Automatic Emotion Recognition: Quantifying Dynamics and Structure in Human Behavior

by

Yelin Kim

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering-Systems)
in The University of Michigan
2016

Doctoral Committee:

Assistant Professor Emily Mower Provost, Chair
Associate Professor Jason Corso
Professor Alfred O Hero III
Associate Professor Honglak Lee
Associate Professor Siwei Lyu, University at Albany, State University of New York

To my parents and brother

ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my advisor, Emily Mower Provost, for her continuous support and guidance throughout my PhD journey. This dissertation would not have been possible without her support and I can never thank her enough for her exemplary mentorship. It has been the huge privilege to be her first PhD student, and I am immensely grateful for having the best advisor and mentor that a PhD student can possibly imagine. She taught me the true joy and beauty of research, also pushing me to keep the high standards of my work with insightful and constructive comments. I could not be more grateful for all of her positive impact on my life, both personally and professionally.

I would also like to thank the researchers and mentors who significantly influenced my research and career. Siwei Lyu has always provided insightful advice and support whenever I make important career decisions. I also deeply thank Roddie Cowie, one of the most inspiring psychologists in emotion research, for his support on my work and his generous help on this journey. Also, Alfred Hero has helped me since my first year at Michigan, and has continued to provide his thoughtful insights on my work. Lastly, I would like to thank Sehyun Oh, who has closely mentored me since my freshman year and has shared her wisdom and advice throughout my journey.

I also greatly appreciate my thesis committee members, Jason Corso and Honglak Lee. Their insightful comments have been invaluable, and have challenged me to continue improving the quality of this dissertation. I would also like to thank many great professors at Michigan, including Demosthenis Teneketzis, Sandeep Pradhan, Laura Balzano, and Clayton Scott. Also, I deeply appreciate huge support on my

work from the top-tier GE researchers, including Jixu Chen, Peter Tu, Ming-Ching Chang, Guiju Song, and Jiwon Mok. Special thanks to Becky Turanski, Kimberly Mann, Michele Feldkamp, José-Antonio Rubio, Anne Rhoades, Don Winsor, and Laura Fink for their invaluable support in administration and computing resources.

My five years in Ann Arbor have been truly a joyful experience, and the beginning of priceless friendships with many inspiring people. I am thankful especially to Chun Lo, Takanori Watanabe, Chansoo Lee, Jong-Jin Park, Kihyuk Sohn, Grace Tsai, Zhen Zeng, Donghwan Kim, Daeyon Jung, Kyemin Lee, and Hyun-Jung Cho, for their helpful feedback on many of my papers and presentations. I would also like to thank my beloved CHAI labmates–Duc Le, Biqiao Zhang, John Gideon, June Shangguan, Zakaria Aldeneh, and Soheil Khorram. I will also always have fond memories and sincere gratitude for the countless special moments that I shared with Chaerin Jin, Nellie Kim, Wonhyung Lee, Shao-Yuan Chen, Victor Chan, and Mike Allison; with my friends in EE: Systems, including Taehyung Kim, Mooyoung Shin, Taewon Kim, Kibum Bae, Jaekyu Hyun, Pavan Datta, Rinarchi Garg, Paridhi Desai, Mai Le, Parinaz Naghizadeh, Eugene Wu, Connie Qiu, Joyce Liu, and Jean Young Song; and my potluck crews, Ha Nguyen and Hyesun Jun. I also thank my roommates, Korean Student Association-Graduate members, IBM Sapphire folks, GE friends, 89ers, my rock-climbing friends, Korean-Chinese eat-out crews, and the incredible people at BBB 3945, with whom I shared many great memories and exciting adventures.

Above all, this dissertation is dedicated to my wonderful parents, Taehoon Kim and Kyung-Mi Lee, for being my life-long mentors. I also thank my brother, Minseok Kim, for his endless love and support throughout my life. They have always been there for me throughout the ups and downs of my journey. Their love, optimism, and thoughtfulness have influenced every aspect of who I am today.

Lastly, I greatly thank for the financial support from KETEP, NSF, and IBM.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

ABSTRACT

Automatic Emotion Recognition: Quantifying Dynamics and Structure in Human Behavior

by

Yelin Kim

Chair: Emily Mower Provost

Emotion is a central part of human interaction, one that has a huge influence on its overall tone and outcome. Today's human-centered interactive technology can greatly benefit from automatic emotion recognition, as the extracted affective information can be used to measure, transmit, and respond to user needs. However, developing such systems is challenging due to the complexity of emotional expressions and their dynamics in terms of the inherent multimodality between audio and visual expressions, as well as the mixed factors of modulation that arise when a person speaks. Given these complex variations in affective behaviors, how can we capture emotion expressed over time? How can we effectively fuse the pieces of information from audio and visual expressions? How can we tease apart non-emotional behaviors that are mixed throughout the course of emotion expressions, such as mouth movement due to speech articulation or an eyebrow raise for emphasis?

To overcome these challenges, this thesis presents data-driven approaches that can quantify the underlying dynamics in audio-visual affective behavior as follows:

- **Motivational Studies:** The first set of studies lay the foundation and central motivation of this thesis. We discover that it is crucial to model complex non-linear interactions between audio and visual emotion expressions, and that dynamic emotion patterns can be used in emotion recognition.

- **Mixed Factors of Behavior:** The understanding of the complex characteristics of emotion from the first set of studies leads us to examine multiple sources of modulation in audio-visual affective behavior. Specifically, we focus on how speech modulates facial displays of emotion. We develop a framework that uses speech signals which alter the temporal dynamics of individual facial regions to temporally segment and classify facial displays of emotion.

- **Localization of Salient Events:** We present methods to discover regions of emotionally salient events in a given audio-visual data. We demonstrate that different modalities, such as the upper face, lower face, and speech, express emotion with different timings and time scales, varying for each emotion type. We further extend this idea into another aspect of human behavior: human action events in videos. We show how transition patterns between events can be used for automatically segmenting and classifying action events.

Our experimental results on audio-visual datasets show that the proposed systems not only improve performance, but also provide descriptions of how affective behaviors change over time. We conclude this dissertation with the future directions that will innovate three main research topics: machine adaptation for personalized technology, human-human interaction assistant systems, and human-centered multimedia content analysis.

CHAPTER 1

Introduction

Human-human and human-machine interactions often evoke and involve affective and social cues, such as emotion, social attitude, engagement, conflict, and persuasion. These signals, such as words, head and body movements, and facial and vocal expressions, can be inferred from both verbal and nonverbal human behaviors [107]. The signals profoundly influence the overall outcome of interactions [107, 175], and hence the understanding of these signals will enable us to build human-centered interactive technology tailored to an individual user's needs, preferences, and capabilities.

Emotion is an essential component of human interaction. It affects and regulates how we communicate with each other, and how we perceive, judge, and react to the outside world [71, 114, 172]. Therefore, if a machine can automatically recognize a user's emotion, it can enable natural and human-centered user experience and help automatic behavior assessment systems, namely, social and affective human-machine interaction systems, wellness and health-related systems that help individuals better monitor their emotional landscape, as well as intelligent surveillance systems that can automatically detect anomalous behaviors based on machine recognition of nervousness or anxiety.

Emotion expressions are complex and dynamic, and are often difficult to decode computationally. The expressions are inherently multimodal [172]– they involve behavior [9, 51, 91, 144, 195], physiology [23, 78, 83, 111, 183, 238], and language [174, 186, 207, 243]. These expressions also continue to change over time [21, 32, 165, 220], moreover, they are often mixed with other factors of modulation [113, 114, 172]. For instance, when a person is speaking, facial movements change

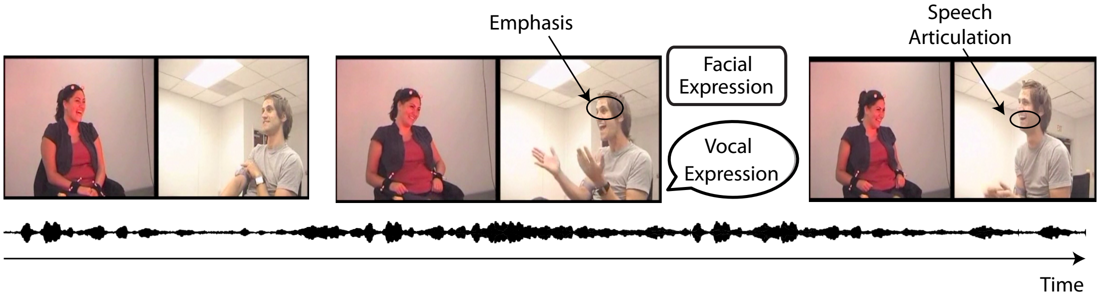


Figure 1.1: *Example of two people interacting with each other. This example addresses three main challenges in automatic emotion recognition that we tackle in this thesis: (i) complex multimodal interactions between facial and vocal expressions, (ii) continuous changes in emotion and behavior, and (iii) multiple factors of behavior modulation, including emotion (e.g., smiling), emphasis (e.g., eyebrow raise), and speech articulation (e.g., mouth movement changes). The figure is generated from one of the datasets used in this dissertation, called IEMOCAP [27].*

to communicate not only emotion, but also other types of signals, such as emphasis or lexical content (Figure 1.1). Facial changes may be related to emotion, such as when a person raises his/her eyebrows in surprise, however, they may result from other sources of modulation, e.g., an eyebrow raise due to speech emphasis. Similarly, changes in the mouth region due to smiling are similar to the changes when a person is saying ‘cheese’ [113, 114, 142, 148].

The goal of this dissertation is to develop emotion recognition systems that capture complex interactions between audio and visual expressions during speech. To this aim, we design frameworks that can control for variations in emotion expression by capturing cross-modal interactions, modeling temporal emotion and behavior, and controlling for non-emotional behavior.

This thesis is organized into three main parts: (i) motivational studies, (ii) mixed factors of behavior, and (iii) localization of salient events (Table 1.1). We first present two studies that motivate the importance of capturing complex non-linear interactions between audio and visual emotion expressions (Chapter 3), and describe how

	Part	Research Problem	Chapter	Reference
I	**Motivational Studies**	Cross-modal interaction	3	[117]
		Emotion dynamics	4	[112]
II	**Mixed Factors in Behavior**	Temporal segmentation	5	[113, 114]
		Informed segmentation and labeling	6	[116]
III	**Localization of Salient Events**	Emotion spotting	7	[115]
		Event detection	8	[118]

Table 1.1: *Outline of this Dissertation.*

emotion flows over time for automatically recognizing emotion (Chapter 4). The findings led us to explore how these audio and visual modalities interact over time, which constitutes the central topic of this dissertation. We study emotion changes in the upper face, lower face, and speech modalities over time, and discover that temporal methods capable of controlling for non-emotional behavior improve the system performance (Chapters 5 and 6). We then study methods to detect regions of salient affective behavior, varying for different modalities and emotion types (Chapter 7). We further test the importance and application of modeling temporal dynamics to recognize another aspect of human behavior: human action events from videos. We find that the modeling of transition patterns between behaviors of interest improves the performance of the systems (Chapter 8).

1.1 Emotion Background

1.1.1 Emotion: Definitions, Assumptions, and Quantifications

What is an emotion? William James posed this question in his revolutionary book *"What is an Emotion?"* in 1884 [99], defining emotion as "a distinct bodily expression" in which "a wave of bodily disturbance of some kind accompanies the perception of the interesting sights or sounds, or the passage of the exciting train of ideas. Surprise, curiosity, rapture, fear, anger, lust, greed, and the like, become then the names of the mental states with which the person is possessed" [172]. Since then, the definition of an emotion is still an ongoing discussion and many researchers have attempted to answer this question [2, 29, 58, 63, 67, 127, 170, 172, 193, 201–203].

A recent study by Scherer presents the component process model, which considers emotions as "the synchronization of many different cognitive and physiological components [201]". In his model, he distinguishes emotion from other low-level cognitive appraisals (e.g., processing of relevance) in that emotions require the overall process rather than just a trigger of bodily expressions and feelings. In [29], Cabanac defines emotion as any human experience with "high intensity and high hedonic content (pleasure/displeasure)".

A book by Ortony et al. [172] offers a comprehensive summary of the cognitive structures of emotion and studies related to emotion in the cognitive process. The authors present two main views on emotion: one is the *appraisal theory* that describes emotions as appraisals and arousals of events [63, 67, 172, 193, 202]. The fundamental assumption behind the appraisal theory is that emotion happens based on individuals' interpretations or appraisals on events rather than physiological arousals. The other view on emotion claims that any emotion can be described using continuous values of different dimensions. The most two widely used dimensions are activation (excited vs. calm) and valence (positive vs. negative) [2, 127, 203]. Scholsberg found that three dimensions of emotion, valence (pleasantness), activation, and attention, can be described based on facial and bodily expressions [203]. More recent studies in 1995 by Lang et al. explored the effect of valence and arousal of television messages on viewers' memory of the messages [127]. Oatley et al. distinguishes emotion from mood or preference by the duration of each kind of state [170]. They present different time courses for affective phenomena, such as expressions, emotions, moods, emotional disorders, and personal traits (Figure 1.2). In this description of emotion, expressions and autonomic changes span from seconds to minutes, whereas self-reported emotions

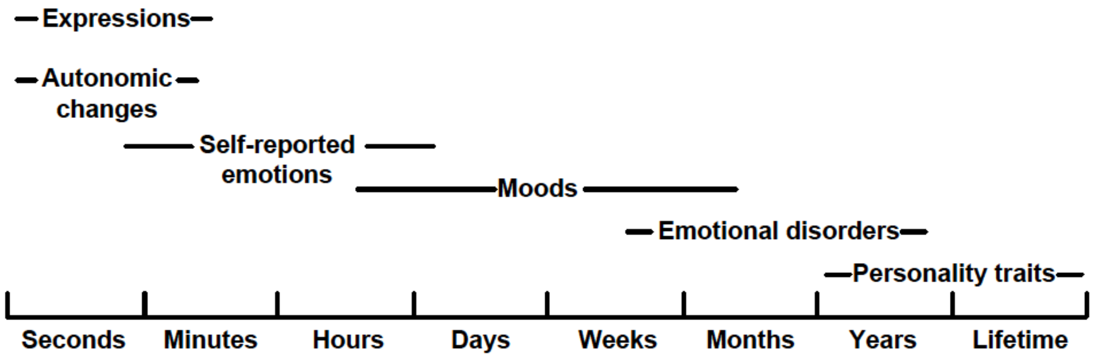


Figure 1.2: *A spectrum of affective phenomena in terms of the time course of each. Figure from [170, 222].*

last from minutes to hours. Moods last from a few hours to months, whereas emotional disorders can span up to years and personality traits range between years to lifetime.

In this dissertation, we focus on engineering approaches to understand and recognize expressive behaviors of emotion. We assume that the emotion labels in a given dataset are related to the emotional phenomena in the dataset. Based on this assumption, we can quantitatively measure and represent emotion, i.e. *affective labeling*. Affective labeling provides our ground truth data. There are two widely used approaches to affective labeling: categorical and dimensional approaches [74, 79].

Categorial Modeling of Emotion Categorical modeling of emotion represents emotion in terms of a finite number of discrete emotion labels such as *Angry, Happy, Sad*, etc. The fundamental assumption behind categorical modeling is that a small, finite number of 'basic' emotions is hard-wired in the human brain and is recognized universally [62]. Paul Ekman, who studied relationships between facial expressions and emotions, defined a basic emotion that is differentiated from other emotions by the following properties [57]:

- Distinctive universal signals
- Distinctive physiology

- Automatic appraisal
- Distinctive universals in antecedent events
- Distinctive appearance developmentally
- Presence in other primates
- Quick onset
- Brief duration
- Unbidden occurrence
- Distinctive thoughts, memories, images
- Distinctive subjective experience

Many researchers support the ‘palette theory’, which assumes that any emotion can be represented as a mixture of *primary emotions*, Anger, Disgust, Fear, Joy, Sadness, and Surprise [47].

Dimensional Modeling of Emotion In contrast to categorical modeling, dimensional modeling of emotion, introduced in 1954, defines emotion in terms of continuous values corresponding to different dimensions. The two most common dimensions are: *Valence* (positive vs. negative) and *Activation* (calm vs. excited) [203]. Previous studies suggested that these two dimensions can capture most emotion expressions [2]. The valence-activation dimensional space or a variation of this have been used in previous studies, such as [36, 39, 77, 192, 245].

In this thesis, we adopt a categorical approach to describe emotion since it is more straightforward to compare with previous systems with more identifiable descriptions.

Perceived Emotion In this dissertation, we use perceived emotion from multiple human annotators (greater than or equal to three annotators) to label the emotion of a given stimulus. This enables us to develop systems that can capture the expressive behavior of human emotion. Such emotional capability is critical for the real-world applications of systems, since expressive behaviors largely influence the overall tone and quality of human interactions [100, 106, 232]. For instance, a personal assistant system can analyze emotion expressions of a user, infer the underlying emotion based on how humans would perceive these expressions, and respond correspondingly, resulting in more natural interactions with the user.

1.1.2 Emotion Expressions

Emotion is expressed in multiple modalities. The primary modalities studied in this dissertation are speech and the facial expressions of emotion. Mehrabian has found the relative importance of verbal and nonverbal messages, called the "the 7%-38%-55%" rule [145]. He found three basic elements in any face-to-face communication: tone of voice, nonverbal behavior, and words. His findings suggest that these three elements account differently for our perception of the other person's emotions: words account for 7%, tone of voice accounts for 38%, and visual cues like facial and body language accounts for 55%. For the perception of emotion, these non-verbal elements, the audio and video cues, are particularly important. Specifically, when the messages are inconsistent, i.e., if words disagree with nonverbal behavior (e.g., "I do not have a problem with you!" uttered with an angry face and voice), people tend to believe the tonality and nonverbal behavior.

Physiological signals are often used to recognize self-reported emotion [184, 189] or to analyze mental disorders [183]. Studies on emotion expressed through natural language are often called 'sentiment analysis', where the attitude or emotion behind the text is inferred [174, 186, 207].

Humans integrate emotional information from different modalities to infer emotion [9, 51, 91, 144]. De Gelder and Vroomen demonstrated that humans cannot completely ignore emotion expressions from specific modalities, for instance, their perception is affected by vocal expressions even when they are asked to ignore speech and focus only on facial displays [51]. The finding demonstrates that bidirectional links exist between emotion perception systems in vision and audition. A recent survey discusses how and why the timings of emotion expressions in different modalities are not synchronized in empirical studies [91]. The survey concludes that the complex nature of behavioral, phychophysiological, and cognitive components of emotion, including the "intensity of emotions elicited in the laboratory, nonlinearity, between- versus within-subject associations, the relative timing of components," would result in a theory-data mismatch.

1.1.3 Emotion Perception

The recognition of emotion from behavioral expressions vary in time for different emotion types [179, 231]. Pell and Kotz studied how quickly listeners can identify a speaker's emotion from the speech patterns, and whether different types of emotion require different durations for the recognition [179]. They divided utterances conveying each emotion into different segments based on the number of syllables starting from the beginning of the utterance. They analyzed how much information an observer requires to recognize each emotion as a speech utterance unfolds. This study found that humans recognize anger, sadness, fear, and neutral expressions more accurately in a shorter time frame than they do happiness and disgust; however, the findings demonstrate that the recognition of happiness improves significantly as speech progresses [179].

Tracy and Robins explored a Darwinian aspect of emotion, which generally assumes that humans evolved to "communicate needs that facilitate survival and reproduction" [231]. They designed experiments to study the different speed that perceivers require to recognize different types of emotion, such as anger and happiness, and the effect of cognitive load on emotion recognition. They found that even under cognitive load, perceivers can recognize most emotion expressions quickly and efficiently. Also, consistent with previous work [54, 95, 120], this study found that humans are quicker to recognize positive emotion than negative emotion.

Individuals perceive emotion expressions differently, which results in inter-rater disagreement in emotion perception [221]. Gabrielsson et al. studied confusion in music emotion perception for 'perceived' and 'induced' emotion, and found that the relationship between these emotions varies depending on multiple factors, such as music and human evaluators [68]. Perception can also change due to disorders [3], e.g., in patients with a mental disorder such as schizophrenia [109, 123], or brain disorders [8, 85, 93], or autism spectrum disorder [173]. Different lighting conditions during emotion displays can also affect humans' perception of emotion [6].

1.2 Problem Statement and Methods

1.2.1 Multimodal Interactions in Emotion Expressions

Human emotion expressions are inherently multimodal. A key challenge is to understand how computational systems can combine multiple modalities. We explore complex non-linear interactions between speech and facial emotion expressions (Chapter 3). Previous emotion recognition systems have seen great improvements in classification accuracy, due in part to advances in feature selection methods. However, many of these methods capture only linear relationships between features, or, alternatively, require the use of labeled data. Our hypothesis is that deep learning techniques, which can explicitly capture complex non-linear feature interactions in multimodal data, can improve the system performance compared to traditional feature selection methods. We test this hypothesis by evaluating a suite of Deep Belief Network models, and demonstrate that these models show improvement in emotion classification performance over traditional methods that do not employ deep learning. This result suggests that the learned high-order non-linear relationships are effective for emotion recognition.

1.2.2 Dynamic Patterns of Emotion

Human emotion changes continuously and sequentially, resulting in dynamics intrinsic to affective communication. One of the challenges in developing automatic emotion recognition is to computationally represent and analyze these dynamic patterns. We explore how global, utterance-level dynamic patterns can be automatically captured by emotion classification systems (Chapter 4). Our hypothesis is that these dynamic patterns are specific to different emotion classes, and the systems that use these patterns can outperform baseline methods that use static patterns or low-level feature dynamics. We quantitatively represent emotion flow within an utterance by estimating short-time affective characteristics, using techniques introduced in [155, 158]. We compare the time-series estimates of these characteristics using Dynamic Time Warping, a time-series similarity measure. We demonstrate that this measure can be used to classify the affective label of the utterance. We test our hypothesis and show that similarity-based pattern modeling outperforms both a feature-based method and static modeling, particularly for ambiguous emotion content (defined as no rater consensus). Our results also provide insight into the typical high-level patterns of emotion. We visualize these dynamic patterns and the similarities between the patterns to gain insight into the nature of emotion expression.

1.2.3 Reducing Speech-Related Variability for Facial Emotion Recognition

We consider the problem of speech-related modulation for facial emotion recognition (Chapter 5). Real-world emotion recognition faces a central challenge when a user is speaking. In particular, facial movements arising from speech are often confused with facial movements related to emotion. In this chapter, we first focus on facial movements modulated by emotion and speech articulation. Facial emotion recognition systems aim to discriminate between emotions, while reducing the speech-related variability in facial cues. This aim is often achieved using two key features: (1) *phoneme segmentation*: facial cues are temporally divided into units with a single phoneme, and (2) *phoneme-specific classification*: systems learn patterns associated with groups of visually similar phonemes, e.g. /P/, /B/, and /M/. In this work, we hypothesize that proper segmentation that can effectively capture emotion-specific variation is critical, and empirically compare the effects of different temporal segmentation and classification schemes for facial emotion recognition. We propose an unsupervised segmentation method that does not necessitate costly phonetic transcripts. We show that the proposed method bridges the accuracy gap between a traditional sliding window method and phoneme segmentation, achieving a statistically significant performance gain. We also demonstrate that the segments derived from the proposed unsupervised and phoneme segmentation strategies are similar to each other. This paper provides new insight into unsupervised facial motion segmentation and the impact of speech variability on emotion classification.

1.2.4 Temporal Framework for Controlling Sources of Modulation in Audio-Visual Affective Behavior

This chapter extends Chapter 5 and considers the problem of multiple sources of modulation in audio-visual affective behavior (Chapter 6). As shown in Chapter 5, recent studies have found that the use of phonetic information can reduce speech-related variability in the lower face region. However, it has been less explored how to distinguish between the upper face movements due to emotion and speech. This gap leads us to the proposal of the Informed Segmentation and Labeling Approach (ISLA). ISLA uses speech signals that alter the dynamics of the lower and upper face regions. We demonstrate how pitch can be used for estimating emotion from the upper face, and how this estimate can be combined with emotion estimates from the lower face and speech. Our emotion classification results on the IEMOCAP and SAVEE datasets show that ISLA improves overall classification performance. We also demonstrate how emotion estimates from different modalities correlate with each other, providing insights on the difference between posed and spontaneous expressions.

1.2.5 Emotion Spotting: Discovering Regions of Salient Audio-Visual Affective Behavior

This chapter aims to discover consistent patterns of emotion in time across the lower face, upper face, and speech modalities (Chapter 7). Previous studies have found that humans require different amounts of temporal information to accurately perceive emotion expressions. This varies as a function of emotion classes. For example, Pell and Kotz found that the recognition of happiness requires longer speech data than the recognition of anger [180]. In this chapter, we hypothesize that a data-driven system can leverage these patterns of emotion and achieve similar performance to traditional systems with less data. To test this hypothesis, we develop a system that automatically detects emotion evidence for different emotion classes and different modalities. We use a combination of four binary emotion classifiers to estimate short-time emotion, and explore patterns (timings and durations) of emotion evidence. Our results demonstrate similar patterns for each emotion class across different training folds of emotion corpora. In addition, we show that our proposed method that only uses a portion of the data (e.g., 59%) can achieve comparable accuracy to a system that uses all of the data within each utterance. Our data-driven method has a higher accuracy compared to a baseline method that randomly chooses a portion of the data. We show that the performance gain of the method is mostly from prototypical emotion expressions (defined as expressions with rater consensus). The key novelty of the proposed method is that it provides understanding of how multimodal cues reveal emotion over time.

