y_a, y_d) \approx 2$). α is the weighting term for f-SPL. The test phase does not depend on triplets.

performance given soft labels. The classification loss provides direction for the optimization, while f-SPL, calculated on the output of an intermediate layer, enforces that the learned representation preserves pairwise similarity.

Recall that the theoretical guarantee of the f-SPL is subject to a *balanced condition*:

$$\sum_{p \in I}(C_p - 1) = 0,$$

where $C : A_Y \times A_Y \mapsto [0, 2]$ is the label similarity. We wish to satisfy this condition regardless of data shuffling or the selection of batch size, while still allowing for randomness. Therefore, we generate the pairs in a triplet form. For each anchor (x_a, y_a), we pick a similar example (x_s, y_s) and a dissimilar example (x_d, y_d) that satisfy $C(y_a, y_s) - 1 \approx 1 - C(y_a, y_d)$. Specifically, we calculate the *label similarity* between the anchor and all other examples (can be reduced to a subset of examples, if the training set is very large) and generate a dictionary with unique similarity values (rounded to two decimal point) as keys and utterance indices as values. We keep a key only if 2-key is also in the dictionary. When generating a triplet, we randomly select a key c, and two examples, each from c and $2 - c$, respectively. As a result, every batch of triplets

$$T = \{tri = [(x_a, y_a), (x_s, y_s), (x_d, y_d)], ...\}$$

naturally implies a balanced set

$$I_T = \{s = [(x_a, y_a), (x_s, y_s)], d = [(x_a, y_a), (x_d, y_d)], ...\}.$$

The overall loss function for each triplet, tri, is

$$\begin{aligned}L(tri; g, \omega) = L_{cls}(y_a, \hat{y_a}) + \\ \alpha(f\text{-SPL}(S_s(g), C_s) + f\text{-SPL}(S_d(g), C_d)),\end{aligned}$$

where $\hat{y_a} = \omega(g(x_a))$ is the prediction over classes.

In the loss function, L_{cls} is the classification loss calculated on the anchor only, and α is the trade off between L_{cls} and f-SPL. C_s and C_d are the label similarity between y_a and y_s, y_a and y_d, respectively. $S_s(g)$ and $S_d(g)$ are the similarity between $g(x_a)$ and $g(x_s)$, $g(x_a)$ and $g(x_d)$, respectively.

The total loss of the batch is the mean of all triplets' losses: $L(T; g, \omega) := \frac{1}{N}\sum_{tri \in T} L(tri; g, \omega)$, where N is the batch size. Note that the f-SPL portion of $L(T; g, \omega)$ equals $\frac{\alpha}{N} f\text{-SPL}(I_T; g)$.

The multi-task framework is only used in the training phase. In the testing phase, the trained network takes batches of individual examples as the input.

8.4 Experiments

We experiment on IEMOCAP [25] and MSP-Improv [26]. We select these datasets because: (1) they are relatively large, which allows us to train neural networks; (2) they provide ordinal evaluations of valence and activation; (3) they use similar emotion elicitation methods, but differ in speakers, lexical content, recording conditions, and the number of evaluations per utterance.

All experiments use cross-corpus evaluation. This results in four experiments (2 training-testing combination × 2 dimensions). We introduce the data, model, and experimental settings in more detail in the following subsections.

8.4.1 Data

8.4.1.1 Labels

We focus on predicting binary valence and activation, where the classifiers are trained using soft labels. We consider each evaluation as a vote to the two classes, weighted by the distance to the opposite class. For example, an evaluation value of 2 on the 5-point scale is converted to [0.75, 0.25]. For each utterance, we average over the converted evaluation from each annotator and use the resulting two-dimensional vector as the final soft label. The label similarity, $C \in [0, 2]$, is calculated by $2 - 2d$, where d is the total variation distance ($\in [0, 1]$) between a pair of labels. Given the way we generate the soft labels, d is equivalent to the scaled Euclidean distance between the average of the raw evaluations on the one-dimensional space.

8.4.1.2 Features

We preprocess the data such that the audio sampling rate is 16,000 Hz for both datasets. We then extract 40-dimensional log Mel-frequency Filterbank energy (MFB) using Kaldi [124], with a frame size of 25ms and a step size of 10ms, as in [3, 4, 212]. We perform z-normalization for each feature dimension at the frame-level over each dataset, individually.

8.4.2 Classification Model

We use temporal Convolutional Neural Networks with global pooling (*Conv-Pool*) as our model. The Conv-Pool structure has been demonstrated to be the state-of-the-art on categorical emotion recognition in [3], and has shown good performance on predicting the distribution of emotion perception in [212]. Figure 8.2 shows the architecture the network. It consists of a 1D convolutional layer over time with 128 number of kernels and a kernel width of 16, a global max pooling, two fully-connected

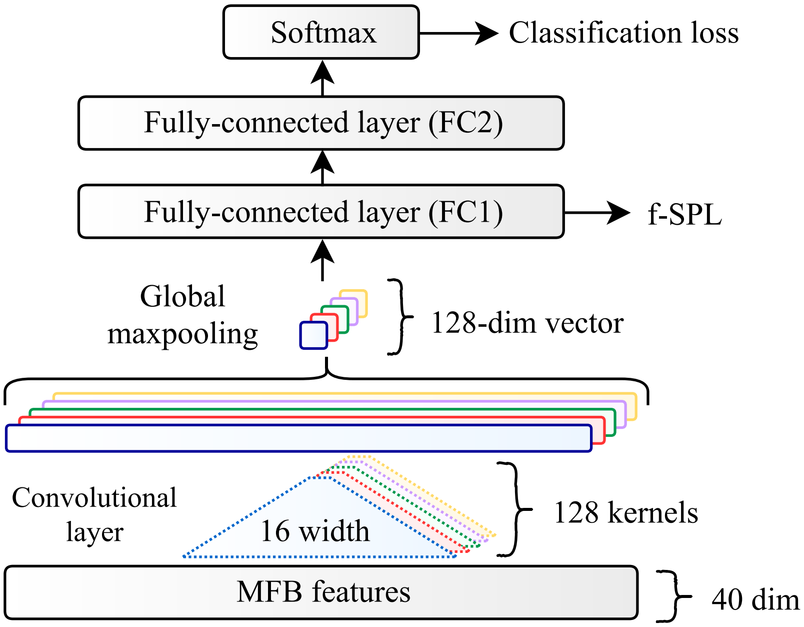


Figure 8.2: The Conv-Pool network structure.

layers with a layer size of 128, and a final fully connected softmax layer. These hyper-parameters were selected according to [3, 212]. The inputs to network are the variable-length MFBs. The global max-pooling layer summarizes the output of the 1D convolutional layer and generates a fixed-length representation. This representation is then fed into the fully-connected layers. We use Rectified Linear Units (ReLU) as the activation functions, except in the last fully-connected layer, where softmax is used instead.

We calculate f-SPL on the output of FC1 (see Figure 8.2). In this way, we allow room for modeling non-linearity on both sides of the intermediate representation. We first normalize the output of FC1 to unit vectors and then calculate the Euclidean distance between the embeddings. It is worth noting that although the distance, D, between two unit vectors has a range of $[0, 2]$, our embeddings have non-negative entries due to ReLU and thus $D \in [0, \sqrt{2}]$. Therefore, we scale the distance and convert it to the embedding similarity $S \in [0, 2]$ by $2 - \sqrt{2}D$.

The structure of the model is kept the same in all experiments. We use cross-entropy computed using the soft labels as the classification loss. The vector representing a soft label always sums to one. We weigh the two classes using $N/\left(2\sum_{i=1}^{N} y_i^c\right)$ in the loss calculation to reduce the influence of data imbalance. Here, N is the total number of training utterances, y_i^c is the value for class c in the label vector of data point i. We consider a loss function containing only the cross-entropy classification loss as the baseline. For the multi-task loss, we select f-SPL based on the convex functions corresponding to five common f-divergence measures, including KL divergence, Reverse KL divergence, Pearson χ^2, Squared Hellinger, and Jensen-Shannon Divergence, as shown in Table 8.1. An epsilon value of 1e-12 is added to the denominators and the input of log in f-SPL in implementation for numerical stability.

For the multi-task frameworks, every training example is used as the anchor once in each epoch. Therefore, the classification loss is calculated over the same data as in the baseline. The triplets are randomly generated at the beginning of each epoch using the procedure introduced in Section 8.3.3. Empirical results show that the values of f-SPL is about a magnitude smaller than the classification loss, because triplets with extreme similarity values are rare in our data. Therefore, we use a α value of 10 in the loss function.

8.4.3 Performance Measure and Cross-Validation

In the testing phase, we convert the output of the network to a class prediction. We use Unweighted Average Recall (UAR) as the performance measure due to data imbalance, as discussed in [131]. In the case that the ground truth labels are tied (i.e., [0.5, 0.5]), we consider predictions for either class as correct, as in [4]. This is true for both the baseline CNN and f-SPL approaches. As a result, the chance performance calculated by generating predictions uniformly at random is higher than 50%.

We conduct the experiments using PyTorch[1], with a learning rate of 0.0001

[1] http://pytorch.org

with the Adam optimizer [81] and a batch size of 100. We select weight decay in $\{0, 0.0001, 0.001, 0.01\}$ and the number of epochs to train in the range of $[1, 50]$ by leave-one-session-out cross-validation (LOSOCV). In each experiment (e.g., valence classification, train on IEMOCAP and test in MSP-Improv), the weight decay and number of training epochs that lead to the highest LOSOCV UAR of the baseline model (averaged over three runs) are used for all models. In the cross-corpus training and testing, we run each experiment 30 times to reduce performance fluctuations. We report the average UAR and conduct significance tests using the results from all runs.

8.5 Results and Discussion

8.5.1 Performance Comparison

We present the UAR of the four experiments (2 training-testing combinations × 2 dimensions) of all the models in Table 7.3. Each reported UAR is averaged over 30 runs. All cross-validation experiments selected the same weight-decay value of 0.001. The models include:

- CE: Conv-Pool network (Figure 8.2) with only cross-entropy classification loss. This is used as the baseline.

- CE+f, where $f \in$ KL, RKL, PS, HLG, JS: Conv-Pool network using the multi-task framework illustrated in Figure 8.1, with the convex functions corresponding to KL divergence, Reserve-KL, Pearson χ^2, Squared Hellinger, and JS divergence as f for f-SPL.

For each experiment, we first test if the influence of model is significant, using a one-way Analysis of variance (ANOVA) test and asserting significance at p <0.05. We treat the result of each run as a random example, and group them by the model.

	MSP→IEMOCAP		IEMOCAP→MSP	
	Valence	Activation	Valence	Activation
Epoch	23	33	20	16
Chance	58.41	62.74	54.43	54.21
CE	64.45	80.79	<u>**61.31**</u>	72.74
CE+KL	<u>**66.01**</u>*	<u>**81.67**</u>*	60.81	74.17*
CE+RKL	65.73*	81.16	60.71	<u>**74.30**</u>*
CE+PS	65.55*	81.53*	60.48	74.17*
CE+HLG	65.76*	81.10	60.78	74.04*
CE+JS	65.94*	81.41	60.44	74.03*

Table 8.2: UAR (%) for the four cross-corpus experiments. The best performance in each experiment is marked by bold and underline. Epoch: the number of epochs trained; CE: cross-entropy loss only (baseline); CE+KL, CE+RKL, CE+PS, CE+HLG, and CE+JS: multi-task with cross-entropy and f-SPL, where f is KL divergence, Reserve-KL, Pearson χ^2, Squared Hellinger, and JS divergence, respectively. "*" indicates that the marked performance is significantly better than CE, where significance is assessed at p <0.05 using the Tukey's honest test on the ANOVA statistics.

This results in 180 examples (30 runs×6 models) in each test. We find that the influence of model is significant for valence when training on MSP-Improv and testing on IEMOCAP (denoted as *MSP→IEMOCAP Valence*), and for activation with both training-testing combinations. The statistics are F(5,174)=8.1, p=6.9e-7 for *MSP→IEMOCAP Valence*, F(5,174)=3.7, p=0.0033 for *MSP→IEMOCAP Activation*, and F(5,174)=9.7, p=3.3e-8 for *IEMOCAP→MSP Activation*, respectively.

We find that in three out of four experiments, all the five CE+f models show consistent performance improvement over the baseline CE model, with the only exception of *IEMOCAP→MSP Valence*. For the experiments where the influence of model is significant, we conduct pairwise comparisons using the Tukey's honest test on the statistics of the ANOVA and assert significance at p <0.05. We find that in *MSP→IEMOCAP Valence*, all five CE+f models are significantly better than CE, with p = 6.9e-7, 1.0e-4, 0.0017, 6.9e-5, and 2.6e-6 for CE+KL, CE+RKL, CE+PS, CE+HLG, and CE+JS, respectively. In *MSP→IEMOCAP Activation*, CE+KL and

CE+PS has significantly higher UAR than CE (p=0.0028 and 0.022, respectively). In *IEMOCAP→MSP Activation*, all the five CE+f models have significantly better performance than CE. The p-values are 9.1e-7, 6.7e-8, 8.8e-7, 1.2e-5, and 1.3e-5 for CE+KL, CE+RKL, CE+PS, CE+HLG, and CE+JS, respectively. We do not observe any significant difference between the performances of the five CE+f models in any experiments.

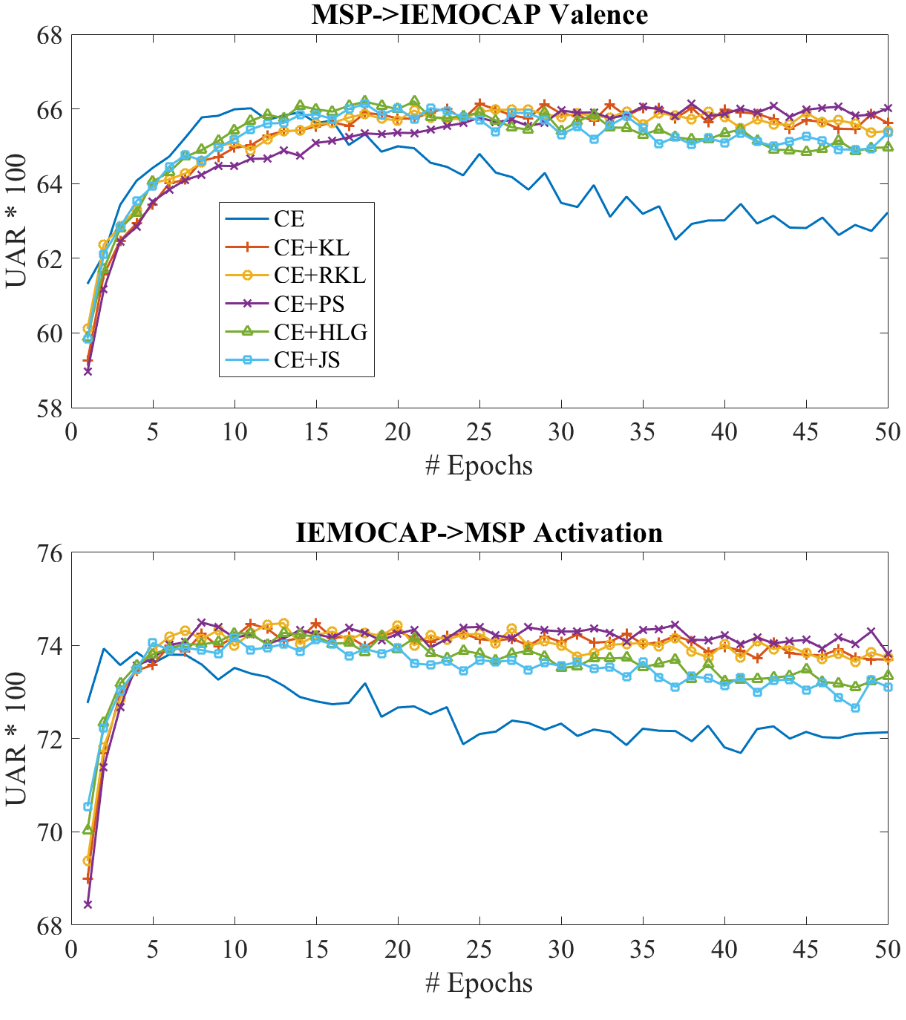


Figure 8.3: Test UAR against the number of training epochs for *MSP→IEMOCAP Valence* and *IEMOCAP→MSP Activation*.