1.2.6 Transition Patterns in Temporal Behavior

Previous chapters demonstrated the importance of modeling temporal dynamics of emotion expressions. The question arises, are these temporal dynamics important to other aspects of human behavior? Chapter 8 explores temporal approaches for recognizing human action events in videos. Our hypothesis is that there exist transition patterns between these behaviors and a system that models this transition patterns will improve the system performance on behavior recognition. To this end, we propose a temporal segmentation and classification framework that accounts for transition patterns between events of interest. We apply this method to automatically detect salient human action events from videos. A discriminative classifier (e.g., Support Vector Machine) is used to recognize human action events, and an efficient dynamic programming algorithm is used to jointly determine the starting and ending temporal segments of recognized human actions. The key difference from previous work is that we introduce the modeling of two kinds of event transition information, namely *event transition segments*, which capture the occurrence patterns between two consecutive events of interest, and *event transition probabilities*, which model the transition probability between the two events. Experimental results show that the proposed approach significantly improves the segmentation and recognition performance for the two datasets we tested, in which distinctive transition patterns between events exist.

1.2.7 Emotion Evaluation Strategies

To train and test our systems, we employ leave-one-speaker-out cross-validation. In each dataset, we hold out each speaker as a "test speaker" and deploy the remaining speakers to train the system. We evaluate the performance of the system on each held-out test speaker. To be consistent with previous multi-class emotion recognition research [130, 158], we use unweighted average recall (UAR) over all speakers in a given dataset to measure performance. We employ paired t-tests to test the significance of the difference between systems, to be consistent with previous work in emotion recognition [142]. We particularly use a paired t-test method for k-fold cross validation in [52] to compare the accuracy of each test fold (subject). We claim significance when the p-value is less than 0.05.

1.3 Related Work and Contributions to this Topic

In this section, we provide a summary of related work to this dissertation. Comprehensive surveys of methods in automatic emotion recognition can be found in [30, 31, 62, 79, 96, 121, 163, 210].

1.3.1 Related Work in Multimodal Emotion Recognition

Emotion recognition systems use either unimodal (i.e., only using speech or facial features) and multimodal (i.e., using both speech and facial features) data. In this section, we review the unimodal and multimodal systems and present common feature selection techniques.

Unimodal Emotion Recognition. Speech is one of the most important methods of human communication [62]. The progress made in speech recognition has sparked new research directions into methods to extract and analyze the emotional content from speech [132, 152, 156, 209, 210]. The studies include investigations into speech features and feature selection methods [72, 249, 249], proper units of analysis [73, 124], and classification methods [70, 131].

Emotion has also been modeled using visual cues. The goal is to automatically extract and analyze salient visual information. A majority of visual emotion recognition systems is based on facial expressions, since facial displays arguably contain more discriminative features for emotion recognition than body gestures [175, 240]. Recent studies have focused on inferring emotion based on salient visual cues, including not only facial expression features, but also other types of visual features, such as aesthetic features (introduced by [16]), to understand perceived emotions [40, 102].

Multimodal Emotion Recognition. Researchers found that the joint use of speech and facial cues can improve the overall accuracy in emotion recognition. Many studies have investigated how to combine these two modalities and how to build a classification system that could effectively fuse this information [103, 146, 197, 200]. Decision-level fusion methods, which combine emotion estimates from different modalities at the decision level, have been widely used. Kächele et al. presented a hierarchical emotion and depression recognition system that trained ensembles of weak learning algorithms and fuses the audio and facial data using a Kalman filter at the decision level [103]. Savran et al. showed that particle filtering methods can also effectively combine emotion estimates from audio, facial, and lexical modalities, where the estimates are treated as measurement variables in the filtering methods [200]. A Bayesian network topology to combine facial and vocal expressions in a probabilistic manner was also proposed by Sebe et al. [214]. They found that the performance of the proposed multimodal system is higher than both facial and vocal emotion recognition systems.

Feature Selection in Emotion Recognition. Feature selection techniques used extensively in emotion research include: Forward Selection, Information Gain (IG), and Principal Component Analysis (PCA). These techniques are either supervised (forward selection and IG) or use representations based on the linear dependencies between the original features (PCA).

Forward feature selection is a greedy algorithm that sequentially selects features that increase the overall classification accuracy. This method has been widely used in many machine learning applications, including emotion recognition tasks [132]. Although this method can identify a subset of good features for classification, it may not be suitable if there are groups of features with complex relationships due to the greedy nature of the approach. IG-based feature selection methods are also commonly used in emotion recognition [158, 186]. This method ranks features by calculating the reduction in the entropy of class labels given knowledge of each feature. In general, however, it does not search for feature interactions. Furthermore, both forward selection and IG methods require labeled data during the feature selection process.

PCA and its variants (e.g., Principal Feature Analysis, or PFA [137]) are broadly used in the emotion recognition literature [149, 221, 246]. PCA finds a linear projection of the base feature set to a new feature space where the new features are uncorrelated. The feature set can be reduced to retain a majority of the variance in the original feature space. Although this unsupervised method has been widely used in many emotion applications, the limitation is in its linear projection of the base features, which tends to obscure the emotion content [24]. PFA is an extension of PCA. It clusters the data in the PCA space and returns the final features closest to the center of each cluster. This results in a feature set that maintains an approximation of the variance of the original set, while minimizing correlations between features. However, an open question remains whether complex non-linear interactions between audio and video modalities would benefit emotion recognition systems, as shown to be useful in audio-visual speech classification [166]. We explore this research question in Chapter 3.

1.3.2 Related Work in Emotion Dynamics

This section discusses related work that developed quantitative models of human emotion dynamics. A proper understanding of the dynamic nature of emotion has led to modeling advancements and a greater understanding of the temporal patterns that underlie our affective communication. There is a large body of work in tracking feature-level emotion structure, including Hidden Markov Models (HMMs) and Bidirectional Long Short-Term Memory (BLSTM) systems. For instance, coupled HMMs were used to take dynamics from vocal and facial expressions into account in emotion recognition [151]. Mower and Narayanan also have demonstrated that short-term estimates of affective flow could also be modeled dynamically using HMMs, which suggested that emotion has definable structure [155]. BLSTM models are neural networks with memory blocks that can capture variable amounts of context [206, 246]. These models are effective in capturing long-range context [76]. In these methods, and commonly within the community, the common practice for modeling emotion dynamics considers the feature-level fluctuations of the signal [182, 247]. Our work differs from previous studies in that we directly compute the time-series similarity between trajectories of emotion using Dynamic Time Warping (Chapter 4). This provides flexibility in the analysis of emotion stimuli and permits an analysis of the temporal patterns that are similar.

1.3.3 Related Work in Reducing Sources of Modulation

Previous studies in emotion recognition have attempted to reduce the effect of variability in audio-visual behavior. These studies mostly focused on facial emotion recognition systems to reduce the effect of speech-related variability. The studies have shown that methods using speech production knowledge can improve emotion recognition performance on specific regions of the face tied to speech production, such as the mouth [113, 114, 142, 148]. These methods segmented facial data at the phoneme level, and separately trained emotion classifiers for each visually similar phoneme group (e.g., /P/, /B/, /M/). However, questions still remain on how the upper face region changes over time and how these changes are associated with emotion. Previous work has shown that the upper face region changes in a longer-range duration than the lower face region, however it is under-explored whether these changes can be accurately captured from speech signals [24, 25].

Temporal Segmentation of Audio-Visual Data. Audio-visual data are often temporally segmented into smaller units to obtain more meaningful features [124, 205], to build dynamic classifiers [155, 198], and to find semantically meaningful regions [5, 17, 124, 194].

Temporal segmentation is commonly employed in speech emotion recognition. One approach is to use phonemes for segmentation (a comprehensive survey on phoneme segmentation can be found in [230]). In speech recognition, sub-word units such as phonemes are often used, since word-level or whole-word models are challenging to build due to the large vocabulary sizes in natural language [73]. Several recent works have approached phoneme segmentation problems as well [105, 110, 188]. In the pioneering study of Lee et al., the authors designed and compared the standard emotion-specific HMMs and HMMs trained on individual phoneme groups for each emotion and found that vowel sounds were the most effective for emotion classification compared with the other four phoneme groups [133]. Ringeval et al. also proposed a speech feature extraction method based on a pseudo-phonetic speech segmentation technique combined with a vowel detector [191]. They compared MFCC acoustic features from these pseudo-phonetic segments (vowels, consonants) with segments created by identifying the regions of voiced and unvoiced speech. They showed that the voiced segments could be modeled more accurately than the vowel or consonant segments for emotion recognition.

Signals other than phonemes have also been explored for segmenting emotional speech. Koolagudi et al. studied methods to segment speech for emotion recognition based on the prosody of speech segments. They used words and syllables as units of the segments [124]. They found that the system-level performance using prosody-based speech segments was not high, but that the performance significantly improved when combined with spectral features. Batliner et al. treated words as the basic unit of emotion expression. They combined words either into syntactically and semantically meaningful chunks or into sequences of words that belonged to the same emotion class [12]. Jeon et al. investigated different sub-sentence segment units (words, phrases, time-based segments) using a two-level system that focused both on segment-level and utterance-level emotion prediction. They found that time-based segments achieved the best performance [101]. Schuller et al. also investigated different timing patterns for segmentation using absolute and relative time intervals. Utterances were either segmented at fixed time intervals (absolute) or at fixed relative positions such as halves or thirds (relative) [205]. They demonstrated that absolute time intervals of one second achieved the highest accuracy (also demonstrated in [155]). Additionally, they found that systems based on relative time intervals were more accurate than those that used absolute time intervals.

There have also been research efforts in temporal segmentation for facial emotion expression recognition. As seen in audio modeling, these methods include phoneme-based segmentation [43, 90] and the standard fixed-length and multiple fixed-length segmentation [155, 168, 198]. Cohen et al. proposed a multi-level HMM for the automatic segmentation and classification of facial expressions [43]. The proposed method automatically segments and classifies a video recording of six sequences that display each of the six basic emotions (*anger, disgust, fear, happiness, sadness, and surprise*). The multi-level HMM makes an assumption that the transitions between emotions pass through the *neutral* state. They compare this with the standard emotion-specific HMM, where the input video is pre-segmented for each emotion. They found that the accuracy of their proposed system was similar to that of the standard emotion-specific HMM. Hoey used a manual segmentation of facial cues presenting a subject's underlying emotion [90]. He proposed a multi-level weakly supervised dynamic Bayesian network that learns the high-level dynamics of facial expressions.

There has also been work focused on unsupervised segmentation of facial expression data. Zhou et al. presented an unsupervised segmentation and clustering method of facial events [250]. They used k-means clustering with a Dynamic Time Alignment Kernel [216] to segment and cluster posed and unposed facial events. They found moderate intersystem agreement with the Facial Action Coding System (FACS). However, most of the previous work used facial data not modulated by spoken content, rendering it challenging to understand the impact of speech-related variability (a notable exception includes [250]). Our work is differentiated from the previous studies focusing on how we can better estimate emotion class by reducing the variability of facial movements caused by speech, while using unsupervised techniques.

Phoneme or Viseme-Dependent Modeling. Previous studies found that facial cues are difficult to model when facial movement is modulated by both emotion and speech production [113, 142, 148]. These studies have approached these challenges by building emotion classification systems that train classifiers on specific groups of phonemes with similar facial movement. This construction allows for a focus on modulations due to emotion rather than due to articulation and emotion. Metallinou et al. first conducted phoneme-dependent modeling on the IEMOCAP database for facial emotion recognition [148]. They presented an emotion classification system based on HMM that separated the classifiers into 14 similar viseme groups, the groups also used in our study. The highest unweighted accuracy they achieved was 55.74%, when using viseme-specific HMMs. Mariooryad and Busso studied two types of methods to reduce or compensate for speech variability in facial emotion recognition: feature-level and model-level compensation [142]. The feature-level method normalizes phoneme-dependent patterns in facial movement using the whitening transformation to compensate for the difference in phoneme-dependent patterns in the features. The model-level method separates emotion classifiers into viseme-dependent groups. The study found that both the feature and model-level compensation methods improve overall performance. In particular, their results showed a larger performance gain for the model-level method compared to the feature-level method. The previous studies demonstrated the benefits of phoneme segmentation and viseme-group classification, however, an open question remains as to how similar levels of accuracy can be achieved without segmenting based on phoneme transcript and whether phonemes are the correct unit for segmentation. We address this question in Chapter 5.

Interrelation Between the Lower Face, Upper Face, and Speech. Previous studies have found that there exist different characteristics in emotion expressions from the lower face, upper face, and speech modalities. Bassili studied the role of facial movement in human emotion recognition [10]. He found that humans rely relatively different on the upper and lower facial expressions to recognize emotion. For instance, human recognition of happiness is related to changes in the mouth and cheek regions, and humans confused happiness with sadness when only the upper face region was presented. Busso and Narayanan conducted a single-subject study that investigated the correlation between the recorded and estimated facial features, derived from speech features [24]. The speech features, which include pitch, energy, and MFCCs, are used to estimate facial features using an affine minimum mean square error (AMMSE) estimator. The results reveal that the lower face region provide the highest activeness and correlation levels. The highest correlation was $r = 0.8$. The findings demonstrated that facial gestures are linked at different resolutions, however, they modeled the upper face region simply at the sentence level.

Based on these findings, recent emotion recognition studies modeled different modalities separately [82, 140, 150]. Metallinou et al. studied the decision-level fusion of speech, facial and head movements [150]. They extracted features separately for the upper and lower face regions, with an intuition that the two regions have low correlation due to different underlying muscles and communicative roles. The authors found that this separate modeling of the two face regions improves the emotion classification of emotional states, i.e., anger, happiness, and sadness, compared to joint modeling. However, the proposed separate modeling decreased the accuracy for neutrality. Mansoorizadeh and Charkari studied feature-level fusion methods between the upper and lower face, and speech signals [140]. They proposed a buffer method to deal with asynchrony between the signals, where they either repeated the last frame value or took the median before fusion, similar to filling in missing values in the data. Hakim et al. also modeled the upper and lower face regions separately for facial emotion recognition, and demonstrated that such separate modeling achieves higher recognition rate than joint modeling of the whole face [82]. Questions still remain on how to (1) choose the right temporal scale (segmentation) for emotion expressions and (2) control for sources of modulation.

1.3.4 Modeling Transition Patterns in Temporal Behavior

Human action recognition is an active research area in computer vision [187, 233, 242].

Video segmentation. Segmentation of videos into salient events is an important task in video analysis that facilitates the retrieval, indexing, annotation, and representation of video data [125]. Traditionally, it entails shot boundary detection, i.e., the complete segmentation of a video into continuously imaged temporal segments [46].

Video event recognition. A recent research trend in temporal segmentation is based on salient events of interest rather than continuously recorded images, e.g., [89, 167, 225, 251]. Tang et al. studied HMM-based models to learn the temporal structure of complex events in Internet videos [225]. They utilized a variable-duration HMM to model the durations and transitions of an event segment of interest, where the model is trained in a discriminative, max-margin fashion. They achieved competitive accuracies on activity recognition and event detection tasks. However, their work differs from ours in that a video clip with a single event label is analyzed instead of a video sequence with multiple events. Hoai et al. [89], Cheng et al. [41], and Zhou et al. [251] studied the temporal segmentation of human action videos that contain multiple action events. Hoai et al. jointly localized and classified action events using a max-margin classifier and DP, which is most relevant to our work [89]. The main difference is that our approach benefits from the inclusion of transition events (i.e. events between two salient events of interest). The introduction of event transitions, the probabilistic modeling, and an efficient implementation are the key novelties of our work. Cheng et al. demonstrated the importance of temporal dependencies between events in joint segmentation and classification tasks [41] by applying the Sequence Memorizer [248]. The main difference of our work is that our system identifies events at the individual frame level, whereas the work of Cheng et al. represents a video using visual words of fixed-length sub-sequences. Zhou et al. studied unsupervised temporal clustering of human motion using the kernel k-means algorithm with the generalized dynamic time alignment kernel [251]. Our work differs from [251] in that we utilize the event-level transition information to capture longer-range temporal information of human motions.

Generative and discriminative event modeling. Transition events have been handled using generative models (e.g., transition matrix in HMM) [69] and modeled as individual transition events in specific domains, for example the onset and offset states in facial Action Unit recognition. Galata et al. used variable-length Markov models that temporally segmented human activities into atomic behavior components [69]. Valstar et al. presented a hybrid SVM/HMM system to segment a facial action into temporal phases (e.g. onset, offset, peak, and neutral states), with a noticeable performance gain [234, 235]. They used a sigmoid function operating on the SVM outputs as an emission probability for HMMs (instead of traditional Gaussian mixture models, since SVMs discriminate extremely well). Several studies have demonstrated the efficacy of using transition information for temporal segmentation of videos [225].

Event transition in facial movements. Studies in facial Action Units (AU) detection have demonstrated the utility of event transition information [53, 122, 234, 234]. AUs are anatomical facial muscle actions based on the Facial Action Coding System (FACS), where 9 upper face AUs and 18 lower face AUs are defined [235]. The set of AUs can be categorized by their transition states into *onset* (muscles contracting and expression becoming stronger), *peak* (with consistently strong expression), and *offset* (muscles relaxing back to neutral appearance) phases. The order of the phases is often "*neutral-onset-peak-offset-neutral*", whereas spontaneous facial expressions with multiple peaks and other orderings are also possible [44, 234]. Koelstra et al. introduced a combination of discriminative frame-based GentleBoost ensemble learners and used a dynamic generative HMM to detect AU and its temporal segments [122]. The 'cascade of tasks' of Ding et al. combines outputs of different tasks (frame, segment, and transition detection) linearly for the final AU event detection [53]. The combination parameters are learned by cross-validation, and independent onset and offset detectors were trained using a linear SVM for transition detection.

To our best knowledge, the use of transitions in discriminative learning has not been extensively exploited for event recognition, in particular for the purpose of joint localization and classification of complex video events.

1.4 Main Contributions

The main contributions of this dissertation are:

- We explain the importance of modeling complex non-linear interactions between audio and visual emotion expressions (Chapter 3). We show improvement in recognition rate when using deep learning approaches that capture these interactions compared to traditional feature selection methods.
- We provide interpretable descriptions of how emotion flows over time in Chapter 4.
- In Chapters 5 and 6, we explore how speech alters the dynamics of different regions of the face, and how the speech signals can be used to inform the design of audio-visual emotion recognition systems when a person is speaking.
- We discover subject-independent consistent patterns in time regions of emotion evidence in audio-visual affective behavior in Chapter 7.
- We demonstrate the importance and applicability of modeling temporal dynamics in human action events, and show that modeling transition patterns between behaviors can benefit behavior recognition systems in Chapter 8.

1.5 Organization of the Dissertation

This dissertation is composed of three main parts: (i) motivational studies (Chapters 3 and 4), (ii) mixed factors of behavior (Chapters 5 and 6), and (iii) localization of salient events (Chapters 7 and 8). We review datasets and features that we use in this dissertation in Chapter 2 and summarize the main contributions of this dissertation and discuss the potential directions for future work in Chapter 10.

CHAPTER 2

Data and Features

2.1 IEMOCAP Database

The majority of this dissertation uses the Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database (Figure 2.1 and 2.2 [27]). This database has been widely used in the field of automatic audio-visual emotion recognition [119, 130, 151, 158].

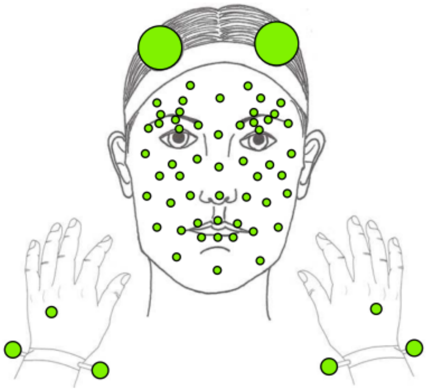
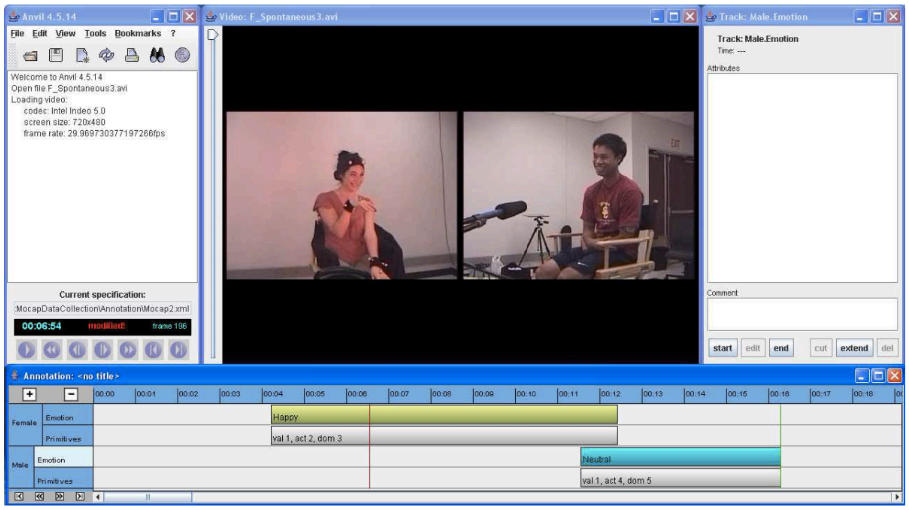


Figure 2.1: *IEMOCAP motion capture data [27]*

The database approximately contains 12 hours of dyadic conversations between five pairs of actors (each pair contains one male and one female). Each session contains both acted and improvised scenarios. For acted scenarios, actors were given three scripts which contain clear emotional content. For improvised scenarios, actors were asked to elicit specific emotions using high-level scene descriptions (e.g., you are at an airport and the airline has lost your baggage).

The data are captured using audio-visual cameras and a nine-camera Vicon recoding system, providing 3-D marker coordinates at 120 frames per second (fps). The



```
Welcome to Anvil 4.5.14
Open file F_Spontaneous3.avi
Loading video:
    codec: Intel Indeo 5.0
    screen size: 720x480
    frame rate: 29.969730377197266fps
```

Current specification: :MocapDataCollection\Annotation\Mocap2.xml

Figure 2.2: *Human annotation for emotion labels. The high-level affective content of each utterance (manually segmented) are evaluated based on categorical (e.g., happiness, sadness) and dimensional (e.g., activation, valence) labeling.* [27]

data include 53 motion capture markers on the actor's face as shown in Figure 2.1. The five nose markers are excluded due to their limited movement, and the two eyelid markers are also excluded due to their frequent occlusions, as in [113, 148]. The data are manually segmented into utterances (defined as speaker turns), resulting in over 10,000 utterances. The turn level is defined as the time a speaker is actively speaking.

The data were evaluated by human evaluators using both categorical and dimensional labeling schemes. The categorical labels are used in my approaches. The labels include: *Anger, Happiness, Neutrality, Sadness, Excitement, Surprise, Frustration, Fear, Disgust, Other*. They were evaluated by at least three evaluators. We use utterances with majority voted categorical labels from the set: *Anger, Happiness+Excitement, Neutrality, Sadness*, in line with previous studies [142, 148]. There are 43.0 ± 26.2 angry, 91.5 ± 37.5 happy, 44.6 ± 27.4 neutral, and 51.8 ± 28.1 sad utterances per speaker, totaling 3,060 utterances over all speakers. The mean length of an utterance is 4.73 ± 3.34 seconds. Utterances have an average of 0.75 seconds of silence at the beginning of an utterance and 0.86 seconds at the end of an utterance.

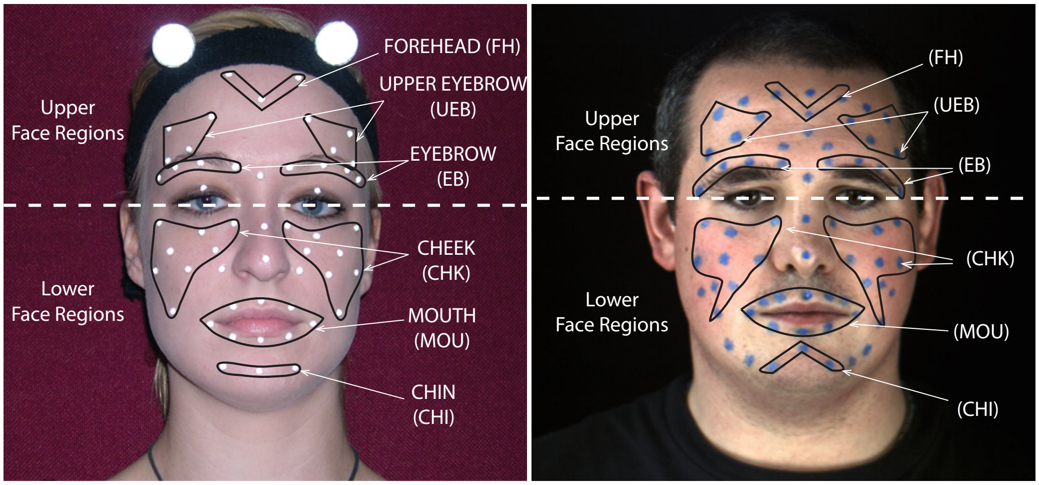


Figure 2.3: *Positions of face markers and six face regions: chin (CHI), forehead (FH), cheek (CHK), upper eyebrow (UEB), eyebrow (EB), and mouth. The images are from the IEMOCAP (left, [27]) and SAVEE datasets (right, [84]).*

Evaluators agreed with the assigned ground truth labels approximately 72% of the time. The dimensional attributes include valence (positive vs. negative), activation (calm vs. excited), and dominance (passive vs. dominant). They were evaluated by at least two evaluators.