8.5.2 Analysis of Results

We further analyze the results to better understand the reasons behind the improvement in performance. We plot the test UAR against training epochs in Figure 8.3 for the two experiments where the CE+f models achieved the highest performance gain over the baseline CE model. We find that while the optimal results from different models do not differ much, the CE+f models are more stable over time. More specifically, the CE model reaches the best UAR around epoch 10 in *MSP→IEMOCAP Valence* and within 5 epochs in *IEMOCAP→MSP Activation*. It starts to show signs of over-training after that, even before reaching the number of epochs to train we set and with weight-decay, when both hyper-parameters are selected by cross-validation. In contrast, the proposed CE+f models with the exact same hyper-parameters do not show obvious performance decline after reaching the highest UAR.

We visualize the learned feature embeddings at epoch 10, 30, and 50 for *IEMOCAP → MSP Activation* with t-Distributed Stochastic Neighbor Embedding in Figure 8.4.

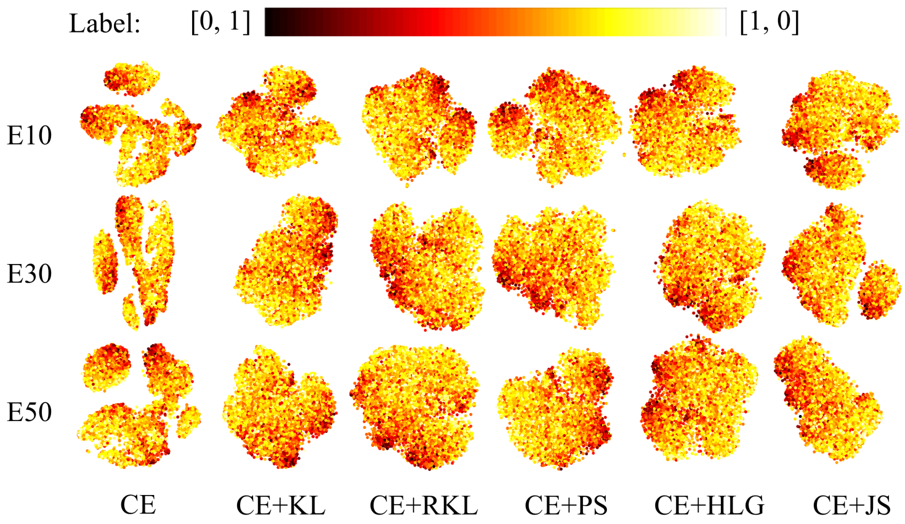


Figure 8.4: *IEMOCAP→MSP Activation* embedding visualization. Dots are data points in MSP. Colors represent the activation labels, the darkest are [0, 1] and the lightest are [1,0]. The rows are the embeddings at epoch 10, 30, 50 (e.g., E50). The columns correspond to the six models.

The color of the dots in the figure represents the soft labels. The dark end of the color gradient represents [0, 1] and the light end represents [1, 0]. We find that the baseline CE models lead to several clusters, but the clusters do not correspond to labels. On the other hand, the CE+f models often lead to a single cluster where the opposite labels are more well separated and the data that are more uncertain (e.g., label $\sim$ [0.5, 0.5]) are in between. This shows that we can learn an embedding that has emotional meaning using a multi-task framework combining classification loss and f-SPL.

8.6 Conclusions

In this work, we propose a family of loss functions, f-Similarity Preservation Loss, based on the dual form of f-divergence. These loss functions are designed for deep metric learning with soft labels, label with continuous values along one or multiple dimensions. We prove mathematically that the minimizer of the proposed loss functions, a set of nonlinear mappings through neural networks, preserves the pairwise label similarities in the learned feature embeddings when the pairs of data satisfy a balanced condition. We propose a pair sampling method that guarantees the balanced condition regardless of shuffling and batch size without losing randomness. Finally, we introduce a framework that combines f-SPL with the traditional classification loss.

We apply the proposed methods on the task of cross-corpus speech emotion recognition with dimensional emotion descriptors. We show that our methods significantly outperform the baseline model, which uses only the classification loss for optimization. This demonstrates the efficacy of our f-SPL in the multi-task framework. Further analysis shows that our methods are more robust to over-training and are able to learn an emotionally-meaningful embedding space.

CHAPTER 9

Conclusions and Future Directions

In this dissertation, we have explored methods for handling the variability that presents in the data and labels used in speech emotion recognition, with the goal of creating more robust and generalizable emotion recognition systems. This chapter summarizes the main results and contributions of our work and discusses potential future directions.

9.1 Main Results and Contributions

The first part of this dissertation focused on handling data variability. In Chapter 4, We approximated variability caused by a mixture of different influences by considering the training corpus as an explicit factor and took the impact of gender into account. These factors defined our tasks in a multi-task learning approach. By allowing information sharing across the tasks, we were able to create models that generalize better across datasets with different expression styles (i.e., acted and spontaneous) and languages (i.e., English and German). Our models outperformed the state-of-the-art cross-corpus speech emotion recognition systems from literature. We explored the influence of phonetic information on valence recognition from speech in Chapter 5. We proposed a multi-stage fusion approach that repeatedly added phonetic features into a single model. The proposed model was able to exploit both the acoustic and the lexical properties of phonemes and showed robustness to unscripted speech data with imperfect transcriptions.

The second part of this dissertation explored methods for addressing the variability present in emotion labels, including the discrepancy between self-perception and the perception of others, and the differences in emotion perception across a group of individuals. We investigated how self-reported and perceived emotion could be modeled jointly to enhance the overall recognition ability (Chapter 6). We showed that jointly modeling the two types of labels using multi-task learning, combined with unsupervised feature learning, could improve the performance of emotion recognition systems on unseen speakers for both types of labels. We subsequently modeled how a group of people perceive the same emotional message in Chapter 7. We proposed a label processing method that can generate probability distributions on the two-dimensional valence-activation space using a limited number of ordinal labels. We further demonstrated that using a dynamic approach and collectively predicting the distributions as a whole significantly improved system performance on unseen speakers, compared to a static approach that optimized for different regions on the distribution space separately. Finally, we showed that the combination of acoustic and visual modalities results in better performance compared to using each individual modality.

The final part of this dissertation takes both data and label variability into account. We proposed a family of similarity preservation loss for deep metric learning with soft labels, i.e., labels with continuous values along one or multiple dimensions. The minimizer of the proposed loss functions preserves the real-valued pairwise label similarities. We further presented a pair sampling method for implementing the loss functions in neural networks. We demonstrated the efficacy of the proposed methods on cross-corpus speech emotion recognition, using soft labels that embodied the uncertainty in emotion perception. We showed that our methods result in systems that generalize better across corpus and more robust to over-training. The proposed methods were for preserving emotional similarity and reducing the influence of non-emotional variability.

9.2 Future Directions

The work presented in this dissertation aims to improve the generalizability of speech emotion recognition. The ultimate goal behind this is to develop systems that can work well in real-world applications. Future work will explore the following directions, moving towards emotion recognition “in the wild”.

9.2.1 Online Speech Emotion Detection

The unifying factor of our work up to this point has been the focus on the offline setting. The data are cleaned and segmented to include a single speaker per utterance, and the systems can make predictions based on a (semi-)complete speaker turn, sentence, or affective burst. However, many applications of speech emotion recognition, such as augmented driving, augmented homes and call centers, may require the systems to predict users’ emotional states in real-time. These applications call for the development of online speech emotion recognition systems.

A challenge that arises with such online systems is the need to integrate emotion detection with other speech signal processing tasks, such as speech detection, speaker verification, and speech recognition. For example, emotion recognition systems depend on segments with sufficient speech signals; the system may need to link emotion to the person producing it; the understanding of speech and emotion co-occur in most cases. Besides, previous work [4, 72] has demonstrated that incorporating the lexical modality greatly improves the accuracy of speech emotion recognition, especially for valence.