2.2 SAVEE Database

The SAVEE dataset contains read speech of four male British-English speakers, eliciting six emotions: *Anger, Disgust, Fear, Happiness, Sadness,* and *surprise*. Each emotion was expressed in 15 phonetically-balanced sentences, and *Neutrality* in 30 sentences. This results in 480 utterances in total. In our work, we use four classes for consistency with the IEMOCAP database: *Anger, Happiness, Neutrality, and Sadness*, resulting in 300 utterances in total. The average length of an utterance within the subset is 3.85 ± 0.33 seconds. The utterances have an average of 0.51 seconds of silence at the beginning of an utterance and 0.55 seconds at the end of an utterance. The facial data include 2D coordinates of 60 markers on the forehead, eyebrows, cheeks, lips and jaw (Figure 2.3). The sampling rate was 44.1 kHz for audio, and 60 fps for video.

The provided emotion labels of the SAVEE dataset are the labels given to the actors, rather than the intended target emotion (annotated labels were not available). This is different from IEMOCAP, in which we use emotion labels derived from perceptual evaluations. However, the authors presented a high level of agreement between the intended target emotion and perceived emotion: 441 out of 480 total sentences in the data were perceived as the intended target emotion by at least 8 out of 10 annotators, indicating good agreement between the actor's intended emotion and the annotator's perception [84]. Additional differences between the SAVEE and IEMOCAP databases include: (i) 2-dimensional vs. 3-dimensional motion capture data, (ii) motion-capture frame rate of 60 fps vs. 120 fps, (iii) four speakers, each with scripted utterances, vs. ten speakers, each embedded within a dyadic interaction.

I use a subset of the 60 motion capture markers to have a configuration similar to the IEMOCAP database. The subset totals 46 markers (Figure 2.3). Further, we address the difference in fps between the two databases by interpolating the SAVEE motion capture data using cubic spline interpolation (described in Section 5.2.1) to increase the frame rate to 120 fps. We discuss the impact of this interpolation in Section 5.2.1. Finally, we scale the SAVEE motion capture data to have the same minimum and maximum values as in the IEMOCAP database.

2.3 Smartroom Database

We created a new Smartroom Dataset with volunteers performing a series of upper body actions, where the challenge is that both the temporal durations of events and the number of events are unknowns [118]. The dataset contains six subjects performing a mix of the following actions in 8 videos: Crossing arms on chest (CC), Touching face (TF), Arms on hip (AH), and Normal (N). Each action is repeated two to three times in each video. Normal action represents the case of hands down in a resting position. The average length of the videos is 47.8 seconds. Each of the {CC, TF, AH} actions was enacted sequentially following the "neutral-onset-peak-offset-neutral" pattern for the right arm, left arm, and both arms. The enacted events share a large extent of variations in terms of temporal durations and spatial locations.

For ground truth segment configurations, two human annotators labeled both (1) the start and end timing of peak segment, and (2) the action label of the three pre-defined actions. We add three frames prior-to and post-to each peak boundary, and define non-overlapping onset, peak, offset, and neutral segments. The onset and offset segments are always chosen to be 7 frames in length.

2.4 CMU-MAD Database

CMU-MAD dataset [94] contains 35 human actions of 20 subjects recorded using a Microsoft Kinect sensor. Similar to the Smartroom Dataset, we use the joint angles of elbows and shoulders as frame-level features, and utilize the same segment-level features φ mapping as in the Smartroom Dataset, i.e. mean, standard deviation, and linear regression slope. The start and end time of each action are provided in this dataset. However, the timings can not be directly used in our *neutral-onset-peak-offset-neutral* model, since the action between the start and end time contain all of the *neutral, onset, peak, offset,* and *neutral* events. Due to the specific labeling scheme of this dataset, it is reasonable to separate each labeled action segment into three sub-sequences: [0-33.3%] for onset, [33.3-66.6%] for peak, and [66.6-100%] for offset. we focus on the evaluation of 9 actions that contain meaningful transitions and exclude actions such as running (where the action peak as well as onset/offset transitions are not clearly defined). These selected 9 actions are: left/right arm waving, left/right arm pointing to the ceiling, crossing arms on the chest, basketball shooting, both arms pointing to both sides and left/right side.

Part I: Motivational Studies

CHAPTER 3

Multimodal Feature Learning for Emotion

3.1 Introduction

Emotion recognition is complicated by the inherent multimodality of human emotion expression (e.g., facial and vocal expression). This multimodality is characterized by complex high-dimensional and non-linear cross-modal interactions [228]. Previous research has demonstrated the benefit of using multimodal data in emotion recognition tasks and has identified various techniques for generating robust multimodal features [25, 177, 236, 237, 244]. However, although effective, these techniques do not take advantage of the complex non-linear relationship that exists between the modalities of interest, or alternatively require the use of labeled data. In this chapter, we apply deep learning techniques, which can overcome these limitations, in order to provide robust features for audio-visual emotion recognition.

Emotion recognition accuracy relies heavily on the ability to generate represen-

The work presented in this chapter has been published in the following article: **Yelin Kim**, Honglak Lee, and Emily Mower Provost. ”Deep Learning for Robust Feature Generation in Audio-Visual Emotion Recognition.” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

tative features. However, this is a very challenging problem. Emotion states do not have explicit temporal boundaries and emotion expression patterns often vary across individuals [4]. This problem is further complicated by the high dimensionality of the audio-visual feature space. Consequently, accurate modeling generally requires a reduction of the original input feature space. This reduction is commonly accomplished using feature selection, a method that identifies a subset of the initial features that provide enhanced classification accuracy [62]. However, it is not yet clear whether it is more advantageous to select a subset of emotionally relevant features or to capture the complex interactions across all features considered. In this chapter, we demonstrate the effectiveness of Deep Belief Networks (DBN) for multimodal emotion feature generation. We learn multi-layered DBNs that capture the non-linear dependencies of audio-visual features while reducing the dimensionality of the feature space.

There has been a substantial body of work on feature representation, extraction, and selection methods in the emotion recognition field in the last decade. Our work is motivated by the discovery of methods for learning multiple layers of adaptive features using DBNs [13]. Research has demonstrated that deep networks can effectively generate discriminative features that approximate the complex non-linear dependencies between features in the original set. These deep generative models have been applied to speech and language processing, as well as emotion recognition tasks [153, 154, 217]. In speech processing, Ngiam et al. [166] proposed and evaluated deep networks to learn audio-visual features from spoken letters. In emotion recognition, Brueckner et al. [22] found that the use of a Restricted Boltzmann Machine (RBM) prior to a two-layer neural network with fine-tuning could significantly improve classification accuracy in the Interspeech automatic likability classification challenge [212]. The work by Stuhlsatz et al. [223] took a different approach for learning acoustic features in speech emotion recognition using Generalized Discriminant Analysis (GerDA) based on Deep Neural Networks (DNNs). While the present study is related to recent approaches in multi-modal deep learning and the application of deep learning techniques to emotion data, it focuses on non-linear audio-visual feature learning for emotion, which has not been extensively explored in the emotion recognition domain.

In the current work we present a suite of DBN models to investigate audio-visual feature learning in the emotion domain. We compare two methodologies: (1) unsupervised feature learning (DBN) and (2) secondary supervised feature selection. We first build an unsupervised two-layer DBN, enforcing multi-modal learning as introduced by [166]. We augment this DBN with two types of feature selection (FS): 1) before DBN training to assess the benefit of feature learning exclusively from an emotionally-salient subset of the original features and 2) after DBN training to assess the advantage of reducing the learned feature space in a supervised context. We compare this to the performance of a three-layer DBN model. Our baseline is a Support Vector Machine that uses subsets of the original feature space selected using supervised and unsupervised feature selection. The results provide important insight into feature learning methods for multimodal emotion data.

The results show that the DBN models outperform the baseline models. Further, our results demonstrate that the three-layer DBN outperforms the two-layer DBN models for emotionally subtle data. This suggests that unsupervised feature learning can be used in lieu of supervised feature selection for this data type. In addition, the relative performance improvement of the three-layer model for subtle emotions suggests that these complex feature relationships are particularly important for identifying subtle emotional cues. This is an important finding given the challenges inherent in and need for recognizing emotions elicited in realistic scenarios [208].

3.2 Related Work

3.2.1 Unsupervised Feature Learning and Deep Learning

Deep learning techniques (See [13] for a survey) have become increasingly popular in various communities including speech and language processing [153, 154, 217] and vision processing [126, 134, 135, 219, 226]. This progress has been facilitated by the recent discovery of more effective learning algorithms for constructing DBNs in an unsupervised context, for example exploiting single-layer building blocks such as Restricted Boltzmann Machines (RBMs) [218]. DBNs [88] learn hierarchical representation from data and can be effectively constructed by greedily training and stacking multiple RBMs.

RBMs are undirected graphical models that represent the density of input data $\mathbf{v} \in \mathbb{R}^D$ (referred to as "visible units") using binary latent variables $\mathbf{h} \in \{0,1\}^K$ (referred to as "hidden units"). In the RBM, there are no connections between units in the same layer, which makes it easy to compute the conditional probabilities.

In this chapter, we use Gaussian RBMs that employ real-valued visible units for training the first layer of the DBNs. We use Bernoulli-Bernoulli RBMs that employ binary visible and hidden units for training the deeper layers. In a Gaussian RBM, the joint probability distribution and energy function of $\mathbf{v}$ and $\mathbf{h}$ is as follows:

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp(-E(\mathbf{v},\mathbf{h})) \tag{3.1}$$

$$E(\mathbf{v},\mathbf{h}) = \frac{1}{2\sigma^2}\sum_i v_i^2 - \frac{1}{\sigma^2}\left(\sum_i c_i v_i + \sum_j b_j h_j + \sum_{i,j} v_i W_{ij} h_j\right) \tag{3.2}$$

where $\mathbf{c} \in \mathbb{R}^D$ and $\mathbf{b} \in \mathbb{R}^K$ are the biases for visible and hidden units, respectively, $\mathbf{W} \in \mathbb{R}^{D\times K}$ are weights between visible units and hidden units, σ is a hyperparameter, and Z is a normalization constant. The conditional probability distributions of the Gaussian RBM are as follows:

$$P(h_j = 1|\mathbf{v}) = sigmoid\left(\frac{1}{\sigma^2}(\sum_i W_{ij}v_i + b_j)\right) \tag{3.3}$$

$$P(v_i|\mathbf{h}) = \mathcal{N}\left(v_i; \sum_j W_{ij}h_j + c_i, \sigma^2\right) \tag{3.4}$$

The posteriors of the hidden units given visible units (Equation 3.3) form the generated features used in the classification framework. The parameters of the RBM ($\mathbf{W}$, $\mathbf{b}$, $\mathbf{c}$) are learned using contrastive divergence as in [87]. We use sparsity regularization [134] to penalize a deviation of expected activation of the hidden units from a low fixed level p. Given a training set $\{\mathbf{v}^{(1)}, ..., \mathbf{v}^{(m)}\}$, we include a regularization penalty of the form:

$$\lambda \sum_{j=1}^{K} \left| p - \frac{1}{m}\sum_{l=1}^{m} \mathbb{E}\left[h_j^{(l)}|\mathbf{v}^{(l)}\right]\right|^2 \tag{3.5}$$

where $\mathbb{E}[\cdot]$ is the conditional expectation given the data, λ is a regularization parameter, and p is a constant that specifies the target activation of the hidden unit h_j [134].

3.3 Utterance-Level Features

We use audio and motion capture data of the IEMOCAP database. The audio features include both prosodic and spectral features: pitch, energy and mel-frequency filter banks (MFBs). Prosodic features such as pitch and energy features have been demonstrated to be highly effective in speech emotion recognition [62]. Previous research also found that spectral features capture a significant amount of emotional contents from speech; in addition, MFBs have been shown to be better discriminative features than mel-frequency cepstral coefficients (MFCCs) in emotion recognition [26]. The video features are based on Facial Animation Parameters (FAP), part of the MPEG-4 standard. FAPs describe the movement of the face using distances between particular points on the face. They have been widely used to capture facial expressions in the emotion recognition literature. The subset is chosen to include emotionally meaningful movements (e.g., eye squint, smile, etc.).

Our input features are statistics of the raw features calculated at the utterance level. They include: mean, variance, lower quantile, upper quantile, and quantile range. The initial feature set contained 685 features, where 145 are audio and 540 are video features. The features are normalized on a per-speaker basis to mitigate speaker variation [158].

3.4 Proposed Method

3.4.1 Cross-Validation and Performance Evaluation

We use leave-one-speaker-out cross validation to ensure that the models are not overtraining to the affective styles of a particular speaker. We pre-train the DBN models (unsupervised) and search for the best hyper-parameters including: sparsity parameters and the number of final output nodes. We select our hyper-parameters using cross validation over the training data. We fix the number of hidden nodes of the two-layer DBNs, the sigma parameter for the first-layer Gaussian RBMs, and the L2 regularization parameter (Section 3.4.3). We select the best hyper-parameters for each data type: prototypical, non-prototypical, and combined.

We evaluate the performance of the baseline and DBN systems using Unweighted Accuracy (UA). UA is an average of the recall for each emotion class [208]. The unweighted accuracy better reflects overall accuracy in the presence of class imbalance.

3.4.2 Baseline Models

Our baseline models are two SVMs with radial basis function (RBF) kernels. The SVMs do not use features generated via deep learning techniques. The SVMs have radial basis function (RBF) kernels and are implemented using the Matlab Bioinformatics Toolkit. We train four emotion-specific binary SVMs in a self-vs.-other approach. The final emotion class label is assigned by identifying the model in which the test point is maximally far from the hyperplane as in as in [158].

Both models employ feature selection. The first uses IG [55] and the second uses PFA [137] feature selection (a supervised and unsupervised feature selection technique, respectively). IG is applied to each emotion class, resulting in four sets of emotion-specific features. Each emotion-specific SVM uses the associated emotion-specific feature subset. The number of features is chosen over {60, 120, 180} for each data type.

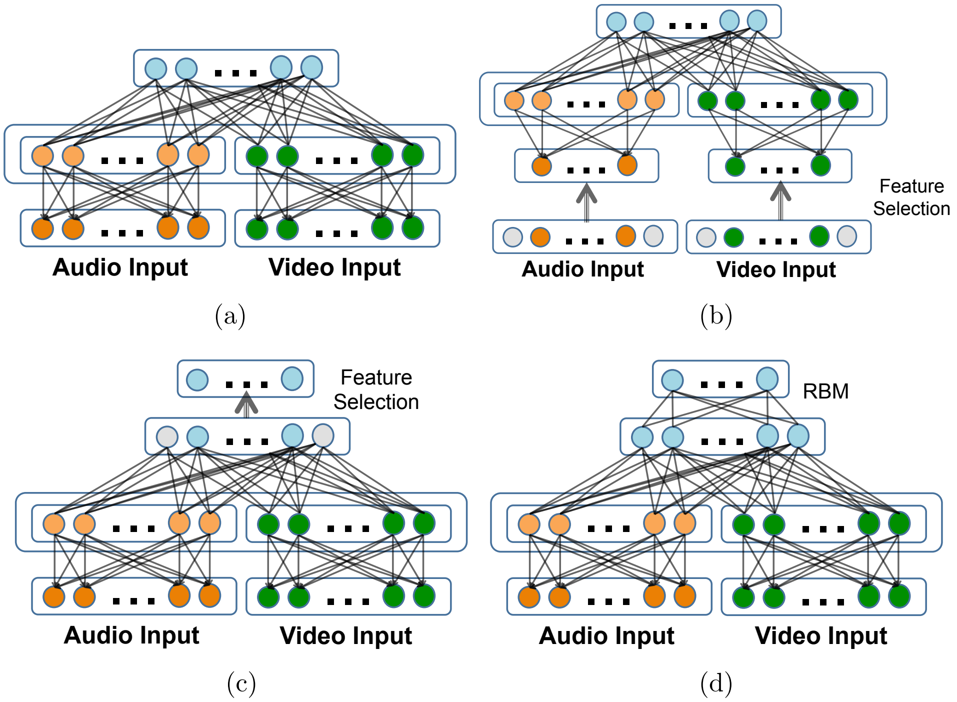
We optimize the baselines using leave-one-subject-out cross-validation for each data type (prototypical, non-prototypical, and combined data). The parameters include the number of selected features using IG and PFA, the value of the box constraint (C=1) for the soft margin in the SVM, and the scaling factor (sigma=8) in the RBF kernel.

We also compare our results with the maximal accuracy achieved from a previous work of Metallinou et al. [150], which utilizes the same IEMOCAP database as our work and introduces a decision-level Bayesian fusion over models using face, voice, and head movement cues. Although Metallinou's work used a different subset of the IEMOCAP database, this comparison supports the strong performance of our proposed method.

3.4.3 Deep Belief Network Models

We experiment with four different DBN models in order to explore different non-linear dependencies between audio and motion-capture features. We also assess the utility of feature selection methods in these deep architectures (Figure 3.1).

Our basic DBN is a two-layer model and is a building block for the other variants. It learns the audio and video features separately in the first hidden layer. The learned features from the first layer are concatenated and used as the input to the second



(a) (b) (c) (d)

Figure 3.1: Illustration of proposed models: (a) DBN2, (b) FS-DBN2, (c) DBN2-FS, and (d) DBN3.

hidden layer as introduced in [166]. We call this the *DBN2* model (Figure 3.1(a)).

The other three DBN models are based on DBN2. Two involve feature selection and one is a three-layer DBN model. The two-layer models use supervised feature selection (IG) either prior to or post the unsupervised pre-training. The three-layer model reduces the feature dimensionality using a third RBM layer, invoking unsupervised feature learning. Thus, the three-layer model captures additional high-order non-linear dependencies of all features, whereas the models employing supervised feature selection use only emotionally salient features. The variants are defined as follows:

- FS-DBN2 is a two-layer DBN with feature selection prior to the training of the DBN2 model (Figure 3.1(b)).
- DBN2-FS is a two-layer DBN with feature selection on the final RBM output nodes (Figure 3.1(c)).
- DBN3 is a three-layer DBN that stacks an additional RBM on the second-layer RBM output nodes of the DBN2 model (Figure 3.1(d)).

The number of hidden units in the first layer is approximately 1.5x overcomplete for each audio feature (300 units from 145 audio features) and video feature (700 units from 540 video features), resulting in 1000 concatenated first layer hidden units. The number of second hidden units is fixed at 200 for DBN2, DBN2-FS, and DBN3. For FS-DBN2, the number of second hidden units is fixed to 150 because the number of visible units is smaller compared to the other three DBN models.

The sparseness parameters are selected using leave-one-speaker-out cross-validation, while all other parameters (including hidden layer size and weight regularization) are kept fixed (See Section 3.4.1 for details). Since the number of video features is larger than the number of audio features, we select the sparsity parameters of bias for audio data and video data over {0.1, 0.2} and {0.02, 0.1}, respectively. Also, the sparsity parameters of λ are selected over {2, 6, 10} for audio features, while λ sparsity parameters are fixed at 5 for video features. Our preliminary results demonstrated that the λ value for the video features did not noticeably affect the results. The number of features selected at the final level (DBN2-FS) and the number of hidden units at the final level (DBN3) are selected over {50, 100, 150}.

For FS-DBN2, a total of 100 audio features and 200 video features are chosen using IG. We first pre-train a sparse RBM with 100-200 nodes for the audio features and 200-600 nodes for the video features. We select the sparsity parameters of bias over {0.1, 0.5} for each RBM. λ is fixed as 5. Next, we concatenate the learned features and pre-train a first layer of DBN with 800 output nodes and the second layer with 150 nodes (Bernoulli-Bernoulli).

The output of each DBN is classified using the same SVM structure used in the baseline (Section 3.4.2).

3.5 Results and Discussion

A summary of the emotion classification results can be seen in Table 3.1. The DBN models for the combined data achieve UAs ranging from 65.25% (DBN2) to 66.12% (DBN2-FS). All DBN models outperform the baseline models (the two baseline models perform comparably). The performance gap between the maximal UAs of proposed models and the PFA baseline is 1.67%.

The DBN models for the non-prototypical data achieve accuracies ranging from 56.70% (FS-DBN2) to 57.70% (DBN3). All DBNs outperform the baseline models (which again perform comparably). The performance gaps between the UAs of proposed models and baseline models range from 1.71% to 1.89%. We obtain a slight performance gain when using DBN3 compared to both DBN2-FS and FS-DBN2 for subtle or non-prototypical utterances (0.73% and 1.63% increase, respectively). This result is important given that the DBN3 model does not use any labeled data (unsu-

	Baseline		**Proposed DBNs**			
	IG	PFA	DBN2	DBN2-FS	DBN3	FS-DBN2
Combined	64.42	64.45	65.25	66.12	65.71	65.89
Non-Prot	55.81	55.99	56.89	56.97	57.70	56.07
Prot	73.38	70.02	70.46	72.96	73.78	72.77

Table 3.1: *Unweighted classification accuracy (%) for combined, non-prototypical, and prototypical data*

pervised feature learning), whereas the FS-DBN2 model learns a new set of features from a previously identified subset of emotionally salient features and the DBN2-FS invokes feature selection at the output. This demonstrates that we can effectively use unsupervised feature learning, rather than supervised feature selection, for emotion recognition, even for emotionally subtle utterances (non-prototypical).

The DBN models for the prototypical data achieve accuracies ranging from 70.46% (DBN2) to the maximum of 73.78% (DBN3). The performance gap between the maximal UAs of the proposed models and maximal UAs of the baseline models (73.38% with IG) is 0.40%. The baseline models themselves achieve differing levels of accuracy; the IG baseline outperforms the PFA baseline by 3.36%. This may suggest that in emotionally clear utterances, supervised feature selection (emotion-specific IG) is preferable to unsupervised feature selection (PFA). The accuracy of the DBN3 model indicates that unsupervised feature learning can achieve comparable performance to supervised feature selection for emotionally clear utterances. Further, the DBN3 outperforms unsupervised feature selection (PFA baseline) by 3.76%, highlighting the potential importance of feature learning rather than unsupervised feature reduction for emotionally clear data.

The deep learning method performs comparably to the previous work of Metallinou et al. [150], 62.42%. Direct comparisons are not possible due to differences in the data subsets considered.

We present the utility of deep learning techniques for unsupervised feature learning in audio-visual emotion recognition. Our results demonstrate that DBNs can be used to generate audio-visual features for emotion classification, even in an unsupervised context. The comparison of the classification performances between the baseline and the proposed DBN models demonstrate that it is important to retain complex non-linear feature relationships (using deep learning techniques) in emotion classification tasks. The strongest performance gain is observed in the non-prototypical data. This is important in applications of automatic emotion recognition systems where emotional subtlety is common.

CHAPTER 4

Analysis of Emotion Dynamics

4.1 Introduction

A proper understanding of the dynamic nature of emotion will lead to modeling advancements and a greater understanding of the structure that underlies our affective communication. There is a large body of research on modeling and assessing such dynamic structure. One of the most common methods is using HMMs. This technique gained popularity in the speech recognition community and has been effectively used in the emotion recognition community as well. However, in this chapter, we take a different approach and focus on methods that will provide interpretable descriptions of emotion dynamics. We quantify how emotion flows over an utterance

The work presented in this chapter has been published in the following article: **Yelin Kim** and Emily Mower Provost. "Emotion Classification via Utterance-Level Dynamics: A Pattern-Based Approach to Characterizing Affective Expressions." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013.

and demonstrate how patterns of this flow can effectively be used to predict an emotional state. Our goal is to identify these patterns, which we call "flow patterns," and use them in an emotion classification framework. We further hypothesize that the salient characteristics of these patterns are the long-term utterance-level dynamics in addition to the short-term fluctuations. We expect to see common patterns repeating over utterances of the same emotion class. We propose a simple quantitative method to model the flow patterns and demonstrate how these patterns of estimated emotion dynamics furthers our understanding of human emotion expression. We estimate emotion flow by extracting features related to emotion dynamics. The features are sequential short-term estimates of emotion states extracted using methods introduced in [155, 158]. Each estimate describes the utterance in terms of blends of emotional cues. Previous work demonstrated that these sequential emotion estimates can be used to classify and identify affective states in a dynamic classification setting [155]. In this chapter, we present an emotion modeling technique that leverages the intra-utterance flow patterns to capture the emotional similarity between utterances. This method natively provides insights into the flow patterns and their relationship to emotion state.