However, the time-dependency between these tasks, specifically, between automatic speech recognition (ASR) and emotion detection, may result in significant delay in emotion predictions. Our work in Chapter 5 provide a potential direction to lessen the time-dependency. We modeled speech content as phoneme sequences instead of word sequences (a lexical representation). This may lead to a reduced requirement for ASR systems: a language model may not be necessary if the phonetic information can be extracted with a sufficient degree of accuracy. As a result, the decoding process of ASR using the language model can be parallelized with emotion recognition. Future work will explore the feasibility of performing emotion recognition by combining audio and phoneme predictions output by the acoustic model of ASR systems.

9.2.2 Personalized Speech Emotion Recognition Systems

Our work has focused on creating speech emotion recognition systems that are generalizable, i.e., systems that capture the emotion-related signals that are common across individuals. Future work will explore ways to personalize these systems given a small amount of emotion data from an individual.

Many potential applications of speech emotion recognition involve close interactions between the system and a single individual or a small number of people, given a user terminal (e.g., a car for augmented driving, an Amazon Echo or Google Home device for augmented home). This requires attention to the personalization of speech emotion recognition systems, in order to better suit the need of each user. A challenge of personalized systems is the lack of emotion data from end users: (1) the spontaneity of emotion expression makes it hard to collect sufficient amount of emotion data to train a system from scratch; (2) it may not be reasonable for the systems to require ground truth labels from the users when an emotional expression is captured. Therefore, it may be more feasible to adapt an existing model that already work well with a small amount of data, data with uncertain labels that come from implicit user feedback.

In Chapter 8, we introduced a family of loss functions for preserving real-valued pairwise similarity. We demonstrated that this method could learn an embedding space with emotional meanings. In the future, we are interested in exploring whether our proposed methods can also be effectively applied to personalization using transfer learning, with a small set of uncertain labeled data from each user.

9.3 Work Published

The work presented in this dissertation was published in or included in the submission of the following papers or book chapters:

- Part of Chapter 1, in: Biqiao Zhang and Emily Mower Provost. "Automatic recognition of self-reported and perceived emotions". In: *Multimodal Behavior Analysis in the Wild: Advances and Challenges*. Ed. by Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. To appear. Elsevier, 2018

- Part of Chapter 1 - 3, and Chapter 4, in: Biqiao Zhang, Emily Mower Provost, and Georg Essl. "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences". In: *IEEE Transactions on Affective Computing* (2017). Early Access. DOI: 10.1109/TAFFC.2017.2684799 ©2017 IEEE. Reprinted, with permission.

- Part of Chapter 1 - 3, and Chapter 5, in Biqiao Zhang, Soheil Khorram, and Emily Mower Provost. "Exploiting Acoustic and Lexical Properties of Phonemes to Recognize Valence from Speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. In submission. 2019

- Part of Chapter 1 - 3, and Chapter 6, in: Biqiao Zhang, Georg Essl, and Emily Mower Provost. "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?" In: *ACM International Conference on Multimodal*

Interaction. 2016, pp. 217–224. DOI: 10.1145/2993148.2993173

- Part of Chapter 1 - 3, and Chapter 7, in: Biqiao Zhang, Georg Essl, and Emily Mower Provost. “Predicting the distribution of emotion perception: Capturing inter-rater variability”. In: *ACM International Conference on Multimodal Interaction*. 2017, pp. 51–59. DOI: 10.1145/3136755.3136792
- Part of Chapter 1 - 3, and Chapter 8, in: Biqiao Zhang et al. “f-Similarity Preservation Loss for soft labels: A demonstration on cross-corpus speech emotion recognition”. In: *AAAI Conference on Artificial Intelligence*. To appear. 2019

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Mohammed Abdelwahab and Carlos Busso. “Supervised domain adaptation for emotion recognition from speech”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2015, pp. 5058–5062.

[2] Arvind Agarwal, Samuel Gerber, and Hal Daume. “Learning multiple tasks using manifold regularization”. In: *Advances in Neural Information Processing Systems.* 2010, pp. 46–54.

[3] Zakaria Aldeneh and Emily Mower Provost. “Using regional saliency for speech emotion recognition”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2017, pp. 2741–2745.

[4] Zakaria Aldeneh et al. “Pooling acoustic and lexical features for the prediction of valence”. In: *ACM International Conference on Multimodal Interaction.* 2017, pp. 68–72.

[5] Sharifa Alghowinem et al. “From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech.” In: *FLAIRS.* 2012.

[6] Syed Mumtaz Ali and Samuel D Silvey. “A general class of coefficients of divergence of one distribution from another”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1966), pp. 131–142.

[7] Timur Almaev, Brais Martinez, and Michel Valstar. “Learning to transfer: Transferring latent task structures and its application to person-specific facial action unit detection”. In: *IEEE International Conference on Computer Vision.* 2015, pp. 3774–3782.

[8] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. “Convex multi-task feature learning”. In: *Machine Learning* 73.3 (2008), pp. 243–272.

[9] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. “Multi-task feature learning”. In: *Advances in neural information processing systems.* 2007, pp. 41–48.

[10] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. “Openface: An open source facial behavior analysis toolkit”. In: *IEEE Winter Conference on Applications of Computer Vision.* 2016, pp. 1–10.

[11] Tanja Bänziger, Sona Patel, and Klaus R Scherer. “The role of perceived voice and speech characteristics in vocal emotion communication”. In: *Journal of Nonverbal Behavior* 38.1 (2014), pp. 31–52.

[12] A Batliner, S Steidl, and E Nöth. “Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus”. In: *Satellite Workshop of Language Resources and Evaluation Conference.* 2008, pp. 28–31.

[13] Yoshua Bengio. “Practical recommendations for gradient-based training of deep architectures”. In: *Neural Networks: Tricks of the Trade.* 2012, pp. 437–478.

[14] Dmitri Bitouk, Ragini Verma, and Ani Nenkova. “Class-level spectral features for emotion recognition”. In: *Speech communication* 52.7-8 (2010), pp. 613–625.

[15] Zdravko I Botev, Joseph F Grotowski, Dirk P Kroese, et al. “Kernel density estimation via diffusion”. In: *The Annals of Statistics* 38.5 (2010), pp. 2916–2957.

[16] Mátyás Brendel et al. “Towards measuring similarity between emotional corpora”. In: *Satellite Workshop of Language Resources and Evaluation Conference.* 2010, pp. 58–64.

[17] Jane Bromley et al. “Signature verification using a “siamese” time delay neural network”. In: *Advances in neural information processing systems.* 1994, pp. 737–744.

[18] Egon Brunswik. “Representative design and probabilistic theory in a functional psychology.” In: *Psychological Review* 62.3 (1955), p. 193.

[19] Ross Buck. *The communication of emotion.* Guilford Press, 1984.

[20] Felix Burkhardt et al. “A database of german emotional speech”. In: *INTERSPEECH.* Vol. 5. 2005, pp. 1517–1520.

[21] Alec Burmania, Mohammed Abdelwahab, and Carlos Busso. “Tradeoff between quality and quantity of emotional annotations to characterize expressive behaviors”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2016, pp. 5190–5194.

[22] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. “Increasing the reliability of crowdsourcing evaluations using online quality assessment”. In: *IEEE Transactions on Affective Computing* 7.4 (2016), pp. 374–388.

[23] Carlos Busso, Sungbok Lee, and Shrikanth S Narayanan. "Using Neutral Speech Models for Emotional Speech Analysis". In: *Annual Conference of the International Speech Communication Association.* 2007, pp. 2225–2228.

[24] Carlos Busso and Shrikanth S Narayanan. "The expression and perception of emotions: Comparing assessments of self versus others". In: *INTERSPEECH.* 2008, pp. 257–260.

[25] Carlos Busso et al. "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language Resources and Evaluation* 42.4 (2008), p. 335.

[26] Carlos Busso et al. "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception". In: *IEEE Transactions on Affective Computing* 8.1 (2017), pp. 67–80.

[27] Robert J Campbell, Norman Kagan, and David R Krathwohl. "The development and validation of a scale to measure affective sensitivity (empathy)". In: *Journal of Counseling Psychology* 18.5 (1971), p. 407.

[28] Walter B Cannon. "The James-Lange theory of emotions: A critical examination and an alternative theory". In: *The American Journal of Psychology* 39.1/4 (1927), pp. 106–124.

[29] Chih-Chung Chang and Chih-Jen Lin. "LIBSVM: A Library for Support Vector Machines". In: *ACM Transactions on Intelligent Systems and Technology* 2.3 (2011), p. 27.

[30] 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.* 2017, pp. 2746–2750.

[31] Sumit Chopra, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *Computer Vision and Pattern Recognition.* Vol. 1. 2005, pp. 539–546.

[32] Cristina Conati. "Probabilistic assessment of user's emotions in educational games". In: *Applied Artificial Intelligence* 16.7-8 (2002), pp. 555–575.

[33] Koby Crammer and Yishay Mansour. "Learning multiple tasks using shared hypotheses". In: *Advances in Neural Information Processing Systems.* 2012, pp. 1475–1483.

[34] Imre Csiszár, Paul C Shields, et al. "Information theory and statistics: A tutorial". In: *Foundations and Trends® in Communications and Information Theory* 1.4 (2004), pp. 417–528.

[35] Laurence Devillers and Laurence Vidrascu. “Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs”. In: *International Conference on Spoken Language Processing.* 2006.

[36] Ellen Douglas-Cowie et al. “The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data”. In: *International Conference on Affective computing and intelligent interaction* (2007), pp. 488–500.

[37] Andreas Doumanoglou et al. “Siamese regression networks with efficient mid-level feature extraction for 3d object pose estimation”. In: *arXiv preprint arXiv:1607.02257* (2016).

[38] Paul Ekman. “An argument for basic emotions”. In: *Cognition & Emotion* 6.3-4 (1992), pp. 169–200.

[39] Paul Ekman. “Strong evidence for universals in facial expressions: A reply to Russell’s mistaken critique”. In: (1994).

[40] Paul Ekman. *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition).* WW Norton & Company, 2009.

[41] Paul Ekman and Wallace V Friesen. “Facial action coding system”. In: (1977).

[42] Paul Ekman and Wallace V Friesen. “Nonverbal behavior and psychopathology”. In: *The Psychology of Depression: Contemporary Theory and Research* (1974), pp. 3–31.

[43] Paul Ekman et al. “Universals and cultural differences in the judgments of facial expressions of emotion”. In: *Journal of Personality and Social Psychology* 53.4 (1987), p. 712.

[44] Hillary Anger Elfenbein and Nalini Ambady. “On the universality and cultural specificity of emotion recognition: A meta-analysis”. In: *Psychological Bulletin* 128.2 (2002), p. 203.

[45] Hillary Anger Elfenbein and Nalini Ambady. “When familiarity breeds accuracy: Cultural exposure and facial emotion recognition”. In: *Journal of Personality and Social Psychology* 85.2 (2003), p. 276.

[46] Phoebe C Ellsworth and Klaus R Scherer. “Appraisal processes in emotion”. In: *Handbook of Affective Sciences* 572 (2003), p. V595.

[47] Inger Samso Engberg and Anya Varnich Hansen. “Documentation of the danish emotional speech database DES”. In: *Internal AAU report, Center for Person Kommunikation, Denmark* (1996).

[48] Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: the Munich Versatile and Fast Open-Source Audio Feature Extractor". In: *ACM International Conference on Multimedia*. 2010, pp. 1459–1462.

[49] Florian Eyben et al. "On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues". In: *Journal on Multimodal User Interfaces* 3.1 (2010), pp. 7–19.

[50] Rong-En Fan et al. "LIBLINEAR: A library for large linear classification". In: *The Journal of Machine Learning Research* 9 (2008), pp. 1871–1874.

[51] Yin Fan et al. "Video-based emotion recognition using CNN-RNN and C3D hybrid networks". In: *ACM International Conference on Multimodal Interaction*. 2016, pp. 445–450.

[52] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. "Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels". In: *International Joint Conference on Neural Networks*. 2016, pp. 566–570.

[53] Kalani Wataraka Gamage, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. "Modeling variable length phoneme sequences - A step towards linguistic information for speech emotion recognition in wider world". In: *International Conference on Affective Computing and Intelligent Interaction*. 2017, pp. 518–523.

[54] Kalani Wataraka Gamage, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. "Salience based lexical features for emotion recognition". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2017, pp. 5830–5834.

[55] Michael Grimm, Kristian Kroschel, and Shrikanth Narayanan. "The Vera am Mittag German audio-visual emotional speech database". In: *International Conference on Multimedia and Expo*. 2008, pp. 865–868.

[56] Michael Grimm et al. "Primitives-based evaluation and estimation of emotions in speech". In: *Speech Communication* 49.10 (2007), pp. 787–800.

[57] James J Gross and Robert W Levenson. "Emotional suppression: Physiology, self-report, and expressive behavior". In: *Journal of Personality and Social Psychology* 64.6 (1993), p. 970.

[58] Hatice Gunes. "Automatic, dimensional and continuous emotion recognition". In: (2010).

[59] Judith A Hall. *Nonverbal sex differences: Accuracy of communication and expressive style*. 1990.

[60] Wenjing Han et al. “Towards Temporal Modelling of Categorical Speech Emotion Recognition”. In: *INTERSPEECH* (2018), pp. 932–936.

[61] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. “A fast learning algorithm for deep belief nets”. In: *Neural computation* 18.7 (2006), pp. 1527–1554.

[62] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. “Distilling the Knowledge in a Neural Network”. In: *Advances in Neural Information Processing Systems.* 2015.

[63] Elad Hoffer and Nir Ailon. “Deep metric learning using triplet network”. In: *International Workshop on Similarity-Based Pattern Recognition.* 2015, pp. 84–92.

[64] Zhaocheng Huang and Julien Epps. “A PLLR and multi-stage staircase regression framework for speech-based emotion prediction”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2017, pp. 5145–5149.

[65] Zhaocheng Huang and Julien Epps. “An Investigation of Partition-based and Phonetically-aware Acoustic Features for Continuous Emotion Prediction from Speech”. In: *IEEE Transactions on Affective Computing* (2018).

[66] Zhengwei Huang et al. “Speech emotion recognition using CNN”. In: *ACM International Conference on Multimedia.* 2014, pp. 801–804.

[67] Kyung Hak Hyun, Eun Ho Kim, and Yoon Keun Kwak. “Emotional feature extraction based on phoneme information for speech emotion recognition”. In: *International Symposium on Robot and Human interactive Communication.* 2007, pp. 802–806.

[68] E Izard Carroll. “Human emotions”. In: *New York Plenum* (1977).

[69] Carroll E Izard. “The face of emotion”. In: (1971).

[70] William James. “What is an emotion?” In: *Mind* 34 (1884), pp. 188–205.

[71] Qiang Ji, Peilin Lan, and Carl Looney. “A probabilistic framework for modeling and real-time monitoring human fatigue”. In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 36.5 (2006), pp. 862–875.

[72] Qin Jin et al. “Speech emotion recognition with acoustic and lexical features”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2015, pp. 4749–4753.

[73] Patrik N Juslin and Petri Laukka. “Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion.” In: *Emotion* 1.4 (2001).

[74] Zhuoliang Kang, Kristen Grauman, and Fei Sha. “Learning with whom to share in multi-task feature learning”. In: *International Conference on Machine Learning*. 2011, pp. 521–528.

[75] Arvid Kappas, Ursula Hess, and Klaus R Scherer. “6. Voice and emotion”. In: *Fundamentals of Nonverbal Behavior* (1991), p. 200.

[76] Christos D Katsis et al. “Toward emotion recognition in car-racing drivers: A biosignal processing approach”. In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38.3 (2008), pp. 502–512.

[77] Soheil Khorram et al. “Capturing Long-term Temporal Dependencies with Convolutional Networks for Continuous Emotion Recognition”. In: *INTERSPEECH*. 2017, pp. 1253–1257.