There is a large body of work in tracking feature-level emotion structure, including HMMs and BLSTM systems. The BLSTM models are neural networks with memory blocks that can capture variable amount of context [206, 246]. These models are effective in capturing long-range context. However, this context is firmly tied to the multiplicative gate units and may be difficult to interpret [76]. In these methods, and commonly within the community, the common practice for modeling emotion dynamics considers the feature-level fluctuations of the signal [182, 247]. Previous research demonstrated that short-term estimates of affective flow could also be modeled dynamically using HMMs. This suggested that emotion has definable structure [155]. However, our understanding of these underlying dynamics was restricted by the limitations of a finite state space [155]. In this chapter we provide a framework for dynamic modeling of emotion that provides interpretable descriptions of emotion expressions by explicitly focusing on utterance-level dynamics.

Our proposed method captures emotional similarity by estimating time-series similarity between flow patterns of different utterances. This allows us to explicitly take longer-range temporal dependencies into account because our method is focused on variation over the entire utterance (rather than frame level change). We first estimate short-term emotion content over small time windows for each utterance, which approximates emotion dynamics. We calculate the similarity between these estimated dynamics using Dynamic Time Warping (DTW). Unlike HMM, DTW does not make any statistical assumptions about the intrinsic model. Instead, it directly computes the flow pattern similarity between the unknown utterance and known time-series data [178]. We use this DTW similarity measure in an automatic emotion classification system (Figure 4.1).

The novelty of the proposed study is in its focus on interpretable utterance-level dynamic modeling, which furthers our understanding of the structure underlying emotional utterances. The results demonstrate that this modeling is effective for identifying emotion state. The maximal accuracy of flow pattern modeling of estimated emotion in DTW similarity-based classification system is 64.40% (unweighted accuracy). This accuracy is comparable or greater than that of a baseline model that captures the flow patterns at the feature level as well as a static model, 64.02% and 61.20%, respectively. Further it performs comparably to the state-of-the-art results on a different subset of the same database [160]. This suggests that flow pattern of temporal emotion dynamics offers interpretable descriptions on emotion fluctuation, while provides comparable results to the state-of-the-art model.

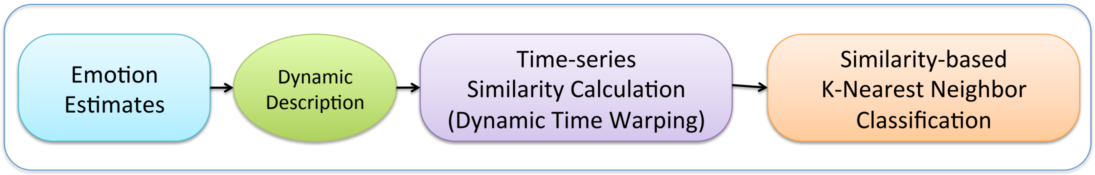

Emotion Estimates
Dynamic Description
Time-series Similarity Calculation (Dynamic Time Warping)
Similarity-based K-Nearest Neighbor Classification

Figure 4.1: *Illustration of the proposed method*

4.2 Audio-Visual Features

As in Chapter 3, we use both audio and motion-capture features of the IEMOCAP dataset. The audio features include pitch, intensity and MFBs. The motion-capture features are based on Facial Animation Parameters. In this chapter, we obtain final features by calculating statistics of the raw audio-visual features over small time windows. These include mean, variance, lower and upper quantiles, and quantile range, giving a total of 685 features. Of these 685 features, 145 are auditory features and 540 are video features, such as mean of pitch, lower quantile of a marker point of a mouth, etc. The features are normalized on a per-speaker basis to mitigate speaker variation [158].

4.3 Audio-Visual Feature Selection

This initial feature was reduced to 180 features as in [155] using Principal Feature Analysis (PFA) [137]. PFA is a variant of Principal Component Analysis (PCA). It projects the input data into the PCA space and clusters the data in this space using k-means. It returns the features closest to the center of each cluster. This ensures that the final features are features in the original space and that a target level of variance in the dataset is retained.

4.4 Emotion Estimation

4.4.1 Emotion Profile (EP)

The short-term affective estimates are made using the Emotion Profiles (EPs) framework. EPs were introduced and demonstrated to be effective for emotion recognition tasks in [155, 158, 160]. EPs describe the emotion content of an utterance by capturing the subtle blends of emotional cues present in that utterance. EPs estimate the degree of confidence in the presence or absence of each of these cues, forming an n-dimensional estimate of affective content. This study uses utterances in the label set: *Angry, Happy, Neutral, Sad.* Thus, the EP for an utterance is a four-dimensional vector describing the confidence, c, in the presence of each emotion from the set: $\overrightarrow{EP}$ $= [c_{angry}, c_{happy}, c_{neutral}, c_{sad}]$. We measure confidence using Support Vector Machines (SVMs). SVMs are binary maximum margin classifiers that find a separating hyperplane that maximizes the distance from the hyperplane to the points closest to the hyperplane. The outputs of SVM are class membership (±1) and distance from hyperplane. We multiply these quantities to arrive at an approximate measure of confidence. The SVMs are trained using leave-one-speaker-out cross-validation (Figure 4.2).

4.4.2 Emotogram

The emotogram of an utterance is the set of EPs extracted over windowed regions of an utterance (See Figure 4.2). They provide a dynamic description of the estimated presence or absence of each emotional cue [155, 160]. This can be seen as estimating the manner in which emotion cues flow in an utterance. Emotograms are four-dimensional time-series of estimated emotion dynamics: $\overrightarrow{Emotogram} = [\overrightarrow{EP_1}$, $\overrightarrow{EP_2}$, ..., $\overrightarrow{EP_T}]$, where T represents the number of sliding windows in an utterance. We use window lengths of 0.25, 0.5, 1.0, 1.5, and 2.0 seconds to evaluate the ef-

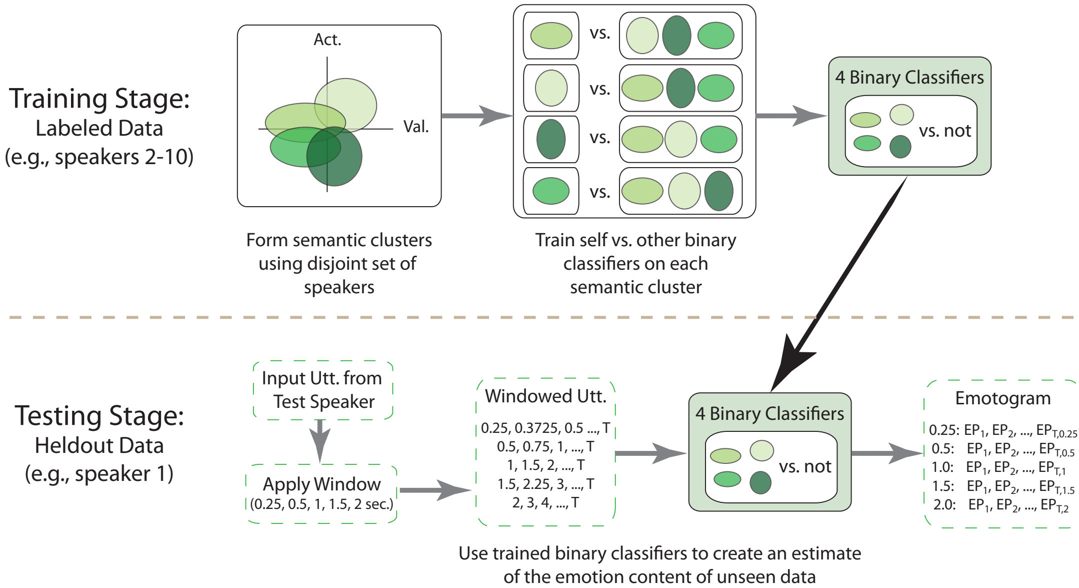

Training Stage:
Labeled Data
(e.g., speakers 2-10)
Act.
Val.
Form semantic clusters using disjoint set of speakers
vs.
vs.
vs.
vs.
Train self vs. other binary classifiers on each semantic cluster
4 Binary Classifiers
vs. not
Testing Stage:
Heldout Data
(e.g., speaker 1)
Input Utt. from Test Speaker
Apply Window
(0.25, 0.5, 1, 1.5, 2 sec.)
Windowed Utt.
0.25, 0.3725, 0.5 ..., T
0.5, 0.75, 1, ..., T
1, 1.5, 2, ..., T
1.5, 2.25, 3, ..., T
2, 3, 4, ..., T
4 Binary Classifiers
vs. not
Emotogram
0.25: EP_1, EP_2, ..., $EP_{T,0.25}$
0.5: EP_1, EP_2, ..., $EP_{T,0.5}$
1.0: EP_1, EP_2, ..., $EP_{T,1}$
1.5: EP_1, EP_2, ..., $EP_{T,1.5}$
2.0: EP_1, EP_2, ..., $EP_{T,2}$
Use trained binary classifiers to create an estimate of the emotion content of unseen data

Figure 4.2: *Emotion Profiles (EPs) and Emotograms generation proposed and described in the previous work [160]*

fect of window size on classification performance [155]. We investigated denoising techniques to mitigate subtle estimation noise. However, both Median Filtering and Kalman Smoothing techniques did not result in performance increases as compared to the *raw* emotograms. Therefore, the emotograms were not smoothed. We hypothesize that this may be because our DTW based method captures high-level emotion flow patterns, rather than the small estimation fluctuations, which would be sensitive to noise.

4.5 Proposed Method

4.5.1 Similarity Measurement Between Emotograms Using DTW

Our hypothesis is that the utterance-level patterns of emotion flow are informative with respect to emotion class. To test this hypothesis we measure the time-series similarity between each emotogram, our estimates of emotion flow, using DTW. DTW is a widely used technique that finds the best alignment between two time series by identifying the warping path between the two sequences that minimize the difference between the sequences. DTW has been widely used in many domains including speech recognition [162] and handwriting recognition [227]. DTW captures utterance-level dynamics, rather than probabilistic transitions in frame-by-frame characteristics, which are seen in HMM modeling. DTW provides flexibility in the analysis of utterances of different lengths since it aligns time series data. In emotion data, contrasted with speech phoneme modeling, the affective data are often of highly varied length. HMMs do not offer this same flexibility because of their innate restriction to an n-state model independent of utterance length. Further, it is difficult to interpret the resulting models generated by HMMs. We present an alternative dynamic modeling technique that facilitates visualizations of affective flow, providing clear measures of emotional similarity. We propose that DTW can be an alternative strategy for emotion recognition.

We align two utterances in the emotion space defined by the emotograms using Multi-Dimensional Dynamic Time Warping (MD-DTW), presented in [92]. MD-DTW uses all emotogram dimensions to identify the best alignment between two utterances in the emotion space. We define the emotion space as $\Phi^{I \times J}$ for a descriptor of length I and dimension J, where J is the number of emotogram dimensions ($J = 4$). Let $T \in \Phi^{M \times J}$and $L \in \Phi^{N \times J}$ be two emotograms in this space. MD-DTW computes the optimal alignment between T and L using dynamic programming ($O(MN)$) [196]. We find the optimal alignment by computing distance between the utterances. The distance measure between any two points in the series is defined as $d : \Phi \times \Phi \rightarrow \mathbb{R} \geq 0$, which can be any p-norm. We use 2-norm, the summation of the squared differences across all dimensions.

The MD-DTW algorithm populates the M by N distance matrix D according to the following equation:

$$D(i,j) = \sum_{k=1}^{K} (T(i,k) - L(j,k))^2, \tag{4.1}$$

where i and j represent the specific short-time estimate of the emotograms, T and L. The distance matrices can be visualized to understand the structural similarities across emotion class (Figure 4.4). We implemented four-dimensional DTW by modifying the one-dimensional code of [65].

4.5.2 *k*-Nearest Neighbor Classification Using MD-DTW

We use the k-Nearest Neighbor (k-NN) classifier to assign a final emotion class label based on the MD-DTW measure. k-NN assigns a label to a given test utterance based on the labels of its k nearest neighbors. The assigned label is a majority vote over the neighbors' labels. We select k using a 10-fold cross-validation hyper-parameter search over values 1, 3, 5, 7, 10, 30, and 50. We did this search over the combined data, which provided access to both the prototypical and non-prototypical examples and found $k = 50$.

We refer to the total framework as DTW-kNN. The algorithm is as follows. During training we calculate the DTW similarity between every pair of testing and training utterances. During testing we find the k closest neighbors to each test utterance using this DTW distance. We label the test utterance with the majority voted label of its k nearest training utterances. In both the DTW-kNN and baseline models we calculate accuracy using leave-one-speaker-out cross-validation. The final reported accuracy measures are the average of the accuracies over all 10 folds.

4.5.3 Baseline Models

We evaluate DTW-kNN by comparing it to three baseline models. The first baseline model tracks emotion similarity using trajectories composed of the compressed feature space ('feature trajectories'), rather than the estimates of affective flow. We reduce our original 180 features using Principal Component Analysis (PCA). The feature dimensionality is selected using leave-one-subject-out cross-validation over compressions that reduce the features space to 4, 10, 20, 30, and 40 dimensions. The best performing model uses ten PCA features. We compare the performances of this compressed feature space to that of the affective estimates to identify the method that best allows us to capture the structure underlying emotional speech. As in the emotion flow model, we calculate the DTW similarity over each utterance, as represented by the feature trajectories, and then identify the emotion state using k-NN with k=50 (selected using hyperparameter search).

The second baseline uses static EP modeling. Static EPs are calculated in the same manner as short-time EPs. However, here the emotion is detected using utterance-level statistics (as compared to windowed statistics, e.g., over 0.25 seconds). This baseline assesses whether the dynamics contribute to our understanding of emotion class. We classify the final label of the static EP estimate using k-NN over the four dimensions ($k = 50$). In the static baseline, the k-NN classifier uses the Euclidean distance between the four-dimensional EP values of the training and test utterances.

The final baseline is a published result that modeled the dynamics of the emotograms using HMMs. HMMs fit these dynamics to an n-state model, where here $n = 3$ (with left-to-right topology) [160]. This baseline is a comparison to a result on a subset of the utterances considered (2,903 utterances vs. 3,018 utterances).

4.6 Results

All results are reported using unweighted accuracy, the average of per class recall. This measure mitigates class imbalance in accuracy reporting. Overall, the DTW-kNN method achieves the highest performance gain for the non-prototypical utterances, the subtle utterances with only majority ground truths. The performance between our proposed method and the baseline methods for the prototypical and combined sets are comparable, however we discuss how our proposed method can provide insight into how emotion changes over time in Section 4.7. The maximal accuracy of our proposed method for the non-prototypical data is 55.79% with a window size of 2 seconds. This is 3.88% higher than the feature trajectory model with the same window size, and 3.15% for the maximally accurate feature trajectory (window size 1.5 seconds). It is 1.68% higher than the accuracy of the static EP. In the prototypical data experiment, the DTW-kNN method achieves the highest accuracy of 68.50% with a window size of 1.5 seconds. This outperforms the feature trajectory model by 1.13% on the same window size. The maximal accuracy for the feature trajectory model is 0.09% higher than our proposed model (window size 0.25 seconds). The combined data has a maximal accuracy of 64.40% with the DTW-kNN method (window size of 1 second). This is 0.48% higher than the feature trajectory with the same window size and 0.38% higher than the feature trajectory with its maximal accuracy with window size of 1.5 seconds. It is 3.20% higher than the static EP estimate. The results are summarized in Table 4.1. The HMM baseline was calculated only over a window size of 0.25 seconds with an accuracy of 64.67%. This is a similar result to our proposed DTW-kNN method, 63.95%, for a window size of 0.25 seconds and suggests this restricted n-state structure may not be necessary for dynamic modeling of emotion.

4.7 Discussion

Our results include two important findings. 1) The new dynamic modeling technique using flow pattern modeling can effectively capture the emotion dynamics. These dynamics can be used to effectively classify utterances. 2) In this framework, the secondary emotogram features outperform the compressed raw feature fluctuations, only for the case of non-prototypical data. This suggests that the secondary features capturing emotion flow offer a targeted compression of the emotion space. More-

	Model	Window size (seconds)				
		0.25	0.5	1	1.5	2
A	Emotogram	66.90	66.82	67.15	68.50	67.76
	Feature	68.59	66.46	66.51	67.38	67.49
	Static EP			67.34		
B	Emotogram	53.96	55.12	55.31	55.49	55.79
	Feature	49.30	50.12	51.64	52.64	51.91
	Static EP			54.11		
C	Emotogram	63.95	64.01	64.40	64.38	64.27
	Feature	62.72	63.91	63.91	64.02	63.44
	Static EP			61.20		

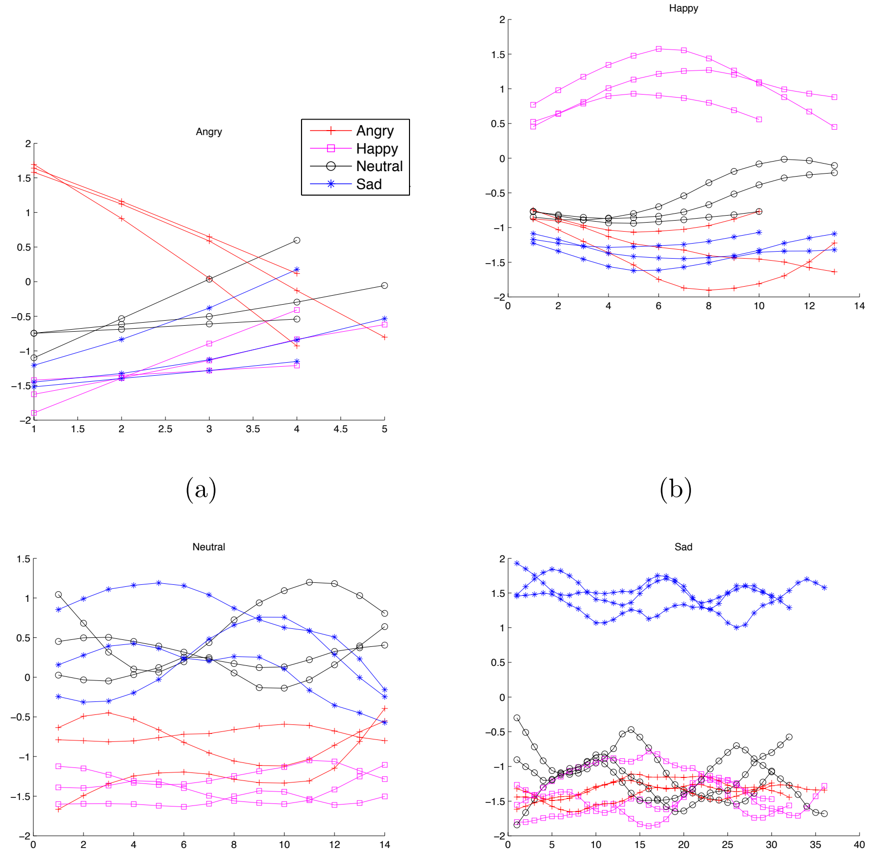
Table 4.1: *Unweighted classification accuracy (%) across different window lengths for each expression type: (A) Prototypical, (B) Non-prototypical, and (C) Combined.*

over, our DTW-kNN frameworks demonstrate the greatest performance gain for the data most challenging for automatic emotion recognition systems, non-prototypical data.

One benefit of our method is that it provides insight into the nature of inter-class similarity. We visualize the time-series similarity distance matrix in Figure 4.4. The DTW distances correspond to the five window sizes of: 0.25, 0.5, 1, 1.5, and 2 seconds (left to right). The diagonal blocks of each distance matrix represent the distance between the utterances with the same emotion class. Darker regions indicate stronger similarity between the dynamics of the utterances. The dark regions on the off-diagonals of the distance matrices demonstrate that there exists confusion between *Neutral* and *Sad*, and between *Neutral* and *Angry*. This confusion mirrors the common classification error between the classes of neutrality and sadness.

The distance matrix also permits an analysis of the structural patterns that are similar. We present utterances that are similar using the MD-DTW formulation. This provides an interpretable description of typical flow patterns for each emotion class (Figure 4.3). In the figures, all utterances shown are correctly classified using the proposed DTW-kNN framework. The angry utterances demonstrate an interesting trend from high confidence in the presence of anger to a more mixed appraisal of emotional message. The happy trends show a peaked happiness behavior. The sad utterances display slight fluctuations in expression. The neutral utterances depicted have irregular flow patterns even though they are correctly classified (See Figure 4.3(c)). This can explain the relatively low classification accuracy of neutral utterances compared to that of the other emotion classes.

In this study we propose a new framework to characterize utterances based on interpretable measures of affective dynamics. We use DTW to align our affective estimates of emotion flow and then classify using kNN with DTW similarity measure. This allows us to evaluate the discriminative power of the framework. The speaker independent experimental results are presented across five different window sizes, 0.25, 0.5, 1, 1.5, and 2 seconds for prototypical, non-prototypical, and combined data. Our results show that the proposed method outperforms the feature trajectory, the static EP, and the HMM baseline models. The highest improvement in our model comes from the classification of non-prototypical, or emotionally subtle, utterances. The novelty of our work is in its explicit modeling of the temporal flow patterns of emotion estimates. By taking into account the long-range dynamics of human emotion, we can have more natural and interpretable modeling techniques for emotion dynamics.



(a)

(b)

(c)

(d)

Figure 4.3: A subset of utterances described by emotograms and chosen for visualization purposes, which are correctly classified by our framework: (a) Angry, (b) Happy, (c) Neutral, and (d) Sad utterances

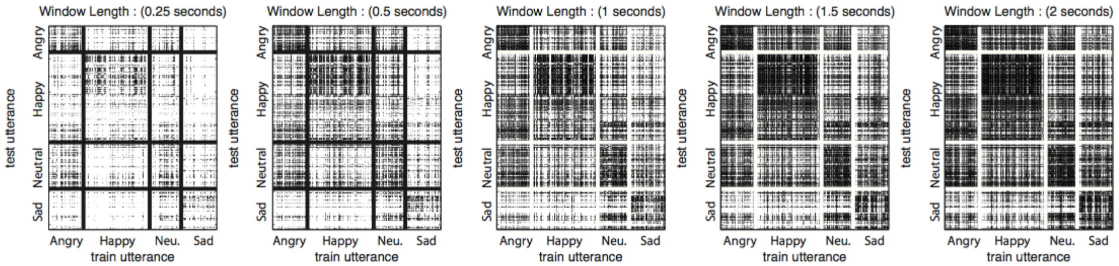


Figure 4.4: *MD-DTW distance matrices between Angry, Happy, Neutral, Sad utterances (combined data). Dark represents similar patterns.*

Part II: Multiple Factors in Behavior

CHAPTER 5

Speech-Related Variability in Facial Expressions

5.1 Introduction

The expression of emotion is complex. It modulates facial behavior, vocal behavior, and body gestures. Emotion recognition systems must decode these modulations in order to gain insight into the underlying emotional message. However, this decoding process is challenging; behavior is often modulated by more than just emotion. Facial expressions are strongly affected by the articulation associated with speech production and speech emphasis. Robust emotion recognition systems must differentiate speech-related articulation from emotion variation (e.g., differentiate someone saying "cheese" from smiling; discriminating between eyebrow raise for emphasis and due to excitement). In this chapter we explore methods to model the temporal behavior of

The work presented in this chapter has been published in the following articles:
Yelin Kim and Emily Mower Provost. "Say Cheese vs. Smile: Reducing Speech-Related Variability for Facial Emotion Recognition." Proceedings of the ACM International Conference on Multimedia (ACM MM), 2014 (Winner, **Best Student Paper Award**; Accepted as an Oral Presentation. Acceptance rate: 19%).
Yelin Kim and Emily Mower Provost. "Emotion Recognition During Speech Using Dynamics of Multiple Regions of the Face." ACM Transactions on Multimedia Computing, Communications and Applications (ACM TOMM), Special Issue on ACM Multimedia Best Papers, 2015. **Invited paper**.

facial motion with the goal of mitigating speech variability, focusing on temporal segmentation and classification methods. The results suggest that proper segmentation and classification are critical for emotion recognition.

One common method for constraining speech variation is by first segmenting the facial movement into temporal units with consistent patterns. Commonly, this segmentation is accomplished using known phoneme or viseme boundaries. We refer to this process as *phoneme segmentation* . The resulting segments are then grouped into categories with similar lip movements (e.g., /P/, /B/, /M/, see Table 5.1). Emotion models are trained for each visually similar phoneme group, a process we refer to as *phoneme-specific classification*. These two schemes have been used effectively in prior work [15, 104, 142, 148]. However, it is not yet clear how facial emotion recognition systems benefit from each of these components. Moreover, these phoneme-based paradigms are costly due to their reliance on detailed phonetic transcripts. In this chapter we explore two *unsupervised* segmentation methods that do not require phonetic transcripts. We demonstrate that phoneme segmentation is more effective than fixed-length sliding window segmentation. We describe a new unsupervised segmentation strategy that bridges the gap in accuracy between phoneme segmentation and fixed-length sliding window segmentation. We also demonstrate that phoneme-specific classification can still be used given unsupervised segmentation by coarsely approximating the phoneme content present in each of the resulting segments.

Studies on variable-length segmentation and the utility of phoneme-specific classification have recently received attention in the emotion recognition field. Mariooryad and Busso studied lexically constrained facial emotion recognition systems using phoneme segmentation and phoneme-specific classification. They introduced feature-level constraints, which normalized the facial cues based on the underlying

We refer to the process as phoneme segmentation when the vocal signal, rather than the facial movement, is used to segment the data. These audio-derived segmentation boundaries are applied to the facial movement.

Group	Phonemes	Group	Phonemes
1	P, B, M	8	AE, AW, EH, EY
2	F,V	9	AH,AX,AY
3	T,D,S,Z,TH,DH	10	AA
4	W,R	11	AXR,ER
5	CH,SH,ZH	12	AO,OY,OW
6	K,G,N,L,HH,NG,Y	13	UH,UW
7	IY,IH, IX	14	SIL

Table 5.1: *Phoneme to viseme mapping.*

phoneme, and model-level constraints, phoneme-specific classification [142]. Metallinou et al. [148] also proposed a method to integrate phoneme segmentation and phoneme-specific classification into facial emotion recognition systems. They first segmented the data into groups of visually similar phonemes (visemes) and found that the dynamics of these segments could be accurately captured using Hidden Markov Models. These methods demonstrate the benefits of phoneme segmentation and phoneme-specific classification. However, the challenge is the need for a phonetic transcript to both identify phoneme boundaries and to assign phoneme content. In this chapter, we explore the first challenge: the identification of phoneme boundaries and ask whether we can find other segmentation strategies for facial movement.

In this chapter, we propose an unsupervised segmentation strategy to circumvent this requirement. We investigate both the application of sliding windows in addition to segmentation using the natural temporal dynamics of the underlying signal. Sliding-window segmentation is a strategy commonly employed in emotion recognition studies [155, 159, 168, 198]. In this strategy, the facial data are segmented into smaller units, all with the same duration. However, this method is not based on the underlying dynamics of the signal and may miss important patterns in the signal. Further, previous work has demonstrated that the use of segmentation based on fixed length windows performs more poorly than phoneme segmentation [142]. To overcome this limitation, we propose an automatic, unsupervised segmentation method based on mouth movement, which utilizes a trajectory segmentation algorithm proposed by Lee et al. for trajectory segmentation and clustering [136]. The algorithm was motivated by Minimum Description Length (MDL) principle, widely used in information theory. Our proposed method does not require a phonetic transcript and achieved comparable performance to phoneme segmentation when used as a component of a facial emotion recognition system.

We also assess the utility of unsupervised segmentation approaches by testing our method using two emotion datasets to understand the impact of variable-length segmentation (i.e., unsupervised MDL-based segmentation and phoneme segmentation) and viseme-group classification on facial emotion recognition systems. We discuss the specific effects of the proposed segmentation and classification strategies across two different motion-capture datasets recorded in different settings: read speech (SAVEE) and two-person conversation (IEMOCAP). We found that when using viseme-group classification it is advantageous to use variable-length segmentation compared to fixed-length segmentation. Further, we analyze the impact of individual facial regions. The results demonstrate that we can increase system-level performance by changing how we integrate information from the facial regions. The results strengthen our argument that both variable-length segmentation and viseme-group classification are critical for facial emotion recognition systems.

5.2 Proposed Method

The overview of our proposed method is shown in Figure 5.1. We first separate the tracked marker positions into six facial regions to capture the facial region-specific characteristics in emotion expression (Section 5.2.1). We then temporally segment the data using three segmentation methods (fixed-length, phoneme, and MDL-based; Section 5.2.2) and measure the time-series similarity between the identified segments using Dynamic Time Warping (DTW). We calculate the distribution of

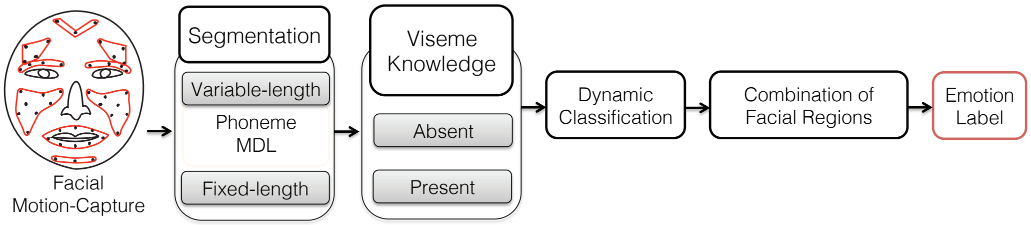


Figure 5.1: *Our system uses facial motion-capture data. It investigates three segmentation methods and explores the benefit of using knowledge of the spoken content of each segment. The system estimates the emotion label by estimating the similarity of the movement in each segment to movement observed in specific emotion classes. Finally, it combines the emotion estimates provided by the individual facial regions to infer a final estimated emotion label.*

emotion classes over each segment and use this information to estimate the emotion class of the segment. We aggregate the segment-level emotion estimates over the utterance to estimate the utterance-level emotion, described in detail in Section 5.2.4. During classification, we explore the benefit of using viseme-group classification given each of the three segmentation strategies. This allows us to understand the impact of using knowledge of the viseme group in classification. We refer to classification as *general* (contrasted with *viseme-group*) when we do not take the knowledge of viseme information into account, described in detail in Section 5.2.3. Finally, we investigate different methods to combine the emotion evidence derived from the individual facial regions, described in Section 5.2.5.

In our experiments, we test six approaches that use combinations of different temporal segmentation and classification methods, originally proposed in [113]. The two rows in Table 5.2 describe the classification scheme: general and viseme-group classification, and the three columns describe the segmentation scheme: phoneme, MDL, and fixed-length sliding window.

		Segmentation		
		Phon	MDL	Win
Classification	General	Gen/Phon	Gen/MDL	Gen/Win
	Viseme-group	VG/Phon	VG/MDL	VG/Win

Table 5.2: *Summary of the abbreviations associated with the six approaches tested in this chapter.*

5.2.1 Motion Capture Preprocessing

Both the IEMOCAP and SAVEE datasets provide facial markers that are (1) translated so that a nose tip becomes the origin of each frame, and (2) rotated to compensate for head movement. In addition, we perform mean-normalization on the facial data of individual speakers to mitigate their different facial configurations. The mean-normalization method was suggested in [148]. We compute the global mean value over all speakers for each marker coordinate and scale each individual speaker's data to make the mean of of each speaker to be the same as the global mean.

We divide the facial motion capture data into six facial regions to study region-specific facial movements, including: chin, forehead, cheek, upper eyebrow, eyebrows, and mouth, as in [113]. As shown in Figure 2.3, there are three markers are in the *Chin* and *Forehead regions*, 16 markers in the *Cheek region*, and eight markers in the *Upper Eyebrow, Eyebrow*, and *Mouth regions*. We track the region-specific marker positions and represent each as a multi-dimensional trajectory. For instance, given a data segment with N motion-capture frames and M marker coordinates (3-D for IEMOCAP and 2-D for SAVEE), the final data are an $N \times M$ trajectory.

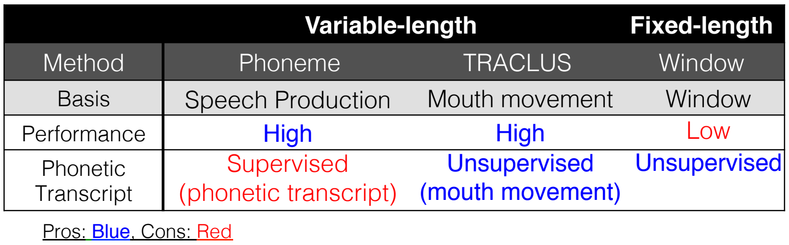
We exclude data with less than seven frames (threshold number of frames were chosen empirically) or 0.058 seconds. Our preliminary work demonstrated that the exclusion of segments with short durations does not make significant changes in emotion classification accuracy, which may due to insufficient temporal information within the segments. The computation time during DTW calculation can be considerably improved by excluding such segments, since 43.5% , 0.67%, and 0.86% of all phoneme, MDL, and fixed-length segments in the IEMOCAP dataset has duration less than seven frames, respectively. The high percentage of excluded phoneme segments occurs because many of the phonemes in the data have very short durations. Further, we drop segments with any missing values in the 46 markers we use. This results in different sets of utterances for each segmentation scheme. We use the set of 3,060 intersecting utterances. This number is slightly higher than in [148] and similar to [142]. In the SAVEE dataset, 34.86% of the phonemes are rejected, 2.23% of the MDL segments are rejected, and 1.44% of the window segments are rejected. The number of utterances remains the same after the exclusion process.

Our preliminary experiments showed that the difference between the SAVEE and IEMOCAP datasets in terms of frame rate (60 fps and 120 fps, respectively) impacted the overall accuracy. We mitigate this effect by increasing the SAVEE frame rate to 120 fps using cubic spline interpolation. This interpolation allows us to apply the same pre-processing steps to SAVEE as applied to the IEMOCAP (e.g. excluding of segments less than seven frames). In addition, our preliminary investigations showed that the SAVEE dataset had marker coordinates had a smaller range than IEMOCAP. This difference in range affected the MDL segmentation process. To mitigate this effect, for each marker coordinate, we scaled the SAVEE data to have the same minimum and maximum value as the IEMOCAP data. After MDL segmentation, we used the original marker values without scaling for the remainder of the classification framework to retain the original characteristics of the SAVEE dataset.

5.2.2 Segmentation

5.2.2.1 Sliding Window Segmentation ("Win")

The Win segmentation method segments each utterance into fixed-length windows. We use window segments without overlapping to enable comparisons with the phoneme and MDL segmentation methods, which do not have overlapping windows.



	Variable-length		Fixed-length
Method	Phoneme	TRACLUS	Window
Basis	Speech Production	Mouth movement	Window
Performance	High	High	Low
Phonetic Transcript	Supervised (phonetic transcript)	Unsupervised (mouth movement)	Unsupervised

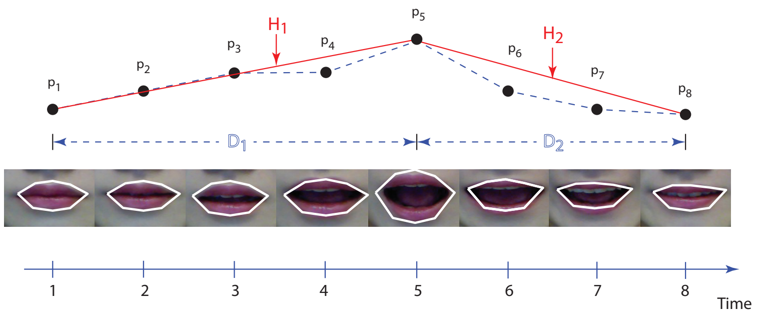
Pros: Blue, Cons: Red

Figure 5.2: *Comparison between the three segmentation methods. Phoneme segmentation has been shown to outperform window segmentation [142]. However, it is a supervised method that requires a phonetic transcript to segment the data. We propose an unsupervised method, MDL segmentation, to segment the data using the natural dynamics of the mouth. MDL, like fixed-window segmentation, does not require a phonetic transcript and, like phoneme-segmentation, captures the intrinsic dynamics of mouth.*

We retain all windows, including segments at the end of an utterance that are shorter than the standard window size. For instance, consider an $N \times M$ trajectory of the eyebrow region over an utterance, where N is the number of frames and M is the number of marker coordinates ($N = 128$ and $M = 24$ for the IEMOCAP data). If we segment this trajectory using 0.1-second window there will be 12 frames per window. The resulting segments of this utterance are ten trajectories, each of size 12×24 and one trajectory of size 8×24.

5.2.2.2 Phoneme Segmentation ("Phon")

The Phon segmentation method segments the facial data within an utterance based on the temporal phoneme boundaries. For instance, if a speaker is saying "hello", we segment the facial trajectories using the phoneme boundaries between /SIL/, /HH/, /AH/, /L/, /OW/, and /SIL/ phonemes. The set of phonemes that we use in this study is in Table 5.1. The boundaries for these phonemes were obtained by force aligning the audio to the known transcript. The average length of phoneme segments is 0.17 ± 0.01 seconds for the IEMOCAP data, and 0.14 ± 0.01 seconds for the SAVEE data.



$$L(H_1) = log_2(len(p_1p_5))$$

$$L(D_1|H_1) = \sum_{k=1}^{5-1} log_2(d_{\perp}(p_1p_5, p_kp_{k+1}) + log_2(d_{\theta}(p_1p_5, p_kp_{k+1}))$$

Figure 5.3: *An example of the MDL segmentation method for the visualization purpose. The x-axis describes time, and y-axis describes example marker positions of a coordinate of one of the mouth markers over an utterance (e.g., a marker on the top of the lip). The blue dashed lines represent the actual marker position changes, whereas the red lines represent the segmented results using the MDL principle. The proposed MDL segmentation method finds temporal points at which the dynamics of the mouth movement change. In this example, the mouth opens widely and then starts to close at frame* p_5. *Therefore, MDL uses* $\{c_1 = 1, c_2 = 5, c_3 = 8\}$ *(including the starting and end point of each utterance) as characteristic points. The hypothesis* H_1 *and* H_2 *correspond to the segmentation based on the characteristic points, lines between* p_1 *and* p_5 *and* p_5 *and* p_8. *The data* D_1 *and* D_2 *are the original mouth movement* $\{p_1p_2, p_2p_3, p_3p_4, p_4p_5\}$, *and* $\{p_5p_6, p_6p_7, p_7p_8\}$.

5.2.2.3 MDL Segmentation ("MDL")

We describe an unsupervised variable-length segmentation that does not require a phonetic transcript. We segment the data using the movement of the mouth (Figure 5.2). This allows us to capture the facial cues that are most highly related to speech production, important due to the focus on viseme-group classification. The segmentation algorithm was originally proposed in the context of a trajectory segmentation and clustering algorithm, called TRACLUS [136]. It uses MDL to automatically find points that should be used to segment regions of the data with different temporal characteristics. The application of this algorithm in the context of facial movement allows us to segment the facial data based on the natural dynamics of the mouth. A mouth-based example is presented in Figure 5.3.

The segmentation is based on the MDL principle, widely used in information theoretic studies. The goal of MDL is to find a hypothesis, H, that describes the original data D trading off between conciseness and preciseness of the description. It aims to minimize the sum of $L(H) + L(D|H)$, where $L(H)$ computes the conciseness of the hypothesis, and $L(D|H)$ the preciseness of the hypothesis. It finds points after which the trajectory changes from the current dynamic pattern. For instance, Figure 5.3 shows the mouth trajectory example where, after frame p_5, the trajectory changes. In this case, MDL would identify $\{c_1 = 1, c_2 = 5, c_3 = 8\}$ (including the starting and end point of the trajectory) as *characteristic points*. Characteristic points mark the beginning and end of regions with consistent dynamics. In [136], the authors proposed to measure $L(H)$ as the length of the proposed segmentation (e.g., in Figure 5.3, $L(H_1)$ is measured as the log of the length of a line connecting p_1 to p_5). The quantity $L(D|H)$ captures the difference between the original line segments, D and the proposed segmentation, H. For example, in Figure 5.3, $L(D_1|H_1)$ is the log of the summation of differences between each of the blue dashed lines $p_1p_2, p_2p_3, p_3p_4, p_4p_5$, and the red line p_1p_5. The segmentation can be formulated as an optimization problem, Equation 5.1.

$$\begin{aligned}
\arg\min_{H} \quad & L(H) + L(D|H) \\
\text{where} \quad & L(H) = \sum_{j=1}^{n-1} \log_2\left(len(p_{c_j}p_{c_{j+1}})\right), \\
& L(D|H) = \sum_{j=1}^{n-1} \sum_{k=c_j}^{c_{j+1}-1} \log_2\left(d_{\perp}(p_{c_j}p_{c_{j+1}}, p_k p_{k+1})\right) \\
& \qquad + \log_2\left(d_{\theta}(p_{c_j}p_{c_{j+1}}, p_k p_{k+1})\right)
\end{aligned} \tag{5.1}$$

In Equation 5.1, $d_{\perp}$ is the perpendicular distance between the line segments and

d_θ is the angular distance [37].

As an approximate solution, the TRACLUS algorithm [136] compares the cost of partitioning, $cost_{par}$, and non-partitioning, $cost_{nopar}$, at each data point, p, Equation 5.2.

$$\begin{aligned} cost_{par} &= L(H) + L(D|H) \\ cost_{nopar} &= L(D) = \sum_{j=1}^{p-1} \log_2 \left(len(p_j p_{j+1}\right) \end{aligned} \tag{5.2}$$

The algorithm advances through the trajectory and estimates whether the data should be segmented at each point. The algorithm makes a segmentation decision based on the equation: $cost_{par} \geq cost_{nopar} + MDL_{Advantage}$. When this equation is true the algorithm identifies the characteristic point as the previous point, marking the end of a segment. The characteristic point is the point prior to the one where the cost of partitioning is suddenly higher than the cost of not partitioning. The point at which the inequality is true then forms the beginning of the next segment. It is important to note that the parameter $MDL_{Advantage}$ controls the granularity of the segmentation and hence the average of segment length. We describe the method that we use to choose $MDL_{Advantage}$ in Section 5.2.4.1. Additional details can be found in [136].

In our work, the input to MDL segmentation is the mouth trajectory (24-dimensional for IEMOCAP and 18-dimensional for SAVEE) smoothed using a median filter with a window size of three (window size chosen empirically), to smooth the 3D-captured mouth movement trajectory.

5.2.3 Knowledge of Viseme Information

Studies of visual speech production have indicated that there are groups of visemes with similar facial movements (Table 5.1) [138]. Recent research has found that it is beneficial to separate emotion classifiers into 14 similar viseme groups, so that each classifier has less speech-related variation [142, 148]. We add to this knowledge by

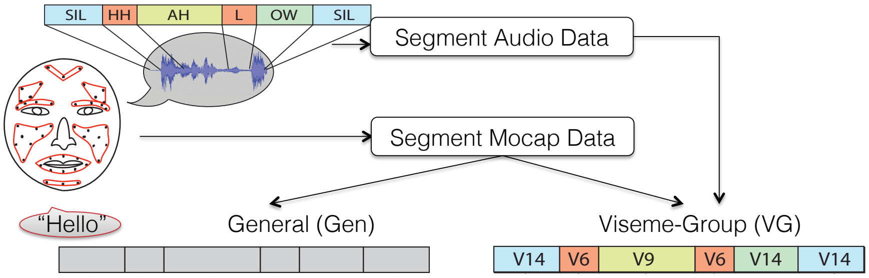


Figure 5.4: *Comparison between general and viseme-group classification methods, for an example in which a speaker is saying "hello." In viseme-group classification, we assign a viseme group that present in the longest duration within each of the segmented data, and separate the segments into different emotion classifiers based on its assigned viseme group.*

understanding how segmentation affects the utility of viseme-group classification.

We use two classification schemes: viseme-group and general classification. In viseme-group classification (VG), it is assumed that the classifier knows which viseme group the segment belongs to. We implement this by assigning a viseme group label to each segment based on the phoneme content. For example, if the speaker says "hello", we have /SIL/ (silence), /HH/, /AH/, /L/, /OW/, and /SIL/ phonemes. The two /SIL/ phonemes will be compared in emotion classifier 14, and /HH/ and /L/ in classifier 6, etc (Figure 5.6 and Table 5.1). In general classification (Gen), it is assumed that this knowledge is absent. This results in a single emotion classifier that has data from all viseme groups (Figure 5.4).

For MDL and Win segmentation, the segment boundaries may not line up with the phoneme boundaries. To estimate the corresponding viseme group of MDL and Win segments, we assign a viseme group label to a segment based on the phoneme that occupies the longest duration within each segment, for VG/MDL and VG/Win. For instance, in Figure 5.5 we consider a VG/MDL example. Note the mismatch between the phonetic transcript (dashed line in the figure) and the MDL segmentation result (hash marks). If the first MDL segment is 85% /SIL/ and 15% /HH/, we assign the phoneme content of the segment to the /SIL/ group, and apply emotion classifier 14.

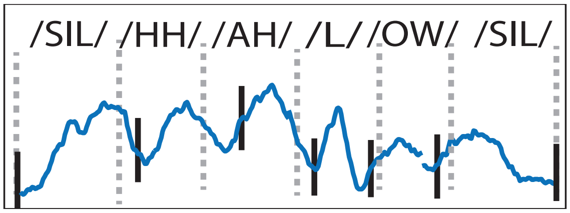


Figure 5.5: *VG/MDL example for describing how to assign phoneme content to MDL segments. The dashed lines show the phoneme boundaries and the hash marks represent different segment boundaries. Notice the potential mismatch between the hash marks and dashed lines in MDL segmentation.*

5.2.4 Emotion classification

5.2.4.1 Cross validation

Our proposed methods have two hyper-parameters: $MDL_{Advantage}$ (MDL segmentation) and window length (Win segmentation). We choose $MDL_{Advantage}$ from the set $\{0, 6, 10, 20\}$. We choose the window length from the set of $\{0.1, 0.25, 0.5, 1, 1.5, 2\}$ seconds.

We build speaker-independent emotion recognition systems using leave-one-speaker-out cross validation, and tune the parameters ($MDL_{advantage}$ and window length) using leave-one-training-speaker-out cross validation. For each speaker in the training set, we hold out a speaker as a validation speaker and train the model using the rest of the training speakers. We repeat this process over each training speaker and calculate the average of the validation accuracy. We choose the value of the parameter that maximizes performance over the set of validation speakers. For the SAVEE dataset, we also do lexical-independent classification to ensure that the same sentence does not appear in both the training and testing sets. This is because SAVEE is a read emotional speech database that has emotion-specific sentences (12 of 15 sentences were emotion-specific for each emotion class: Angry, Happy, and Sad).

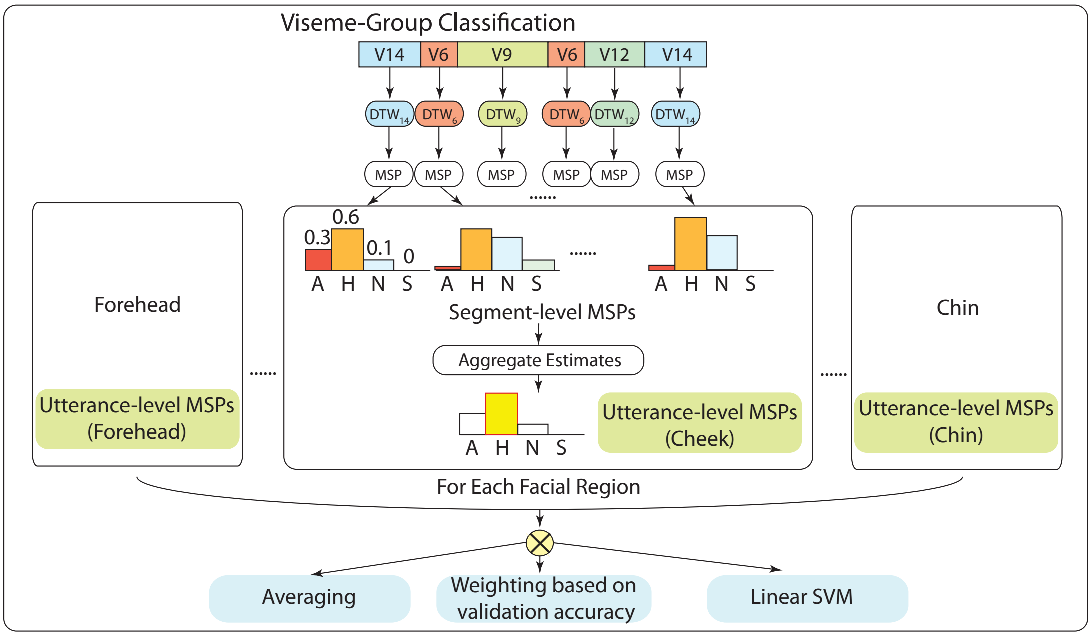


Figure 5.6: *MSP calculation and facial-region combination example for viseme-group classification.*

5.2.4.2 DTW-Motion Similarity Profile Emotion Classification

We use the DTW-Motion Similarity Profile (MSP) method proposed in [113] to infer utterance-level labels based on the temporal similarity between segments, as shown in Figure 5.6. The DTW method is computationally costly in the inference stages since it compares a test data to all training data. However, the method can provide interpretable descriptions about how two facial movements are similar. This method has two steps for emotion classification: (1) segment-level DTW calculation and (2) utterance-level emotion inference.

First, we calculate the segment-level similarity in facial movement between the training and test segments. For instance, if we have two K-dimensional facial movement trajectories of length M_1 and M_2, i.e., $T_1 \in \mathbb{R}^{M_1 \times K}$ and $T_2 \in \mathbb{R}^{M_2 \times K}$, we compute the similarity between the two trajectories using the multi-dimensional the algorithm. It computes the M_1-by-M_2 local cost matrix Q as follows, where i and j denote the frame-level temporal point of T_1 $(1 \leq i \leq M_1)$ and T_2 $(1 \leq j \leq M_2)$, respectively:

$$Q(i,j) = \sum_{k=1}^{K}(T_1(i,k) - T_2(j,k))^2, \tag{5.3}$$

Then, for each facial region, we calculate the emotional distribution of the c closest training segments, called the segment-level MSP, where c is set as 20 based on preliminary analyses. For instance, the first segment-level MSP in Figure 5.6 represents a four-dimensional vector $\{0.3, 0.6, 0.1, 0\}$, calculated based on the emotion labels of the 20 closest training segments: 6 angry, 12 happy, 2 neutral, and 0 sad.