[78] Yelin Kim, Honglak Lee, and Emily Mower Provost. “Deep learning for robust feature generation in audiovisual emotion recognition”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 3687–3691.

[79] Yelin Kim and Emily Mower Provost. “Say Cheese vs. Smile: Reducing Speech-Related Variability for Facial Emotion Recognition”. In: *ACM International Conference on Multimedia*. 2014.

[80] Yoon Kim. “Convolutional neural networks for sentence classification”. In: *Empirical Methods in Natural Language Processing*. 2014.

[81] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *International Conference on Learning Representations*. 2015.

[82] Yuqing Kong and Grant Schoenebeck. “Water from Two Rocks: Maximizing the Mutual Information”. In: *Economics and Computation*. 2018, pp. 177–194.

[83] Hugo Larochelle et al. “Exploring strategies for training deep neural networks”. In: *The Journal of Machine Learning Research* 10 (2009), pp. 1–40.

[84] Chul Min Lee and Shrikanth S Narayanan. “Toward detecting emotions in spoken dialogs”. In: *IEEE Transactions on Speech and Audio Processing* 13.2 (2005), pp. 293–303.

[85] Chul Min Lee et al. “Emotion recognition based on phoneme classes”. In: *International Conference on Spoken Language Processing*. 2004.

[86] Iulia Lefter et al. “Cross-corpus analysis for acoustic recognition of negative interactions”. In: *IEEE International Conference on Affective Computing and Intelligent Interaction*. 2015, pp. 132–138.

[87] Iulia Lefter et al. “Emotion recognition from speech by combining databases and fusion of classifiers”. In: *TSD*. 2010, pp. 353–360.

[88] Lea Leinonen et al. “Expression of emotional–motivational connotations with a one-word utterance”. In: *The Journal of the Acoustical society of America* 102.3 (1997), pp. 1853–1863.

[89] Hongye Liu et al. “Deep relative distance learning: Tell the difference between similar vehicles”. In: *Computer Vision and Pattern Recognition.* 2016, pp. 2167–2175.

[90] Xiaofeng Liu et al. “Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition”. In: *Computer Vision and Pattern Recognition Workshops.* 2017, pp. 522–531.

[91] Reza Lotfian and Carlos Busso. “Formulating emotion perception as a probabilistic model with application to categorical emotion classification”. In: *International Conference on Affective computing and intelligent interaction.* 2017, pp. 415–420.

[92] Jiwen Lu, Junlin Hu, and Jie Zhou. “Deep metric learning for visual understanding: An overview of recent advances”. In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 76–84.

[93] Qirong Mao et al. “Domain adaptation for speech emotion recognition by sharing priors between related source and target classes”. In: *IEEE International Conference on Acoustics, Speech and Signal Processing.* 2016, pp. 2608–2612.

[94] Qirong Mao et al. “Learning salient features for speech emotion recognition using convolutional neural networks”. In: *IEEE Transactions on Multimedia* 16.8 (2014), pp. 2203–2213.

[95] Stacy Marsella, Jonathan Gratch, Paolo Petta, et al. “Computational models of emotion”. In: *A Blueprint for Affective Computing-A sourcebook and manual* 11.1 (2010), pp. 21–46.

[96] Olivier Martin et al. “The enterface’05 audio-visual emotion database”. In: *International Conference on Data Engineering Workshops.* 2006.

[97] Rod A Martin et al. “Emotion perception threshold: Individual differences in emotional sensitivity”. In: *Journal of Research in Personality* 30.2 (1996), pp. 290–305.

[98] David Matsumoto et al. “The contribution of individualism vs. collectivism to cross-national differences in display rules”. In: *Asian Journal of Social Psychology* 1.2 (1998), pp. 147–165.

[99] Gary McKeown et al. “The SEMAINE corpus of emotionally coloured character interactions”. In: *ICME.* 2010, pp. 1079–1084.

[100] Gary McKeown et al. “The Semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent”. In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 5–17.

[101] Albert Mehrabian. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies.* Oelgeschlager, Gunn & Hain Cambridge, MA, 1980.

[102] Joseph A Mikels et al. “Emotional category data on images from the International Affective Picture System”. In: *Behavior research methods* 37.4 (2005), pp. 626–630.

[103] Agnes Moors et al. “Appraisal theories of emotion: State of the art and future development”. In: *Emotion Review* 5.2 (2013), pp. 119–124.

[104] Donn Morrison, Ruili Wang, and Liyanage C De Silva. “Ensemble methods for spoken emotion recognition in call-centres”. In: *Speech Communication* 49.2 (2007), pp. 98–112.

[105] Emily Mower, Maja J Mataric, and Shrikanth Narayanan. “A Framework for Automatic Human Emotion Classification Using Emotion Profiles”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.5 (2011), pp. 1057–1070.

[106] Emily Mower, Maja J Matarić, and Shrikanth S Narayanan. “Evaluating evaluators: A case study in understanding the benefits and pitfalls of multi-evaluator modeling”. In: *Tenth Annual Conference of the International Speech Communication Association.* 2009.

[107] Emily Mower et al. “Interpreting ambiguous emotional expressions”. In: *International Conference on Affective Computing and Intelligent Interaction.* 2009, pp. 1–8.

[108] Shrikanth Narayanan and Panayiotis G Georgiou. “Behavioral signal processing: Deriving human behavioral informatics from speech and language”. In: *Proceedings of the IEEE* 101.5 (2013), pp. 1203–1233.

[109] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. “Estimating divergence functionals and the likelihood ratio by convex risk minimization”. In: *IEEE Transactions on Information Theory* 56.11 (2010), pp. 5847–5861.

[110] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. “On surrogate loss functions and f-divergences”. In: *The Annals of Statistics* (2009), pp. 876–904.

[111] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. “A multi-layer hybrid framework for dimensional emotion classification”. In: *ACM International Conference on Multimedia.* 2011, pp. 933–936.

[112] Mihalis A Nicolaou, Hatice Gunes, and Maja Pantic. "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space". In: *IEEE Transactions on Affective Computing* 2.2 (2011), pp. 92–105.

[113] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. "f-gan: Training generative neural samplers using variational divergence minimization". In: *Advances in Neural Information Processing Systems*. 2016, pp. 271–279.

[114] Keith Oatley and Philip N Johnson-Laird. "Towards a cognitive theory of emotions". In: *Cognition and Emotion* 1.1 (1987), pp. 29–50.

[115] Antonio Origlia, Francesco Cutugno, and Vincenzo Galatà. "Continuous emotion recognition with phonetic syllables". In: *Speech Communication* 57 (2014), pp. 155–169.

[116] Andrew Ortony, Gerald L Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge University Press, 1990.

[117] Maureen O'Sullivan. "Measuring the ability to recognize facial expressions of emotion". In: *Emotion in the human face* 2 (1982), pp. 281–317.

[118] Jaak Panksepp. "Toward a general psychobiological theory of emotions". In: *Behavioral and Brain Sciences* 5.03 (1982), pp. 407–422.

[119] Sona Patel et al. "Mapping emotions into acoustic space: The role of voice production". In: *Biological psychology* 87.1 (2011), pp. 93–98.

[120] Valery Petrushin. "Emotion in speech: Recognition and application to call centers". In: *Proceedings of Artificial Neural Networks in Engineering*. Vol. 710. 1999.

[121] Valery A Petrushin. "Emotion recognition in speech signal: Experimental study, development, and application". In: *studies* 3.4 (2000).

[122] Robert Plutchik. "A general psychoevolutionary theory of emotion". In: *Theories of Emotion* 1 (1980), pp. 3–31.

[123] Florian B Pokorny et al. "Detection of negative emotions in speech signals using bags-of-audio-words". In: *IEEE International Conference on Affective Computing and Intelligent Interaction*. 2015, pp. 879–884.

[124] Daniel Povey et al. "The Kaldi Speech Recognition Toolkit". In: *IEEE Workshop on Automatic Speech Recognition and Understanding*. 2011.

[125] Fabien Ringeval and Mohamed Chetouani. "A vowel based approach for acted emotion recognition". In: *Annual Conference of the International Speech Communication Association*. 2008.