Once we have segment-level MSPs for each segment, we average these to create an utterance-level MSP, a single four-dimensional emotion estimate for each facial region. We combine individual face regions with different methods (described in Section 5.2.5) to obtain the final utterance-level MSP. We normalize each of the four dimensions using speaker-specific z-normalization, to mitigate the imbalance in the emotion class distribution, e.g., there are approximately twice as many happy utterances compared to the other emotions. We assign the final utterance-level label based on the maximum component of the aggregated MSP, e.g., *happiness* in Figure 5.6.

5.2.5 Combination methods of Facial Regions

We investigate three types of decision-level combination methods of individual face regions: (i) simple averaging, (ii) weighted averaging, and (iii) SVM-based aggregation methods. The last stage of Figure 5.6 demonstrates that we combine utterance-level MSPs of individual face regions and explore the three combination methods.

5.2.5.1 Averaging

For the simple averaging method, we use ten different types of experiments to aggregate the MSPs from the individual facial regions.We report the ten AV('AVeraged faces') experimental results of (i) AV6 (averaged over all 6 face regions), (ii) AV4 (averaged over Chin, Cheek, Upper eyebrow, Mouth), (iii) AV Up (averaged over Forehead, Upper eyebrow, and Eyebrow) and (iv) AV Low (averaged over Chin, Cheek, and Mouth), and six individual facial regions of (v) Chin, (vi) Forehead, (vii) Cheek, (viii) Upper eyebrow, (ix) Eyebrow, and (x) Mouth. Unlike the previous work where we used segments with the same parameters (e.g., windows of the same fixed length, segments found using the same $MDL_{Advantage}$ parameter) over all AV experiments, we use the parameters chosen for individual AV experiments based on the cross validation accuracy (as described in 5.2.4.1). Different {speaker, classification (Gen or VG) methods, segmentation (Win, MDL, or Phon) methods} sets have different parameters chosen for each of the AV experiments. For each AV experiment, the individual face regions use the same parameter and are combined to calculate the final MSP.

5.2.5.2 Weighting based on validation accuracy

In the second experiment, we aggregate the emotional evidence using a weighted average. This allows us to more strongly weight information from emotionally expressive areas of the face, compared to less emotionally expressive areas. We first identify the parameters that are associated with the highest performance for each facial region using cross validation (described in Section 5.2.4.1). We calculate the accuracy over the validation speakers and use these accuracies as the initial weights: Val_i. We sum the weights over the six facial regions and normalize each of the weights to ensure that they sum to 1. Then, rather than aggregating MSPs by averaging, we compute a weighted average using the learned weights.

5.2.5.3 Linear-Support Vector Machine

We investigate a third aggregation method, which allows for adaptation based on estimated emotional expressivity of the individual facial regions. We use linear-kernel Support Vector Machine (SVM), in order to find the weighted linear combination of the MSPs that are associated with the individual facial regions. The input to the

SVM is the six four-dimensional MSP estimates (associated with each of the six regions of the face). The goal of the SVM is to estimate the emotion class label. We select the parameter C (10^k) through cross validation, selecting over the set: $k = \{-6, -5, -4, -3, -2, -1, 0, 1\}$.

5.3 Experimental Results

We present results for each database (IEMOCAP and SAVEE). We describe the results in terms of the three combination methods: (i) detailed experiments for each of the 10 averaging methods, where each of the ten methods use the best parameters chosen by cross validation, (ii) weighting of individual facial regions based on cross validation accuracy, and (iii) linear-SVM based weighting. In addition to the three segmentation methods of Win, MDL, and Phon segmentation, we present the utterance-level ('Utt') performance for general classification, where utterances are used without any segmentation. To be consistent with previous multi-class emotion recognition research [130, 158], we use unweighted accuracy, or averaged recall, to calculate the average accuracy.

5.3.1 SAVEE Experiments

Significance Tests To the best of our knowledge, previous work on the SAVEE dataset did not employ significance tests [84]. Since the SAVEE dataset has four speakers, each speaking the same set of utterances, we develop Generalized Linear Mixed Models (GLMM) with binomial link function that predicts the correctness of each utterance and speaker, similar to [19]. The GLMM use mixed-effects models that incorporate both random and fixed-effects parameters. We develop the models that treat both test speakers and utterance IDs as the random effects. We then compare MDL and window segmentation, as well as Phon and window segmentation, each separately within VG and Gen classification. Hence, fixed effects of the GLM models are classification (Gen or VG), segmentation (Win, MDL, or Phon), and the interaction between classification and segmentation. For the random effects, we use both test speakers and utterance IDs. The response of our models is correctness of the emotion inference given each segmentation and classification methods, where the conditional distribution of the response given the random effects is assumed as the binomial distribution. We fit the models using `glmer` function, implemented in `R` [11]. In each experimental result, we claim significance in the accuracy between the MDL and window, as well as Phon and window segments when $p < 0.05$.

5.3.1.1 Averaging

Table 5.3 shows the results of the SAVEE dataset when the MSPs of individual face regions are averaged. We tested the system using the parameter sets chosen over the set of {0.1, 0.25, 0.5, 1, 1.5, 2} seconds for window lengths and over the set of {0, 6, 10, 20} for $MDL_{advantage}$ of MDL segments. We present the parameter sets averaged over the all four test speakers in Table 5.9 and will discuss the interpretation of these results in Section 5.4. We also tested the system using phoneme segments (average segment length of 0.14 seconds), and utterance-length segments (average utterance length of 3.84 seconds).

In the SAVEE dataset, AV 6 outperforms the other methods of averaging different facial regions. For the AV 6 experiment results, we found that VG classification is more accurate than Gen classification for variable-length segmentation. The performance increases, comparing Gen and VG classification, for both MDL segmentation (75.62% to 80.00%, $p < 0.05$)and Phoneme segmentation (75.42% to 79.59%, $p < 0.05$). The window segmentation does not demonstrate any improvement (both methods demonstrate an accuracy of 77.29%).

The results demonstrate that the accuracy increases when variable-length segmentation is used in place of fixed-length segmentation in viseme-group classification.In

Cla	Seg	AV4	AV6	AV up	AV low	Chin	FH	CHK	U.EYE	EB	MOU
VG	Win	71.67	**77.29**	72.92	63.96	39.79	59.79	64.38	55.21	72.92	65.42
	MDL	69.79	**80.00***	75.00*	63.13	37.50	58.13	63.33	56.46	77.08	65.63
	Phon	69.59	**79.59**	74.79	64.59	32.08	62.08	61.67	57.50	73.13	62.71
Gen	Win	71.04	**77.29**	68.54	66.25	38.33	55.00	65.83	54.79	64.79	63.96
	MDL	71.46	**75.62**	62.29	65.21	43.75*	54.58	61.67	56.04	63.96	62.08
	Phon	72.71	**75.42**	61.88	68.34	44.80	56.67	63.55	55.42	63.54	66.88
	Utt	69.17	79.38	76.46	61.04	35.21	56.04	65.63	56.46	**79.58***	58.33

Table 5.3: *SAVEE dataset average accuracy result (%) for AV4, AV6, AV Up, AV Low, Chin (CHI), Forehead (FH), Cheek (CHK), Upper eyebrow (U. EB), Eyebrow (EB), and Mouth (MOU) using the averaging method. Results in* ***bold*** *are the highest accuracy among the ten experiments. '*' indicates significant differences in the accuracy between the MDL and window, as well as Phon and window segments.*

particular, for the AV 6 experiment, the MDL (80.00%) and phoneme (79.59%) segments outperform the window segments (77.29%). The performance improvement of MDL over window segments is statistically significant ($p < 0.02$), whereas phoneme over window segments is not ($p = 0.29$). Moreover, we achieve comparable accuracy ($p = 0.190$) between our proposed MDL segmentation and phoneme segmentation. MDL significantly outperforms window segments in the AV up experiment, achieving 75.00% compared to 72.92% for window segmentation ($p < 0.05$). The VG/MDL method also achieves improvement compared to VG/Win for the eyebrow (EB), achieving 4.16% improvement ($p = 0.07$). The results provides evidence that MDL segmentation can be effectively used in emotion classification.

In Gen classification, the best results of MDL segments (75.62%, $p = 0.071$) and phoneme segments (75.42%, $p = 0.071$) are lower than that of window segments (77.29%), although the results are not significantly different. The highest accuracy is achieved with utterance-length segments (79.38%) for the AV 6 experiment. However, this phenomenon is not consistent over the different facial regions. For instance, in the mouth region, phoneme segments (66.88%) outperform window (63.96%) and

Cla	Seg	Average	Ang	Hap	Neu	Sad
VG	Win	**77.29**	81.67	76.67	76.67	88.33
	MDL	**76.46**	83.33	71.67	71.67	85.00
	Phon	**77.71**	83.33	76.67	62.50	88.33
Gen	Win	**75.00**	86.67	73.33	73.33	88.34
	MDL	**75.00**	93.34	66.67	66.67	90.00
	Phon	**77.08**	91.67	70.00	55.00	91.67

Table 5.4: *SAVEE dataset results of by weighting individual facial region based on its validation accuracy. We report (i) average accuracy, or averaged recall, (ii) angry, (iii) happy, (iv) neutral, and (v) sad class accuracy.*

utterance-length segments (58.33%), whereas MDL segments (62.08%) work slightly worse than window segments. For the chin, both variable-length segmentation strategies, MDL (43.75%) and phoneme (44.80%), outperform fixed-length segments, both window (38.33%) and utterance-length (35.21%) segments. The difference between the MDL and window segments were significant ($p < 0.05$). For the eyebrow, the utterance-length segments achieve significant increase compared to the other segmentation methods, achieving 79.58% accuracy. Overall, phoneme segments perform well for the lower facial regions, mouth and chin, whereas utterance-length segments perform well for regions less modulated by speech, such as the eyebrow.

5.3.1.2 Weighting based on cross validation accuracy

Table 5.4 demonstrates the SAVEE results when we weight the face region-specific MSPs based on validation accuracy. We found that weighting face region-specific MSPs lowered the accuracy of SAVEE (the opposite trend can be observed for IEMOCAP), although the decrease is not significant. VG/Win remains the same 77.29% accuracy, whereas VG/MDL accuracy decreases from 80.00% to 76.46%. For Gen classification, Gen/Win decreases from 77.29% to 75.00% and Gen/MDL decreases from 75.62% to 75.00%. We hypothesize that this is due to the high variability between

Cla	Seg	Average	Ang	Hap	Neu	Sad
VG	Win	**88.75**	80.00	95.00	88.34	91.67
	MDL	**92.08***	93.33	90.00	90.00	95.00
	Phon	**93.12***	95.00	93.33	89.17	95.00
Gen	Win	**88.75**	88.34	90.00	80.00	96.67
	MDL	**88.13**	95.00	81.67	79.17	96.67
	Phon	**88.96**	91.67	88.33	79.17	96.67

Table 5.5: *SAVEE dataset results of SVM based combination where individual facial region uses the best parameter chosen by cross validation. We report (i) average accuracy, or averaged recall, (ii) angry, (iii) happy, (iv) neutral, and (v) sad class accuracy. '*' indicates a significant increase compared to the baseline window segmentation methods*

speakers (e.g., one speaker has a significantly lower recognition rate, with a relative difference of about 20% from the other three speakers), and the lack of training speakers when calculating validation accuracy (i.e., only two speakers for training in cross validation). The per-emotion class accuracies demonstrate that *Anger* ($p < 0.05$, significant) and *Sadness* (not significant, $p = 0.052$) are well recognized compared to *Happiness* and *Neutrality*. The phoneme segments perform well in both VG (77.71%) and Gen (77.08%) classification, showing the highest performance among the three segmentation methods.

5.3.1.3 SVM-based weighting method

Table 5.5 demonstrates the results of linear-SVM based MSP combination. The hyper-parameter C of the SVM is chosen as 10^{-4} using cross validation. It is shown that the results are improved using linear-SVM, achieving up to 92.08% accuracy for VG/MDL, improving from 80.00% of the simple averaging method. This is a significant improvement in accuracy over VG/Win ($p < 0.03$). VG/Win, Gen/Win, and Gen/MDL also improve from 77.29% to 88.75%, 77.29% to 88.75%, and 75.62% to 88.13%, respectively. The phoneme segments perform the best for both VG (93.12%)

and Gen (88.96%) classification. VG/Phon outperforms VG/Win significantly ($p < 0.007$). The per-emotion class accuracies show improved performance for *Happiness* and *Neutrality*. We present the learned SVM weights in Figure 5.7 and will discuss the corresponding findings in the previous psychology studies on emotion perception in Section 5.4.

We hypothesize that the SVM-based weighting method more reliably captures the region-specific temporal characteristics compared to the weighting based on validation accuracy, since the SVM learns more general patterns across training speakers that are associated with emotion prediction compared to the direct validation accuracy. We discuss the learned SVM weights for each emotion prediction task in more detail in Section 5.4.

5.3.2 IEMOCAP

Significance Tests For the IEMOCAP dataset, we use paired t-tests to be consistent with previous work on this dataset [113, 142]. The paired t-test for leave-one-speaker-out cross validation has shown to be useful to test the significance of the difference [52, 113]. We claim significance when the p-value is less than 0.05.

5.3.2.1 Averaging method

Table 5.6 summarizes the average accuracy for each of the 10 different experiments. As in the SAVEE dataset, the two parameters, window length and $MDL_{Advantage}$, are chosen over the set of $\{0.1, 0.25, 0.5, 1, 1.5, 2\}$ seconds and $\{0, 6, 10, 20\}$. In VG classification, variable-length segments outperform window segments in the AV 4 experiment and AV low experiments. MDL segments outperform window segments in most of the experiments except for the upper face regions (the exceptions in upper face regions are not significant). In particular, variable-length segments outperform window segments significantly particularly in the lower face regions. In the AV low

Cla	Seg	AV4	AV6	AV up	AV low	Chin	FH	CHK	U.EYE	EB	MOU
VG	Win	54.30	**54.93**	47.14	52.44	50.36	42.92	43.24	45.21	42.93	53.86
	MDL	55.31	55.07	45.65	55.08*	53.28*	42.43	44.53*	46.04	42.16	**55.46**
	Phon	**56.14**	55.60	45.85	54.05	49.96	40.43	44.92*	45.07	40.14	56.07
Gen	Win	54.20	**54.24**	47.01	52.54	49.70	39.43	42.49	44.15	41.95	53.04
	MDL	**53.51**	51.92	42.39	51.56	46.94	38.97	42.56	42.33	38.33	52.04
	Phon	**54.04**	52.05	42.30	51.77	46.94	39.00	42.71	43.28	38.58	51.13
	Utt	40.02	**40.83**	39.70	40.02	35.51	39.70	25.00	25.00	25.00	37.10

Table 5.6: *IEMOCAP average classification accuracy (%) using six schemes of: (1) VG/Win, (2) VG/MDL, (3) VG/Phon, (4) Gen/Win, (5) Gen/MDL, (6) Gen/Phon, and (7) Gen/Utt (utterance-length) segments. The results are presented as mean over the 10 speakers. ‘*’ indicates a significant increase compared to the baseline window segmentation method.*

experiment, the performance gain when using VG/MDL compared to VG/Win is significant, achieving 2.64% ($p < 0.05$). The chin and cheek regions also showed significant increase compared to the window segments. For the chin, VG/MDL significantly outperforms VG/Win by 2.92% ($p < 0.05$). For the cheek, both VG/MDL and VG/Phon significantly outperform VG/Win by 1.29% and 1.68%, respectively (both $p < 0.05$). The significant improvement in the lower regions of the face when using the MDL segmentation may indicate that the mouth-based segmentation strategy of MDL performs well for the regions that are modulated by speech [34]. VG/Phon outperforms VG/Win by 1.84% in the AV4 experiment. VG/MDL also outperforms VG/Win by 0.74%. However, these differences are not statistically significant. Moreover, the mouth region achieves higher accuracy (55.46%) than the AV6 method for window segments (54.93%), although not significant.

In Gen classification, the accuracy between different segmentation methods was similar. Also, the utterance-length segmentation performed poorly (40.83%), unlike in the SAVEE dataset.

Cla	Seg	Average	Ang	Hap	Neu	Sad
VG	Win	**56.66**	57.59	69.83	33.19	66.04
	MDL	**57.57**	63.03	69.71	34.41	63.13
	Phon	**57.18**	61.76	68.69	40.39	57.90
Gen	Win	**56.02**	55.70	70.23	33.51	64.62
	MDL	**55.00**	59.93	69.48	34.08	56.48
	Phon	**53.76**	58.14	69.96	37.71	49.25

Table 5.7: *IEMOCAP dataset results of weighting individual facial region based on its validation accuracy. We report (i) average accuracy, or averaged recall, (ii) angry, (iii) happy, (iv) neutral, and (v) sad class accuracy.*

5.3.2.2 Weighting using validation accuracy

The weighting method that combines the individual facial regions based on cross validation accuracy improves the performance, up to 57.57% when using the VG/MDL method. This is the highest accuracy in the IEMOCAP dataset and it outperforms the simple averaging method in the AV6 experiment by 1.74% (not significant). This is higher than VG/Win method (56.66%), however the difference is not significant. The VG/MDL result is higher than both of the previous work [142, 148]. VG/MDL and VG/Win perform significantly better using the validation accuracy-based weighting method compared to the simple averaging method (55.07% to 57.57%, $p < 0.05$; and 54.93% to 56.66%, $p < 0.05$). The Gen classification is also improved compared to the simple averaging method, particularly for Gen/MDL. Gen/MDL improves from 51.92% to 55.00% in the AV6 experiment. The average accuracies (Gen/Win 56.02% and Gen/MDL 55.00%) are smaller than seen in the VG classification results. The phoneme segments perform better for VG classification (57.18%) compared to Gen classification (53.76%), as in the other experiments. Gen/Win performs significantly better than Gen/Phon ($p < 0.05$).

Cla	Seg	Average	Ang	Hap	Neu	Sad
VG	Win	**56.06**	66.16	77.28	16.58	64.23
	MDL	**56.63**	68.30	76.18	15.35	66.68
	Phon	**55.06**	67.66	74.94	17.48	60.16
Gen	Win	**53.98**	63.47	77.74	13.75	60.97
	MDL	**53.96**	62.12	76.52	15.85	61.35
	Phon	**52.22**	62.49	76.97	16.12	53.28

Table 5.8: *IEMOCAP dataset results of SVM based combination where individual facial region uses the best parameter chosen by cross validation. We report (i) average accuracy, or averaged recall, (ii) angry, (iii) happy, (iv) neutral, and (v) sad class accuracy.*

5.3.2.3 SVM-based Weighting Method

The hyper-parameter of SVM chosen based on cross validation was 10^{-5}. Linear-SVM slightly improves the accuracy for VG/MDL compared to AV6 in the simple averaging method (55.07%). It also slightly improves the VG/Win accuracy in the AV6 experiment (54.93%). For Gen classification, the accuracy was slightly lower than the averaging method. The differences are not significant.

For the VG classification, MDL segmentation achieves significantly higher accuracy compared to the simple averaging method (55.07% to 56.63%, $p < 0.05$). VG/Win also performs significantly better using the SVM weighting method compared to the simple averaging method (54.93% to 56.05%, $p < 0.05$).

5.4 Discussion

Table 5.9 shows the parameters selected for the SAVEE dataset, averaged over all four test speakers {1, 2, 3, 4}. Note that larger values of the $MDL_{advantage}$ parameter corresponds to longer average segment lengths. The parameters chosen for each facial region demonstrate that in general classification, the upper facial regions such as the eyebrow and forehead perform well with longer segments. This trend may indicate

Cla	Seg	AV4	AV6	AV up	AV low	Chin	FH	CHK	U.EYE	EB	MOU
VG	Win	0.18	0.21	0.14	0.34	0.18	0.33	0.14	0.10	0.53	0.18
	MDL	6	5	6.5	6.5	11.5	2.5	10	6	5	10
Gen	Win	0.28	0.81	1.28	0.61	0.43	0.59	0.10	0.68	1.15	0.14
	MDL	4.5	14	15	11.5	5	6.5	11.5	9	10.5	12.5

Table 5.9: *SAVEE dataset: selected parameters averaged over all four test speakers based on leave-one-training-speaker-out cross validation. For Win segmentation method, the parameter is segment length of each window; and for MDL segmentation method, the parameter is $MDL_{advantage}$, described in Section 5.2.2. Larger $MDL_{advantage}$ corresponds to longer average segment length. (Note that phoneme segmentation methods do not have any parameters that control granularity.)*

	Gen/Win			PS/Win		
Window size (sec)	AV Mou	AV4	AV6	AV Mou	AV4	AV6
0.1	52.41	55.51	52.93	54.24	55.06	55.03
0.25	52.04	55.04	52.49	52.63	54.83	54.92
0.5	52.09	55.24	53.78	52.41	55.04	55.19
1	50.79	53.54	54.29	52.12	53.41	53.98
1.5	50.96	54.36	54.13	51.45	53.95	53.8
2	50.12	53.11	52.11	50.49	51.72	51.75

Table 5.10: *Accuracy result comparisons of Gen/Win (top) and VG/Win (bottom) with different window sizes of the set* $\{0.1, 0.25, 0.5, 1, 1.5, 2\}$ *seconds (IEMOCAP).*

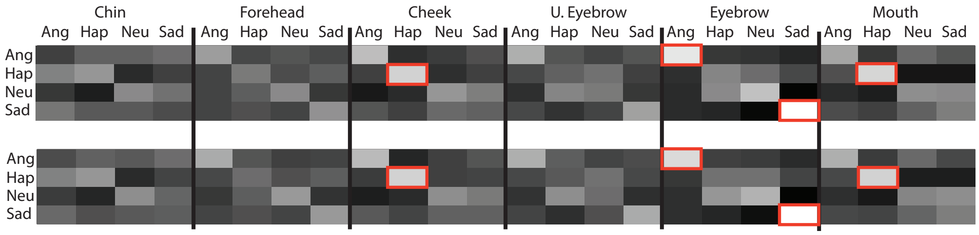


Figure 5.7: *SAVEE dataset: trained SVM weights to combine MSPs from the individual facial regions. Averaged weights for Gen/Win (top) and VG/Win (bottom) methods. Darker weights corresponds to smaller values, and brighter corresponds to larger values. Highlighted red boxes (best shown in color) represent the top four highest weights.*

that the upper facial regions may be characterized by longer-term dynamic patterns.

Figure 5.7 shows the trained SVM weights based on the 24-dimensional features for each face/emotion set for the SAVEE dataset. In the figure, there are six different facial regions and four different emotion classes. We investigate the contribution of each facial region to the final emotion inference. We estimate the contribution based on the SVM weights, e.g., $w_1 face_1 + w_2 face_2 + w_3 face_3 + w_4 face_4 + w_5 face_5 + w_6 face_6$. The weights w_i of each face region $i = \{1, 2, ..., 6\}$ are averaged over the four test speakers. We find that in both Gen and VG classification, the mouth regions have higher weights on the happiness component of the four-dimensional MSPs, whereas eyebrow regions have higher weights on the anger and sadness components. This finding corresponds to the previous emotion perception studies that certain facial regions contribute more to specific emotion perception [215]. The studies on facial action units [59] have shown that certain anatomical regions of the face, or action units, are strongly related to specific emotions. These studies have shown that happiness is strongly related to action units on the mouth (including action unit 6: cheek raiser and action unit 12: lip corner puller), and anger is strongly related to action units on the eyebrow (including action unit 4: brow lowered and action unit 7: lid tightener).

For the IEMOCAP dataset, we observe similar performance between Gen/Win, Gen/MDL, and Gen/Phon. This may imply that without any additional phoneme information it is important to capture longer-term dynamics to understand emotion expression. In addition, we compare the Gen and VG classification accuracies of the fixed-length segments with different window sizes, {0.1, 0.25, 0.5, 1, 1.5, 2} seconds. Table 5.10 summarizes Gen/Win (top) and VG/Win (bottom) accuracy for different window sizes. The Gen/Win accuracy shows statistically insignificant changes across different window sizes. However, the VG/Win accuracy shows significant increase in AV 6 accuracy between 1 and 2 seconds (2.18%, $p < 0.02$) and between 1.5 and 2 seconds (2.02%, $p < 0.03$). Both VG/Win and Gen/Win perform poorly with 2-second windows. This decrease in performance of the 2-second windows compared to smaller window sizes is higher for VG/Win compared to Gen/Win, which may suggest that in VG classification it is critical to use segments that have similar lengths to phoneme segments.

We also compare how many phonemes in each window segment with different sizes. For each window size of {0.1, 0.25, 0.5, 1, 1.5, 2} seconds, the average number of phonemes are 1.71 ± 0.80, 2.67 ± 1.49, 3.87 ± 2.23, 5.47 ± 2.93, 6.54 ± 3.24, and 7.19 ± 3.47.

The consistent increase in accuracy of VG/Win associated with an increase in window length may imply that the window segments that contain phoneme at the closest (i.e. most overlap with phoneme segment) will perform better than the others.

5.5 Conclusions

In this study, we investigate an unsupervised, variable-length segmentation method for compensating for facial movement due to speech, to improve the performance of facial emotion recognition systems. We present detailed results on two different datasets and propose a combination strategy that can account for different temporal characteristics of different facial regions. Our segmentation method is based on the MDL principle. We demonstrated that a hyper-parameter $MDL_{Advantage}$ can change the average segment lengths and how this impacts the system-level performance. Based on this finding, we show how we can combine different hyper-parameters chosen per face regions using cross validation for final emotion inference. We use linear-kernel SVMs to combine facial region-specific emotion evidence and investigate the weights between facial regions and emotions to explore the different contributions of individual facial regions for inference of specific emotion.

Our experimental results on the two IEMOCAP and SAVEE datasets demonstrate that the two variable-length segmentation methods, MDL and phoneme, achieve higher emotion classification rates compared to fixed-length window segmentation in VG classification. We also find that methods to combine estimated emotion from individual face regions, can increase the accuracy significantly.

In our future work, we will investigate the efficacy of MDL segmentation based on different facial regions. In our preliminary study in [113], we found that it is more beneficial to use the mouth region for MDL segmentation, compared to other facial regions. However, it is not yet clear whether this is true for other datasets such as SAVEE. For instance, we found that the SAVEE dataset shows high accuracy using utterance-level eyebrow segments, unlike the chance-level accuracy for the IEMOCAP dataset. This may indicate that the difference between the read speech (SAVEE) and more natural dynamic conversation (IEMOCAP) may have different facial movement characteristics. We will investigate the use of other facial regions for the other datasets in MDL segmentation.

Further, our results indicate that different segmentation strategies per different face regions, e.g., MDL segmentation for mouth region and emphasis-based segmentation for eyebrow region, may benefit the overall facial emotion recognition systems. In addition, we plan to combine the facial emotion recognition system that we developed with audio emotion recognition systems, that can use the multimodal information.

CHAPTER 6

Informed Segmentation and Labeling Approach

6.1 Introduction

Facial movements are highly dependent on multiple factors when a person is speaking, such as lexical content, speech emphasis, and emotion (Figure 6.1). A main challenge in automatic emotion recognition is that facial movements may be modulated by multiple factors. For instance, a person smiling and saying ‘cheese’ both result in similar mouth movements [113]. Likewise, eyebrows raised due to surprise have similar movements to eyebrows to convey emphasis [38].

Traditional systems often neglected the lower face region (particularly the orofacial region) for emotion recognition when a person is speaking, due to its high dependence on speech production [50]. However, recent studies have found that the use of phonetic information can improve the recognition rate on this region [113, 114, 141–143, 148]. Most notably, a set of studies by Mariooryad and Busso has found consistent improvement for facial emotion recognition in the lower face region, when the phonetic information is used both in supervised (phonetic transcript is available) [141, 142] and unsupervised manners [143]. The finding underlying these studies is that the restriction of an emotion recognition task by training separate classifiers for a given unit of speech production, e.g., phoneme, can help reduce speech-related variation in emotion recognition. However, it has been under-explored how the dynamics in

The work presented in this chapter has been submitted in the following article: **Yelin Kim** and Emily Mower Provost. “ISLA: A Framework for Controlling Sources of Modulation in Audio-Visual Affective Behavior.” IEEE Transactions on Affective Computing, 2016.

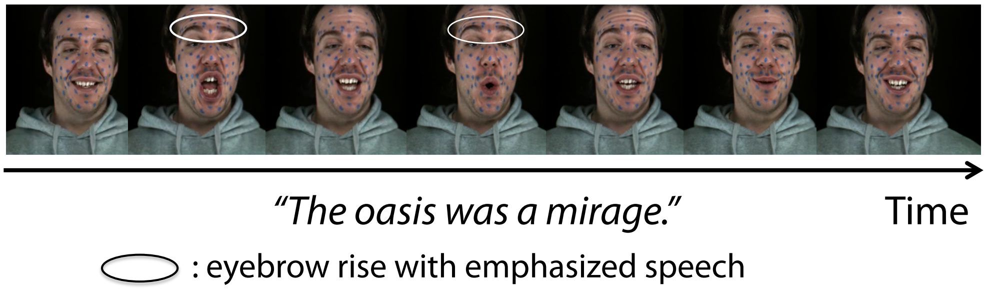

“The oasis was a mirage.”
Time
: eyebrow rise with emphasized speech

Figure 6.1: *An example facial display in the SAVEE dataset [84] that shows a person saying “the oasis was a mirage” with happy emotion. The eyebrow changes when a person is emphasizing (highlighted in dashed circles) and the orofacial region changes to articulate speech.*

the upper face region are altered due to speech [38]. This raises the central research question **(Q1)**, *can we find a signal that modulates the dynamics of the upper face movements and use patterns of this signal for estimating emotion in this region?*

In this chapter, we address this question by proposing the Informed Segmentation and Labeling Approach (ISLA). ISLA captures speech-related variability in both the upper and lower face regions, with an aim of improving the overall emotion recognition rate when a person speaks. A central feature of the proposed ISLA is what we call *ISLA signals*. ISLA uses these speech signals that alter the temporal dynamics of each of the upper and lower face regions, to temporally segment and classify facial displays of emotion. Our ISLA signals are motivated by recent findings in speech prosody that the upper face movements are modulated by speech emphasis [38, 75, 80, 108, 161]. Also, as previously shown to be useful in our prior work [113, 114], we use the phonetic information to segment and classify the lower face region.

The second research question **(Q2)** focuses on *the integrated design of audio-visual emotion recognition system.* As shown in Figure 6.2, we compute Emotion Profiles (EPs), which are vector representations of confidence about presence and absence of *anger, happiness, neutrality, and sadness* [158], for the upper face, lower face, and speech modalities. For the upper and lower face regions, we compute EPs based on

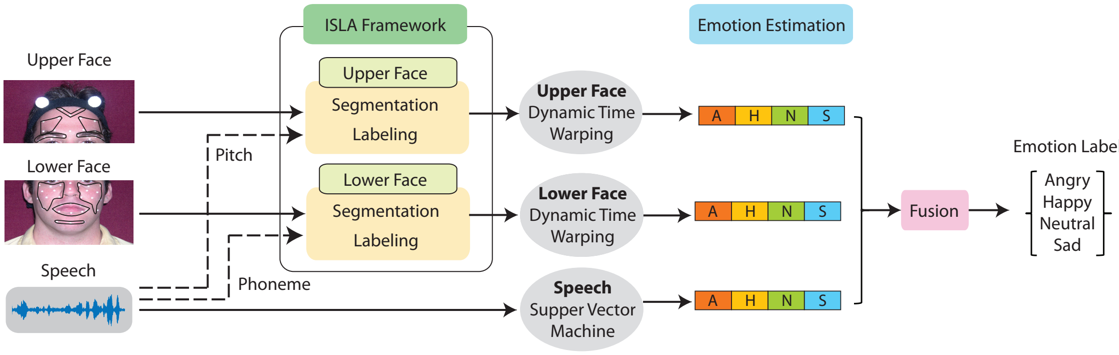


Figure 6.2: *An overview of our proposed audio-visual emotion recognition system. We calculate emotion estimates from the upper and lower face regions by using speech pitch and phoneme patterns for each of the two regions, respectively. The utterance-level emotion estimates from speech are calculated using Support Vector Machines. We combine utterance-level emotion estimates from the upper face, lower face, and speech modalities for two types of experiments: (i) audio-visual classification and (ii) correlation analysis between the modalities.*

time-series similarity between segments as in our prior work [113, 114]. Specifically, we calculate the similarity between training and test segments, and compute test EPs based on the emotion class distribution of the k closest training segments. For speech, we use outputs of binary Support Vector Machines (SVMs) to compute EPs, where each SVM separates '*emo*' vs. everything else ('*emo*' includes *anger, happiness, neutrality*, or *sadness*) as in [158]. The EPs calculated from the three modalities are averaged together to infer the final emotion label.

The third research question **(Q3)** asks *how the emotion expressions in the lower face, upper face and speech modalities coordinate with each other.* Previous studies have identified interrelationships between these modalities, but only for a single subject [24] or with absence of understanding the unique dynamics underlying each region of the face [140, 150]. To this aim, we examine the correlation between EPs from the lower face, upper face, and speech modalities, and provide insight into how the emotion inferences from these modalities correlate with each other.

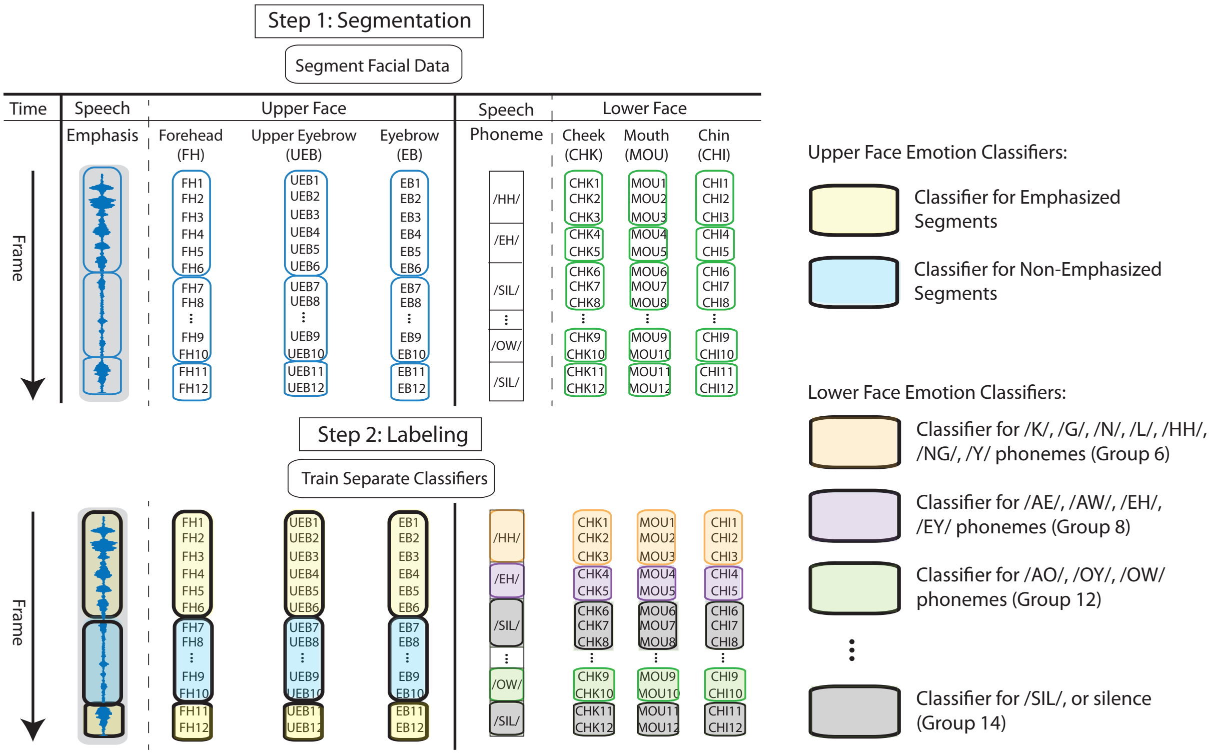

Step 1: Segmentation
Segment Facial Data
Time
Speech
Emphasis
Upper Face
Forehead (FH)
Upper Eyebrow (UEB)
Eyebrow (EB)
Speech
Phoneme
Lower Face
Cheek (CHK)
Mouth (MOU)
Chin (CHI)
Frame
FH1 FH2 FH3 FH4 FH5 FH6 FH7 FH8 FH9 FH10 FH11 FH12
UEB1 UEB2 UEB3 UEB4 UEB5 UEB6 UEB7 UEB8 UEB9 UEB10 UEB11 UEB12
EB1 EB2 EB3 EB4 EB5 EB6 EB7 EB8 EB9 EB10 EB11 EB12
/HH/
/EH/
/SIL/
/OW/
/SIL/
CHK1 CHK2 CHK3 CHK4 CHK5 CHK6 CHK7 CHK8 CHK9 CHK10 CHK11 CHK12
MOU1 MOU2 MOU3 MOU4 MOU5 MOU6 MOU7 MOU8 MOU9 MOU10 MOU11 MOU12
CHI1 CHI2 CHI3 CHI4 CHI5 CHI6 CHI7 CHI8 CHI9 CHI10 CHI11 CHI12
Step 2: Labeling
Train Separate Classifiers
Upper Face Emotion Classifiers:
Classifier for Emphasized Segments
Classifier for Non-Emphasized Segments
Lower Face Emotion Classifiers:
Classifier for /K/, /G/, /N/, /L/, /HH/, /NG/, /Y/ phonemes (Group 6)
Classifier for /AE/, /AW/, /EH/, /EY/ phonemes (Group 8)
Classifier for /AO/, /OY/, /OW/ phonemes (Group 12)
Classifier for /SIL/, or silence (Group 14)

Figure 6.3: *Overview of the ISLA framework with (i) segmentation and (ii) labeling steps (best shown in color). In this framework, speech emphasis becomes the ISLA signal for the upper face region, and speech phoneme becomes the ISLA signal for the lower face region. In Step 1, the framework first segments the facial data based on the dynamic changes of ISLA signal (top). In Step 2, ISLA trains separate classifiers with the segmented facial data based on the characteristics of the signal related to the face region, or ISLA signal, associated with each segment. These classifiers are shown on the right side of the figure.*

Our experimental results on two emotion datasets, IEMOCAP [27] and SAVEE [84], demonstrate that the proposed methods show promising results on emotion recognition tasks.

6.2 Proposed ISLA Approach

In this section, we describe the main idea behind the proposed ISLA framework. The ISLA framework can be used as a pre-processing step for facial emotion classification, and can be integrated with any classifier.

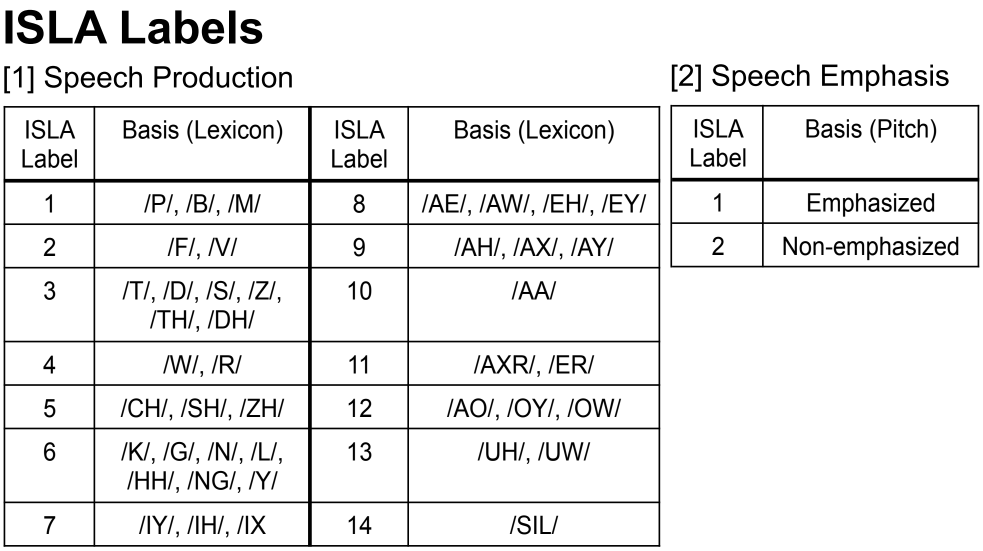
Figure 6.3 shows an ISLA example of how the movements in the upper face (forehead, upper eyebrow, eyebrow) and the lower face (cheek, mouth, chin) regions are segmented and labeled using the ISLA signals. We define ISLA signals as speech signals that are closely related to the dynamics of the target face region: speech emphasis for the upper face and phoneme for the lower face, as identified in previous speech prosody and emotion studies [38, 75, 80, 108, 113, 114, 142, 148, 161].

ISLA first segments the facial movements based on the dynamic changes in the ISLA signal ('informed segmentation'–Step 1 in Figure 6.3). The upper face segments are either emphasized or non-emphasized, and each of the lower face segments is a single phoneme. It then labels each segment where within each group the segments share the same factor of modulation (e.g., both co-occur with the phoneme /IY/; 'informed labeling'–Step 2 in Figure 6.3). These labeled groups of segments can be used to train separate classifiers for each group. This separation enables each classifier to restrict non-emotional factors of modulation and distills out emotion-specific variations between the segments (e.g., differentiating /IY/ movement from happy emotion and from angry emotion).

6.2.1 Informed Segmentation

The informed segmentation is designed to capture the natural dynamics of each facial region using the ISLA signals. As shown in the Figure 6.3 (Step 1), we use two types of ISLA signals, speech emphasis and speech phonemes, to segment the upper face and lower face regions, respectively. If the ISLA framework recognizes changes in emphasis (e.g., changing from non-emphasized speech to emphasized speech), then it segments each of the three upper face regions at the change point. If the ISLA framework recognizes changes in phonemes (e.g., changing from /H/ phoneme to /EH/ phoneme), then it segments each of the three lower face regions.

To estimate emotion from the mouth region, it is important to differentiate emotion-



ISLA Labels

[1] Speech Production

ISLA Label	Basis (Lexicon)	ISLA Label	Basis (Lexicon)
1	/P/, /B/, /M/	8	/AE/, /AW/, /EH/, /EY/
2	/F/, /V/	9	/AH/, /AX/, /AY/
3	/T/, /D/, /S/, /Z/, /TH/, /DH/	10	/AA/
4	/W/, /R/	11	/AXR/, /ER/
5	/CH/, /SH/, /ZH/	12	/AO/, /OY/, /OW/
6	/K/, /G/, /N/, /L/, /HH/, /NG/, /Y/	13	/UH/, /UW/
7	/IY/, /IH/, /IX	14	/SIL/

[2] Speech Emphasis

ISLA Label	Basis (Pitch)
1	Emphasized
2	Non-emphasized

Figure 6.4: *ISLA labels based on (1) speech production (left) and (2) speech emphasis (right). The ISLA labels based on speech production are chosen based on visually similar phoneme groups, as in previous work [113, 114, 142, 148]. We proposed to use pitch signals to assign the ISLA labels based on speech emphasis.*

specific movement from movement due to lexical production. We use phoneme changes as the ISLA signal for the mouth region. Similarly, the other two lower face regions, cheek and chin, are also tied to the lexical production. Thereby, we use the phoneme signals as the ISLA signal for the lower face region.

On the other hand, to estimate emotion from the eyebrow region, it is important to tease apart emotion-related movement from emphasis-related movement. We estimate the speech emphasis patterns and use this as the ISLA signal for the eyebrow region. Similarly, the other two upper face regions, forehead and upper eyebrow, also use the speech emphasis patterns as the ISLA signal.

6.2.2 Informed Labeling

In the informed labeling step, we use the characteristics of the ISLA signals to label each segment (ISLA labels) and train separate classifiers based on these labels (Figure 6.4). For speech emphasis, we use two ISLA labels: emphasized and non-emphasized. We describe how we obtain these labels from pitch information in details in Section 6.3.2. We force-align speech and transcript to obtain phonetic information. Previous visual prosody studies have found that there exist 14 visually similar phoneme groups, and facial emotion recognition systems have followed these 14 groups [113, 114, 142, 148]. Therefore, we use 14 ISLA labels, where each label is for visually similar phoneme groups. For instance, /P/, /B/, /M/ are visually similar and hence have the same ISLA label.

6.3 Methodology

In this section, we present details of our methods and experiments to test three research questions that we introduced. First of all, we describe the data and features that we use in Section 6.3.1. In Section 6.3.2, we present how we use the ISLA framework to tease apart effects of modulation from emphasis and emotion in the upper face region. Section 6.3.3 describes an ISLA framework for the lower face region that uses phoneme signals, obtained from forced alignment between speech and transcript, as the ISLA signals. The facial data was segmented using phoneme boundaries. Lastly, in Section 6.3.4, we discuss how we estimate emotion from speech and how we fuse the outputs of the ISLA framework and speech emotion estimates.

6.3.1 Data Pre-Processing

We use the IEMOCAP and SAVEE datasets described in Chapter 2. We preprocess the markers by first translating to make the nose tip as the origin and then rotating to take into account the head movement, as in Chapter 5. We also mean-normalize each marker position by making the mean of each dimension of marker positions per each subject to be the global mean over all subjects, in order to reduce subject variations in facial configurations [113, 114, 148]. We exclude segments less than seven frames (0.058 seconds) as in Chapter 5.

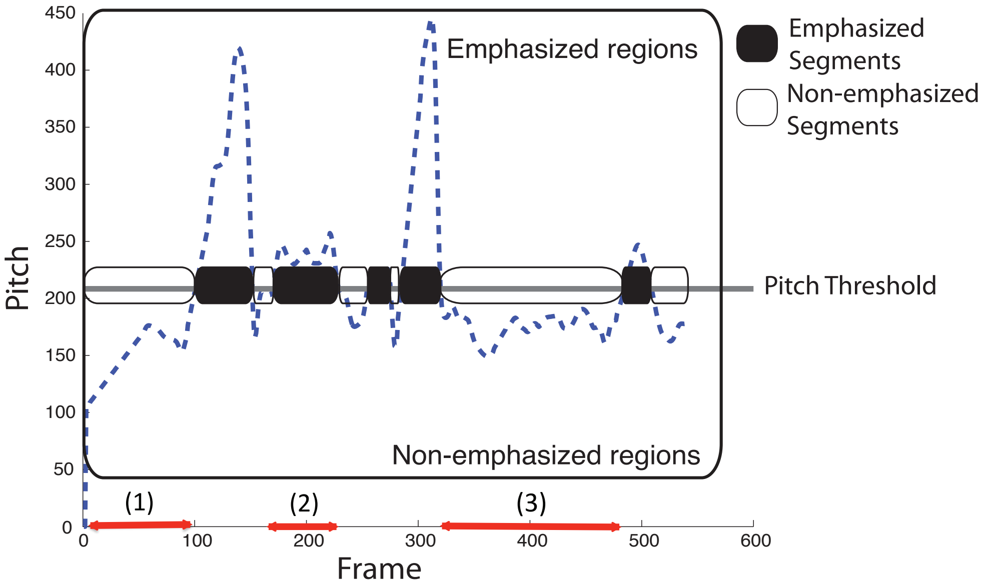


Figure 6.5: *An example pitch contour and estimated emphasized (top) and unemphasized (bottom) regions within an utterance, separated based on the mean pitch (red line), for the IEMOCAP data. Since we exclude segments shorter than 85 frames for the IEMOCAP dataset (indicated as black segments), the remaining segments are segments (1), (2), and (3). Segments (1) and (3) are labeled as non-emphasized segments, and segment (2) is labeled as emphasized segment, and hence these two sets of segments are separated into different classifiers for non-emphasized and emphasized segments, respectively.*

For the SAVEE dataset, as in Chapter 5, we interpolate the facial marker recordings by cubic spline interpolation to 120 frames per second, in order to be consistent with the IEMOCAP data.

6.3.2 Emphasis as the ISLA signal for the Upper Face

We hypothesize that emotion prediction from facial cues will improve when we tease apart the facial movement due to emotion and to emphasis, particularly for eyebrow regions. To estimate emphasis regions during speech, we use pitch signals from the audio modality as in previous work in visual prosody [108]. We first take a mean of pitch of all the spoken utterances for each speaker and then use this mean as a threshold to divide utterances into emphasized (region with pitch higher than

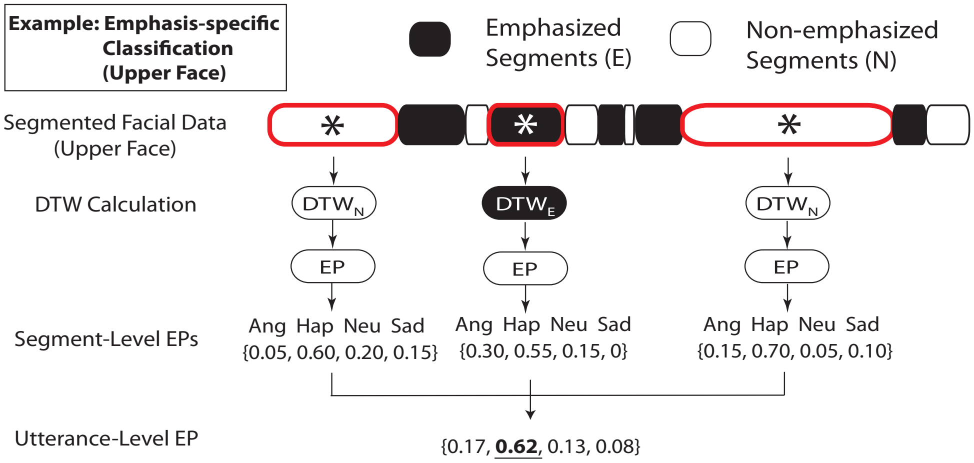

Example: Emphasis-specific Classification (Upper Face)
Emphasized Segments (E)
Non-emphasized Segments (N)
Segmented Facial Data (Upper Face)
DTW Calculation
DTW$_N$
DTW$_E$
DTW$_N$
EP
EP
EP
Segment-Level EPs
Ang Hap Neu Sad {0.05, 0.60, 0.20, 0.15}
Ang Hap Neu Sad {0.30, 0.55, 0.15, 0}
Ang Hap Neu Sad {0.15, 0.70, 0.05, 0.10}
Utterance-Level EP
{0.17, 0.62, 0.13, 0.08}

Figure 6.6: *An example of how utterance-level Emotion Profile (EP) can be obtained from the segment-level emotion estimates from the upper face region using speech emphasis, similar to Chapter 5. We segments the upper face movements based on the pitch threshold, and segments longer than 1 second are remained (noted with '*' and red highlights), as in Figure 6.5.*

mean) and un-emphasized (region with pitch lower than mean) segments.