[126] Fabien Ringeval et al. “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions”. In: *International Conference and Workshops on Automatic Face and Gesture Recognition*. 2013, pp. 1–8.

[127] Fabien Ringeval et al. “Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data”. In: *Pattern Recognition Letters* 66 (2015), pp. 22–30.

[128] R Tyrrell Rockafellar et al. “Extension of Fenchel’duality theorem for convex functions”. In: *Duke mathematical journal* 33.1 (1966), pp. 81–89.

[129] Bernardino Romera-Paredes et al. “Multilinear multitask learning”. In: *International Conference on Machine Learning*. 2013, pp. 1444–1452.

[130] Bernardino Romera-Paredes et al. “Transfer learning to account for idiosyncrasy in face and body expressions”. In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. 2013.

[131] Andrew Rosenberg. “Classifying skewed data: Importance weighting to optimize average recall”. In: *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[132] Robert Rosenthal. *Sensitivity to nonverbal communication: The PONS test*. Johns Hopkins University Press, 1979.

[133] Naomi G Rotter and George S Rotter. “Sex differences in the encoding and decoding of negative facial emotions”. In: *Journal of Nonverbal Behavior* 12.2 (1988), pp. 139–148.

[134] JA Russel. “A circumplex model of affect”. In: *Journal of Personality and Social Psychology* 39 (1980), pp. 1161–78.

[135] James A Russell. “A circumplex model of affect”. In: *Journal of Personality and Social Psychology* 39.6 (1980), pp. 1161–1178.

[136] James A Russell. “Core affect and the psychological construction of emotion”. In: *Psychological Review* 110.1 (2003), p. 145.

[137] James A Russell. “Culture and the categorization of emotion”. In: *Psychological Bulletin* 110 (1991), pp. 426–450.

[138] James A Russell. “Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies”. In: *Psychological Bulletin* 115.1 (1994), p. 102.

[139] James A Russell and Lisa Feldman Barrett. “Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant”. In: *Journal of Personality and Social Psychology* 76.5 (1999), p. 805.

[140] Saurav Sahay et al. “Multimodal Relational Tensor Network for Sentiment and Emotion Classification”. In: *arXiv preprint arXiv:1806.02923* (2018).

[141] Klaus R Scherer. “On the nature and function of emotion: A component process approach”. In: *Approaches to Emotion* 2293 (1984), p. 317.

[142] Klaus R Scherer. “Vocal communication of emotion: A review of research paradigms”. In: *Speech Communication* 40.1 (2003), pp. 227–256.

[143] KR Scherer. “Emotion in action, interaction, music, and speech”. In: *Language, Music, and the Brain: A Mysterious Relationship* (2013), pp. 107–139.

[144] Harold Schlosberg. “Three dimensions of emotion”. In: *Psychological Review* 61.2 (1954), p. 81.

[145] Erik M Schmidt and Youngmoo E Kim. “Modeling Musical Emotion Dynamics with Conditional Random Fields”. In: *International Society for Music Information Retrieval Conference.* 2011, pp. 777–782.

[146] Erik M Schmidt and Youngmoo E Kim. “Prediction of Time-varying Musical Mood Distributions from Audio”. In: *International Society for Music Information Retrieval Conference.* 2010, pp. 465–470.

[147] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “Facenet: A unified embedding for face recognition and clustering”. In: *Computer Vision and Pattern Recognition.* 2015, pp. 815–823.

[148] Björn Schuller, Stefan Steidl, and Anton Batliner. “The INTERSPEECH 2009 emotion challenge.” In: *INTERSPEECH.* 2009, pp. 312–315.

[149] Bjorn Schuller et al. “Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition”. In: *International Conference on Multimedia and Expo.* 2008, pp. 1333–1336.

[150] Björn Schuller et al. “Acoustic emotion recognition: A benchmark comparison of performances”. In: *ASRU Workshop.* 2009, pp. 552–557.

[151] Björn Schuller et al. “Avec 2011–the first international audio/visual emotion challenge”. In: *IEEE International Conference on Affective Computing and Intelligent Interaction.* 2011, pp. 415–424.

[152] Björn Schuller et al. “Avec 2012: the continuous audio/visual emotion challenge”. In: *ACM International Conference on Multimodal Interaction.* 2012, pp. 449–456.

[153] Björn Schuller et al. “Cross-corpus acoustic emotion recognition: Variances and strategies”. In: *IEEE Transactions on Affective Computing* 1.2 (2010), pp. 119–131.

[154] Björn Schuller et al. "Selecting training data for cross-corpus speech emotion recognition: Prototypicality vs. generalization". In: *Afeka-AVIOS Speech Processing Conference*. 2011.

[155] Björn Schuller et al. "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism". In: *INTERSPEECH*. 2013.

[156] Björn Schuller et al. "Using multiple databases for training in emotion recognition: To unite or to vote?" In: *INTERSPEECH*. 2011, pp. 1553–1556.

[157] Nicu Sebe et al. "Emotion Recognition Based on Joint Visual and Audio Cues". In: *ICPR*. Vol. 1. 2006, pp. 1136–1139.

[158] Mohit Shah, Chaitali Chakrabarti, and Andreas Spanias. "Within and cross-corpus speech emotion recognition using latent topic model-based features". In: *EURASIP Journal on Audio, Speech, and Music Processing* 2015.1 (2015), pp. 1–17.

[159] Ismail Mohd Adnan Shahin. "Gender-dependent emotion recognition based on HMMs and SPHMMs". In: *International Journal of Speech Technology* 16.2 (2013), pp. 133–141.

[160] Mohammad Shami and Werner Verhelst. "Automatic classification of emotions in speech using multi-corpora approaches". In: *SPS-DARTS*. 2006, pp. 3–6.

[161] Timothy J Shields et al. "Action-Affect Classification and Morphing using Multi-Task Representation Learning". In: *arXiv preprint arXiv:1603.06554* (2016).

[162] Ingo Siegert, Ronald Böck, and Andreas Wendemuth. "Inter-rater reliability for emotion annotation in human–computer interaction: Comparison and methodological improvements". In: *Journal on Multimodal User Interfaces* 8.1 (2014), pp. 17–28.

[163] Craig A Smith and Phoebe C Ellsworth. "Patterns of cognitive appraisal in emotion". In: *Journal of Personality and Social Psychology* 48.4 (1985), p. 813.

[164] P Smolensky. "Information processing in dynamical systems: Foundations of harmony theory". In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1. 1986, pp. 194–281.

[165] Kihyuk Sohn. "Improved deep metric learning with multi-class n-pair loss objective". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1857–1865.

[166] Marina Sokolova and Guy Lapalme. "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45.4 (2009), pp. 427–437.

[167] Hyun Oh Song et al. "Deep metric learning via lifted structured feature embedding". In: *Computer Vision and Pattern Recognition*. 2016, pp. 4004–4012.

[168] Peng Song et al. "Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization". In: *Speech Communication* 83 (2016), pp. 34–41.

[169] Peng Song et al. "Speech emotion recognition using transfer learning". In: *IEICE Transactions on Information and Systems* 97.9 (2014), pp. 2530–2532.

[170] Nitish Srivastava et al. "Dropout: A simple way to prevent neural networks from overfitting". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 1929–1958.

[171] Alexander Staller and Paolo Petta. "Towards a tractable appraisal-based architecture for situated cognizers". In: *C. Numaoka, D. Canamero, & P. Petta, Grounding Emotions in Adaptive Systems. SAB* 98 (1998).

[172] Stefan Steidl et al. ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency". In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. 2005, pp. I–317.

[173] Yaniv Taigman et al. "Deepface: Closing the gap to human-level performance in face verification". In: *Computer Vision and Pattern Recognition*. 2014, pp. 1701–1708.

[174] Masayuki Tanaka and Masatoshi Okutomi. "A novel inference of a restricted boltzmann machine". In: *ICPR*. 2014, pp. 1526–1531.

[175] Christian Thiel. "Classification on soft labels is robust against label noise". In: *Knowledge-Based Intelligent Information and Engineering Systems*. 2008, pp. 65–73.

[176] Silvan S Tomkins. "Affect theory". In: *Approaches to Emotion* 163 (1984), p. 195.

[177] George Trigeorgis et al. "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2016, pp. 5200–5204.

[178] Khiet P Truong, Mark A Neerincx, and David A Van Leeuwen. "Assessing agreement of observer-and self-annotations in spontaneous multimodal emotion data". In: *INTERSPEECH*. 2008, pp. 318–321.

[179] Khiet P Truong, David A Van Leeuwen, and Franciska MG De Jong. “Speech-based recognition of self-reported and observed emotion in a dimensional space”. In: *Speech Communication* 54.9 (2012), pp. 1049–1063.

[180] KP Truong et al. “Arousal and Valence prediction in spontaneous emotional speech: Felt versus perceived emotion”. In: *INTERSPEECH*. 2009.

[181] Michel F Valstar et al. “Fera 2015-second facial expression recognition and analysis challenge”. In: *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*. Vol. 6. 2015, pp. 1–8.

[182] Konstantinos Veropoulos, Colin Campbell, Nello Cristianini, et al. “Controlling the sensitivity of support vector machines”. In: *IJCAI*. 1999, pp. 55–60.

[183] Dimitrios Ververidis and Constantine Kotropoulos. “Automatic speech classification to five emotional states based on gender information”. In: *EUSIPCO*. 2004, pp. 341–344.

[184] Bogdan Vlasenko, Björn Schuller, and Andreas Wendemuth. “Tendencies regarding the effect of emotional intensity in inter corpus phoneme-level speech emotion modelling”. In: *International Workshop on Machine Learning for Signal Processing*. 2016, pp. 1–6.

[185] Bogdan Vlasenko and Andreas Wendemuth. “Annotators’ agreement and spontaneous emotion classification performance”. In: *Sixteenth Annual Conference of the International Speech Communication Association*. 2015.

[186] Bogdan Vlasenko and Andreas Wendemuth. “Processing affected speech within human machine interaction”. In: *Annual Conference of the International Speech Communication Association*. 2009.

[187] Bogdan Vlasenko et al. “Balancing spoken content adaptation and unit length in the recognition of emotion and interest”. In: *Annual Conference of the International Speech Communication Association*. 2008.

[188] Bogdan Vlasenko et al. “Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications”. In: *Computer Speech & Language* 28.2 (2014), pp. 483–500.

[189] Bogdan Vlasenko et al. “Vowels formants analysis allows straightforward detection of high arousal emotions”. In: *International Conference on Multimedia and Expo*. 2011, pp. 1–6.

[190] Thurid Vogt and Elisabeth André. “Improving automatic emotion recognition from speech via gender differentiation”. In: *Language Resources and Evaluation Conference*. 2006.

[191] Ju-Chiang Wang et al. "Exploring the relationship between categorical and dimensional emotion semantics of music". In: *ACM workshop on Music information retrieval with user-centered and multimodal strategies*. 2012, pp. 63–68.

[192] Ju-Chiang Wang et al. "Modeling the affective content of music with a Gaussian mixture model". In: *IEEE Transactions on Affective Computing* 6.1 (2015), pp. 56–68.

[193] Ju-Chiang Wang et al. "The acoustic emotion Gaussians model for emotion-based music annotation and retrieval". In: *ACM International Conference on Multimedia*. 2012, pp. 89–98.

[194] Qi Wang, Jia Wan, and Yuan Yuan. "Deep metric learning for crowdedness regression". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2017).

[195] John B Watson. "A schematic outline of the emotions". In: *Psychological Review* 26.3 (1919), p. 165.

[196] John Broadus Watson. "Behaviorism". In: (1930).

[197] Kilian Q Weinberger and Lawrence K Saul. "Distance metric learning for large margin nearest neighbor classification". In: *Journal of Machine Learning Research* 10.Feb (2009), pp. 207–244.

[198] Martin Wöllmer et al. "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies." In: *INTERSPEECH*. Vol. 2008. 2008, pp. 597–600.

[199] Wilhelm Max Wundt. *Outlines of psychology*. W. Engelmann, 1897.

[200] Wen-Jing Yan et al. "How fast are the leaked facial expressions: The duration of micro-expressions". In: *Journal of Nonverbal Behavior* 37.4 (2013), pp. 217–230.

[201] Guosheng Yang, Yingzi Lin, and Prabir Bhattacharya. "A driver fatigue recognition model based on information fusion and dynamic Bayesian network". In: *Information Sciences* 180.10 (2010), pp. 1942–1954.

[202] Yi-Hsuan Yang and Homer H Chen. "Prediction of the distribution of perceived music emotions using discrete samples". In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.7 (2011), pp. 2184–2196.

[203] Jufeng Yang, Ming Sun, and Xiaoxiao Sun. "Learning Visual Sentiment Distributions via Augmented Conditional Probability Neural Network". In: *AAAI Conference on Artificial Intelligence*. 2017.

[204] Jufeng Yang et al. "Retrieving and classifying affective images via deep metric learning". In: *AAAI Conference on Artificial Intelligence*. 2018.

[205] Anbang Yao et al. "Capturing au-aware facial features and their latent relations for emotion recognition in the wild". In: *ACM International Conference on Multimodal Interaction*. 2015, pp. 451–458.

[206] Promod Yenigalla et al. "Speech Emotion Recognition Using Spectrogram & Phoneme Embedding". In: *INTERSPEECH* (2018), pp. 3688–3692.

[207] Dong Yi et al. "Deep metric learning for person re-identification". In: *International Conference on Pattern Recognition*. 2014, pp. 34–39.

[208] Masaki Yuki, William W Maddux, and Takahiko Masuda. "Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States". In: *Journal of Experimental Social Psychology* 43.2 (2007), pp. 303–311.

[209] Matthew D Zeiler et al. "On rectified linear units for speech processing". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 3517–3521.

[210] Zhihong Zeng et al. "A survey of affect recognition methods: Audio, visual, and spontaneous expressions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.1 (2009), pp. 39–58.

[211] Biqiao Zhang, Georg Essl, and Emily Mower Provost. "Automatic recognition of self-reported and perceived emotion: Does joint modeling help?" In: *ACM International Conference on Multimodal Interaction*. 2016, pp. 217–224. DOI: 10.1145/2993148.2993173.

[212] Biqiao Zhang, Georg Essl, and Emily Mower Provost. "Predicting the distribution of emotion perception: Capturing inter-rater variability". In: *ACM International Conference on Multimodal Interaction*. 2017, pp. 51–59. DOI: 10.1145/3136755.3136792.

[213] Biqiao Zhang, Soheil Khorram, and Emily Mower Provost. "Exploiting Acoustic and Lexical Properties of Phonemes to Recognize Valence from Speech". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. In submission. 2019.

[214] Biqiao Zhang and Emily Mower Provost. "Automatic recognition of self-reported and perceived emotions". In: *Multimodal Behavior Analysis in the Wild: Advances and Challenges*. Ed. by Xavier Alameda-Pineda, Elisa Ricci, and Nicu Sebe. To appear. Elsevier, 2018.

[215] Biqiao Zhang, Emily Mower Provost, and Georg Essl. "Cross-corpus acoustic emotion recognition from singing and speaking: A multi-task learning ap-

proach". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. 2016, pp. 5805–5809.

[216] Biqiao Zhang, Emily Mower Provost, and Georg Essl. "Cross-corpus acoustic emotion recognition with multi-task learning: Seeking common ground while preserving differences". In: *IEEE Transactions on Affective Computing* (2017). Early Access. DOI: 10.1109/TAFFC.2017.2684799.

[217] Biqiao Zhang et al. "f-Similarity Preservation Loss for soft labels: A demonstration on cross-corpus speech emotion recognition". In: *AAAI Conference on Artificial Intelligence*. To appear. 2019.

[218] Biqiao Zhang et al. "Predicting emotion perception across domains: A study of singing and speaking". In: *AAAI Conference on Artificial Intelligence*. 2015, pp. 1328–1335.

[219] Zixing Zhang et al. "Unsupervised learning in cross-corpus acoustic emotion recognition". In: *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. 2011, pp. 523–528.