In [108], the authors excluded segments less than 1.2 seconds in duration. In our experiments, the average segment length of the IEMOCAP dataset is 0.71 seconds and that of the SAVEE dataset is 0.50 seconds. The exclusion of segments shorter than 1.2 seconds results in significant data loss. Therefore, we removed segments shorter than the average length for each data, i.e., we removed segments shorter than 85 frames, or 0.71 seconds in duration, for the IEMOCAP dataset, and segments shorter than 60 frames, or 0.50 seconds in duration, for the SAVEE data.

Once we segment each utterance into emphasized and non-emphasized regions, we employ an emphasis-specific ('ES') classification strategy. This strategy compares the movement of emphasized test segments to emphasized training segments, and non-emphasized segments for non-emphasized regions. This classification strategy reduces emphasis-related variability in facial movements during emotion classification. Figure 6.5 shows an example pitch contour and estimated emphasis regions for the IEMOCAP data. The gray line shows the threshold of the mean of pitch during speech that separates the utterance into emphasized (black box) and non-emphasized (white box) regions. The boxes show the segmentation results based on this threshold. After the exclusion of segments shorter than 85 frames (0.71 seconds), the remaining segments are indicated as segments 1, 2, and 3 in the figure. The segments 1 and 3 are labeled as non-emphasized segments, and the segment 2 is labeled as emphasized segment. These ISLA labels will be used in classification.

6.3.3 Phoneme as the ISLA signal for the Lower Face

Lexical segmentation and labeling have been used in previous studies to improve facial emotion recognition during speech [113, 114, 142, 148]. This technique segments the facial data using phoneme boundaries. For instance, as shown in Figure 6.7, if a speaker is saying 'cheese', we segment facial motion capture data based on the start and end timing of each phoneme '/SIL/', '/CH/', '/IY/', '/Z/', and '/SIL/'. Phoneme labeling assigns each phoneme segment into visually similar groups of phoneme (Table 6.1). This phoneme assignment is used to group the data, which are used to train separate classifiers. This separation allows each emotion classifier to focus on emotion-specific patterns in the input data, by reducing phoneme-related variations in the data. The lexical segmentation and labeling achieves higher performance compared to fixed-length window segmentation, and this performance gain is mostly from the lower face regions.

Group	Phonemes	Group	Phonemes
V1	P, B, M	V8	AE, AW, EH, EY
V2	F,V	V9	AH,AX,AY
V3	T,D,S,Z,TH,DH	V10	AA
V4	W,R	V11	AXR,ER
V5	CH,SH,ZH	V12	AO,OY,OW
V6	K,G,N,L,HH,NG,Y	V13	UH,UW
V7	IY,IH, IX	V14	SIL

Table 6.1: *Visually similar phoneme groups.*

6.3.4 Audio-Visual Emotion Classification

We develop a speech-based emotion recognition system and propose an audio-visual classification framework that combines the ISLA face-based with speech-based emotion estimates (Figure 6.2). Section 6.3.4.1 describes how we estimate emotion from the upper face, lower face, and speech signals. Section 6.3.4.2 explains the overall audio-visual emotion classification.

6.3.4.1 Emotion Estimation Using Emotion Profiles

Emotion Profiles (EPs) are multi-dimensional vector representations that describe the level of confidence on the presence and absence of each type of emotion [158]. This EP description has been shown to be effective in both representation and classification of emotion [7, 42, 112, 155]. In the presented study, we use the EPs to describe emotion estimates from the upper face, lower face, and speech modalities.

We use ISLA to estimate emotion from both the upper and lower face regions. For each test segment, we calculate four-dimensional emotion estimates, EPs, using a method proposed in Chapter 5.

Dynamic Time Warping (DTW) distances are calculated between test and train segments, for both lower and upper face regions (Figure 6.6), as in [113, 114]. DTW is a time-series similarity measure that aligns two time series to minimize their distance. Unlike Euclidean distance measure, DTW can measure similarity of two time series data with different lengths. For instance, if we have two facial movement trajectories of length M_1 and M_2 with the same feature dimension K, i.e., $T_1 \in \mathbb{R}^{M_1 \times K}$ and $T_2 \in \mathbb{R}^{M_2 \times K}$, we compute a M_1-by-M_2 local cost matrix Q as follows:

$$Q(i,j) = \sum_{k=1}^{K} (T_1(i,k) - T_2(j,k))^2, \tag{6.1}$$

where i and j denote the time points of T_1 and T_2, respectively. We compare DTW

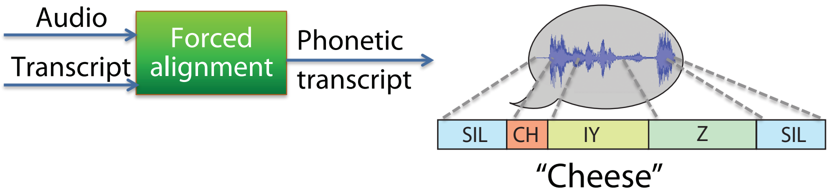


Figure 6.7: *Example of lexical segmentation and labeling for a word 'cheese'. Lexical segmentation segments the facial data using phoneme boundaries, obtained by forced alignment between audio and transcript.*

distances between testing segment and training segments, and calculate the emotion-class distribution over the k-closest neighbors (in chapter, $k = 20$ as in [113, 114]). For instance, if DTW distances of a test segment with $k = 20$ closest training segments have emotion labels with 1 angry, 12 happy, 4 neutral, and 3 sad classes, then we assign a profile of this distribution to the test segment as follows: $\{0.05, 0.60, 0.20, 0.15\}$.

We finally aggregate the segment-level profiles by taking an average of the segments to obtain the utterance-level profiles (Figure 6.6) as in Chapter 5. These profiles will be used as the input to our proposed classification systems.

We estimate the emotion content of the speech using the EP technique. We first extract Interspeech 2013 Paralinguistic features [213], 6,373 features in total. The features are based on 4 energy related low-level descriptors (LLDs), such as loudness, RMS energy, and zero-crossing rate; 55 spectral LLDs, such as MFCC, spectral energy, and spectral variance; and 6 voicing related LLDs, such as F0, probability of voice, log harmonic-to-noise ratio (HNR), jitter, and shimmer. The features are extracted using the openSMILE toolkit [64]. There exists more utterances that contain audio-only data than motion-captured data for the IEMOCAP. We we use all the 6,332 audio-only utterances in this dataset to train the speech emotion estimation system.

We train four binary SVMs: angry vs. not angry, happy vs. not happy, neutral vs. not neutral, and sad vs. not sad. We use the distance from hyperplane to estimate confidence, as in [158]. We convert these distances to probabilistic estimates, by first z-normalizing the outputs over a given speaker and using a sigmoid function to map the distances into values between 0 and 1.

6.3.4.2 Combining Audio and Visual Emotion Estimation

We combine speech-based emotion estimates with emotion estimates from the lower and upper face regions, by first taking the average of utterance-level emotion estimates over each modality. We then take the maximum component of the 4-dimensional averaged emotion estimates. For instance, if we obtain emotion estimates of a test utterance as {0.10, 0.60, 0.20, 0.00} from speech, {0.15, 0.55, 0.23, 0.07} from the lower face region, and {0.03, 0.81, 0.15, 0.01} from the upper face region, we first take the average of these modalities and obtain {0.09, 0.65, 0.19, 0.06}. Since the happy emotion component shows the highest emotion component (0.65), we finally infer the emotion label of this test utterance as happy emotion class.

6.3.5 Classification Setup and Baseline

Our subject-independent emotion classification systems hold all the data from a speaker for testing, and train the systems based on held-out training speakers. For the IEMOCAP dataset, it results in 10 different test sets, and we measure the performance of classification system by taking an unweighted average recall over the ten test performance. For the SAVEE dataset, we have 4 different test sets, and as in the IEMOCAP dataset, we measure the performance using unweighted recall over the four test performance. Also, since the SAVEE dataset is read speech, we conduct sentence-independent classification as well to remove the effect of same lexical information in our emotion recognition. As in Chapter 5, we use a paired t-test proposed in [52] to test the significance of result comparisons, and claim the significance when p-value is less than 0.05. Due to the limited number of speakers, we do not conduct the significance test for the SAVEE dataset. However, the SAVEE dataset provides useful insights into how posed emotion expressions differ from the expressions during conversations.

Our baseline methods assume that fixed-length windows can capture emotionally salient dynamics and that training a single classifier over all data can be effective in emotion classification. The fixed-length segment length is chosen as the average of emphasis and phoneme segments over all ten speakers. For the IEMOCAP dataset, the average segment length is 0.47 seconds, and for the SAVEE dataset, it is 0.34 seconds. We call the traditional classification method that uses a single classifier regardless of segment characteristics as 'general classification' method. These baseline models were also used for comparison in our prior work [113, 114].

6.4 Results and Discussion

6.4.1 Modeling Individual Modalities

In this section, we discuss the classification results of individual modalities. In particular, we compare different segmentation and classification methods for upper and lower face regions, based on our proposed ISLA framework. We also discuss the classification results of the speech signal as well.

6.4.1.1 Face

Tables 6.2 and 6.3 demonstrate the results of each facial region when modeled using three different methods, for IEMOCAP and SAVEE datasets, respectively. We compare our proposed emphasis based segmentation with emphasis-specific classification ('Em/ES') with three traditional methods: emphasis-based segmentation with general classification ('Em/Gen'), fixed-length segmentation with general classification ('Fixed/Gen'), and phoneme-based segmentation with phoneme-specific classification ('Phon/PS'). Fixed/Gen is a traditional method to simply develop a dynamic

Face Region	Face	Method	UW (%)
Upper Face	Forehead	**Em/ES**	**41.84** [*]
		Em/Gen	36.76
		Fixed/Gen	41.42
		Phon/PS	40.63
	Upper Eyebrow	**Em/ES**	**47.30** [*, ◊]
		Em/Gen	41.71
		Fixed/Gen	43.94
		Phon/PS	45.89
	Eyebrow	**Em/ES**	**45.31** [*, ◊, △]
		Em/Gen	38.65
		Fixed/Gen	40.15
		Phon/PS	40.30
Lower Face	Chin	Em/ES	46.51 [*]
		Em/Gen	40.92
		Fixed/Gen	50.20
		Phon/PS	50.10
	Cheek	Em/ES	42.15 [*]
		Em/Gen	35.72
		Fixed/Gen	44.17
		Phon/PS	45.67
	Mouth	Em/ES	46.90 [*]
		Em/Gen	41.26
		Fixed/Gen	51.14
		Phon/PS	57.03

Table 6.2: *IEMOCAP dataset Result Comparisons between our proposed emphasis based segmentation with emphasis-specific classification (‘Em/ES’) and traditional methods, including emphasis-based segmentation with general classification (‘Em/Gen’), fixed-length segmentation with general classification (‘Fixed/Gen’), and phoneme-based segmentation with phoneme-specific classification (‘Phon/PS’). ‘[*]’ indicates statistical significance ($p < 0.05$) between our proposed Em/ES method and Em/Gen. ‘[◊]’ indicates statistical significance ($p < 0.05$) between our proposed Em/ES method and Fixed/Gen. ‘[△]’ indicates statistical significance ($p < 0.05$) between our proposed Em/ES method and Phon/PS. The **bolded** numbers represent highest accuracy in the upper face region.*

Face Region	Face	Method	UW (%)
Upper Face	Forehead	**Em/ES**	**63.49**
		Em/Gen	57.03
		Fixed/Gen	59.79
		Phon/PS	62.08
	Upper Eyebrow	**Em/ES**	**59.88**
		Em/Gen	54.91
		Fixed/Gen	55.57
		Phon/PS	57.50
	Eyebrow	**Em/ES**	**83.96**
		Em/Gen	64.17
		Fixed/Gen	65.21
		Phon/PS	73.13
Lower Face	Chin	Em/ES	36.96
		Em/Gen	38.07
		Fixed/Gen	41.19
		Phon/PS	32.08
	Cheek	Em/ES	64.40
		Em/Gen	61.46
		Fixed/Gen	65.69
		Phon/PS	61.67
	Mouth	Em/ES	63.48
		Em/Gen	59.29
		Fixed/Gen	64.94
		Phon/PS	62.71

Table 6.3: *SAVEE dataset Result Comparisons between our proposed emphasis based segmentation with emphasis-specific classification ('Em/ES') and traditional methods, including emphasis-based segmentation with general classification ('Em/Gen'), fixed-length segmentation with general classification ('Fixed/Gen'), and phoneme-based segmentation with phoneme-specific classification ('Phon/PS'). The* ***bolded*** *numbers represent highest accuracy in the upper face region.*

classifier, and Phon/PS has been shown to be effective particularly for the lower face region in previous work [113, 114, 142, 148].

For the IEMOCAP dataset, our proposed Em/ES method outperforms all the three traditional methods for the upper face region: forehead, upper eyebrow, and eyebrow. The Em/ES method does not benefit the lower face region, which is not as strongly attached to emphasis as in the upper face region. For the emphasized segments (EM), ES classification significantly outperforms Gen classification for all face regions, regardless of upper or lower face regions for the IEMOCAP data. The forehead region achieved 41.84% when using the Em/ES, 5.08% higher than the Em/Gen. The upper eyebrow and eyebrow regions also achieve significantly higher accuracy using the Em/ES, achieving 47.30% and 45.31%, respectively. The lower face regions, chin, cheek, and mouth, also achieve significantly higher accuracy using the Em/ES than Em/Gen, achieving 46.51%, 42.15%, and 46.90%, respectively.

Our proposed Em/ES method also outperforms both Fixed/Gen and Phon/PS for the upper face region. The performance gain from the Fixed/Gen to Em/ES is 0.41% (not significant) for the forehead, 3.36% (significant) for the upper eyebrow, and 5.16% (significant) for the eyebrow regions. The results indicate that we can achieve a greater performance gain for the regions closer to the eyebrow region. However, for the lower face region, Em/ES achieves lower accuracy than Fixed/Gen, indicating that the use of emphasis signal does not contribute to emotion estimation due to its weak relationship with the lower face region. The performance gain from the Phon/PS to Em/ES is 1.19% for the forehead (not significant), 1.42% for the upper eyebrow (not significant), and 5.02% for the eyebrow (significant) regions. As in the performance gain from the Fixed/Gen to Em/ES, the improvement is the most significant in the eyebrow region. Also, Em/ES shows lower accuracy than Phon/PS for the lower face regions.

For the SAVEE dataset, Em/ES outperforms Em/Gen in forehead, upper eyebrow, eyebrow, cheek, and mouth regions– all the regions except for the chin, which is the farthest region from eyebrow. All the upper face regions demonstrate the increased performance of 63.49%, 59.88%, and 83.96%, for forehead, upper eyebrow, and eyebrow regions, respectively, when the Em/ES is used instead of the Em/Gen. The cheek and mouth regions also show the increase in performance of 64.40% and 63.48%, respectively. However, the Em/Gen outperforms the Em/ES for the chin region (38.67% and 36.96%, respectively), and this may indicate that the Em/ES benefits for the region that is highly associated with emphasis and eyebrow muscles (the chin is the farthest region from the eyebrow).

In addition, the SAVEE dataset consistently shows higher accuracy using the Em/ES method compared to any of the three traditional methods in the upper face regions. The Em/ES achieves 63.49%, 59.88%, and 83.96% for the forehead, upper eyebrow, and eyebrow regions, while the best results among the three traditional methods achieves 62.08% (Phon/PS), 57.50% (Phon/PS), and 73.13% (Phon/PS), respectively. For the lower face region, Em/Gen achieves the highest (41.19%), higher than the Em/ES method in the chin region, Fixed/Gen achieves the highest (65.69%) in the cheek region, and Fixed/Gen achieves the highest (64.94%) in the mouth region. Note that the SAVEE dataset has a large standard deviation ($> 5\%$) and previous work [114] found that the outperformance of Fixed/Gen is not significant. The lower performance of Em/ES in the lower face region is consistent as in the IEMOCAP data.

6.4.1.2 Speech

Table 6.4 shows the results when only audio modality is used. We use the speech profile described in Section 6.3.4.1, and make a final emotion inference based on the maximum component of the four emotion profile outputs.

The results demonstrate that audio-only emotion classification achieves higher

Data	UW	A	H	N	S
IEMOCAP	63.51	78.19	50.88	50.29	74.67
SAVEE	80.75	72.98	83.33	75.00	91.67

Table 6.4: *Audio-only classification results (max EP method) for the IEMOCAP and SAVEE datasets. Unweighted (UW) and per-class ('A' for angry, 'H' for happy, 'N' for neutral, and 'S' for sad emotion classes) accuracy.*

	Method	All	Lower Face	Upper Face	Audio	Audio+L.F.	Audio+U.F.	L.F.+U.F.
IEMOCAP	ISLA	61.54	54.64	48.55*	63.51	**67.22***	58.62	54.07
	Baseline	62.56	53.93	45.58	(same)	63.51	57.85	55.44
SAVEE	ISLA	84.38	63.45	78.28	80.75	74.32	**86.01**	81.61
	Baseline	83.09	64.43	64.59	(same)	76.37	74.17	74.76

Table 6.5: *ISLA and baseline (fixed-length segmentation and general classification) results for the IEMOCAP and SAVEE datasets: using all modalities (audio+lower face region+upper face region), audio and lower face region, and audio and upper face region (averaged over all speakers: 10 for the IEMOCAP and 4 for the SAVEE datas). For the IEMOCAP dataset, '*' indicates statistical significance ($p > 0.05$) between our proposed ISLA and the baseline methods. The **bolded** numbers represent highest accuracy achieved in each of the IEMOCAP and SAVEE datasets.*

accuracy than face-only emotion classification for both the IEMOCAP and SAVEE datasets. In particular, the results on the IEMOCAP dataset achieves 63.51% unweighted accuracy and results on the SAVEE dataset achieves 80.75% unweighted accuracy. The highest per-class accuracy of the IEMOCAP dataset is with the angry emotion class, achieving 78.19%, whereas the SAVEE dataset achieves the lowest per-class accuracy with the angry emotion class, achieving 72.98%. The SAVEE dataset achieves its highest per-class accuracy with the sad emotion class, achieving 91.67%. The facial emotion recognition achieves the highest per-class accuracy on the happiness classification for the IEMOCAP data.

6.4.2 Audio-Visual Classification

Table 6.5 shows the result comparisons between our proposed ISLA framework and the baseline for the IEMOCAP and SAVEE datasets. The table summarizes the results when different combinations of lower face, upper face, and audio signals are used in classification. For both of the datasets, our proposed framework outperforms the baseline, validating the efficacy of the ISLA framework for audio-visual emotion classification.

First of all, for the ISLA framework, the results demonstrate that the IEMOCAP and SAVEE datasets show different performance trends for different modalities. For the IEMOCAP dataset, the combination of speech and lower face modalities achieves the highest unweighted accuracy of 67.22%, whereas for the SAVEE dataset, the combination of speech and upper face modalities achieves the highest up to 86.01%. For the IEMOCAP dataset, the lower face region achieves higher emotion recognition accuracy than the upper face region, achieving 54.64% and 48.55%, respectively. On the other hand, for the SAVEE dataset, the upper face region achieves higher emotion recognition accuracy than the lower face region, achieving 78.28% and 63.45%, respectively. Also, the combination of the lower and upper face regions using the ISLA method achieves lower accuracy than the baseline for the IEMOCAP dataset, however, the same combination achieves higher accuracy using the ISLA method for the SAVEE data. Both of the IEMOCAP and SAVEE datasets show higher accuracy when speech is used, compared to cases when an individual modality of lower face or upper face is used.

Next, for the baseline framework, the highest accuracy is achieved in the combination of speech and the lower face region for the IEMOCAP dataset (63.51%), whereas the highest accuracy of 83.09% is achieved in the combination of all modalities, speech, the lower face region, and the upper face region, for the SAVEE data.

Comparing the best results of the ISLA and baseline frameworks across different combination of modalities, our proposed framework outperforms the baseline for both the IEMOCAP and SAVEE datasets, 3.71% (significant) and 2.92%, respectively. IEMOCAP achieves the highest accuracy when combining speech with the lower face region for both the ISLA and baseline frameworks. The proposed ISLA framework outperforms the baseline by 9.13%. The SAVEE dataset achieves highest accuracy when combining the speech with the upper face region for the ISLA framework, and when combining all the modalities for the baseline framework. As in the IEMOCAP dataset, the highest accuracy achieved in the ISLA framework (86.01%) is higher than that achieved in the baseline (83.09%), achieving a 2.92% performance gain.

6.4.3 Correlation Analysis Between Modalities

In this section, we explore how emotion estimates from the lower face, upper face, and speech modalities correlate with each other. To this aim, we use emotion components of EPs to examine the patterns of emotion estimates from the three modalities. When components are correlated across different modalities, for instance, happiness is expressed both in the upper and lower face regions, this suggests that the modalities express similar emotion expressions. If the components are not correlated, this indicates that the emotions expressed across different modalities are different.

The findings can provide valuable insights based on the difference between the IEMOCAP and SAVEE datasets. Posed smiles, such as smiles in the SAVEE dataset, have long been studied with their distinct characteristics of expressions from genuine smiles. In his earlier work, Duchenne found that posed smiles involve only a contraction of the mouth region, whereas genuine smiles, or *Duchenne smiles*, involve movements of both the eye and mouth regions [56, 61]. Recent studies in psychology also support this finding that this combination of muscle contractions in the eye and mouth regions uniquely associates with the positive emotion [147].

Tables 6.6 and 6.7 show the correlation for each emotion component of emotion profiles (‘ang’, ‘hap’, ‘neu’, and ‘sad’ component in each column), between pairs of modalities. This correlation is an average correlation over ten speakers. Each row of the tables show the pairs between (i) lower face and speech (‘LowFace-Audio’), (ii) upper face and speech (‘UpFace-Audio’), (iii) upper face and lower face (‘UpFace-LowFace’), and (iv) upper face, lower face, and speech (‘Up-Low-Audio’). For instance, consider the emotion profiles over 250 utterances from the lower face and speech. We calculate the correlation between the same emotion component, for instance, between $\{a_{1,low}, a_{2,low}, \ldots, a_{250,low}\}$ and $\{a_{1,aud}, a_{2,aud}, \ldots, a_{250,aud}\}$ for anger component. We do the same thing for happiness, neutrality, and sadness components.

In addition, we also investigate correlation between emotion profiles, when only emphasized or non-emphasized segments are aggregated to compute the profiles. The ‘Type’ column of Tables 6.6 and 6.7 show the correlation results when we aggregate over (i) all segments (‘All’), (ii) emphasized segments (‘Emph’), and (iii) non-emphasized segments (‘Non-Em’), to attain utterance-level profiles.

We use the Concordance Correlation Coefficient (CCC) introduced in [129] to analyze correlation between each component. The CCC measures the level of agreement on the profiles, or emotion estimates, obtained by two modalities. Given the Pearson correlation coefficient (σ) and the mean square error, the CCC combines these two measures as follows:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}, \tag{6.2}$$

where σ_x^2 and σ_y^2 are the variance of profiles from each modality, and μ_x and μ_y are the mean profiles of each modality. The CCC has been used in continuous emotion recognition [192].

Type	Pair	ang	hap	neu	sad
All	LowFace-Audio	0.50	0.49	0.21	0.42
	UpFace-Audio	0.37	0.43	0.25	0.27
	UpFace-LowFace	0.32	**0.76**	0.34	0.43
	Up-Low-Audio	0.39	0.56	0.27	0.37
Emph	LowFace-Audio	0.57	0.53	0.22	0.44
	UpFace-Audio	0.31	0.50	0.17	0.23
	UpFace-LowFace	0.35	0.75	0.23	0.39
	Up-Low-Audio	0.41	0.59	0.21	0.35
Non-Em	LowFace-Audio	0.46	0.47	0.17	0.41
	UpFace-Audio	0.14	0.40	0.08	0.24
	UpFace-LowFace	0.12	0.74	0.24	0.36
	Up-Low-Audio	0.24	0.54	0.17	0.33

Table 6.6: *Correlation analysis between utterance-level profiles, aggregated over different types of segments: (i) all segments ('All'), (ii) emphasized segments ('Emph'), and (iii) non-emphasized segments ('Non-Em'), in the IEMOCAP dataset (averaged correlation over all ten speakers)*

6.4.3.1 Duchenne Smiles and Naturalness of Data

The IEMOCAP dataset shows higher correlation between the upper and lower face regions for the happy emotion estimates (0.76), compared to the correlation of the SAVEE dataset (0.40). Our findings from the IEMOCAP and SAVEE datasets support the findings on Duchenne smile, by showing higher correlation in the lower and upper face regions for the IEMOCAP than the posed SAVEE data.

When we only consider emphasized segments, both the IEMOCAP and SAVEE datasets show high correlation between the lower and upper face regions, 0.75 and 0.77, respectively. However, when we consider non-emphasized segments, IEMOCAP still shows high correlation (0.74) but SAVEE does not show correlation (0.17). Considering that emphasis is another modulation source of facial movements, this may indicate that without other sources of modulation, the difference between natural and posed expressions become stronger.

Type	Pair	ang	hap	neu	sad
All	LowFace-Audio	0.63	**0.70**	0.57	0.35
	UpFace-Audio	0.67	0.27	0.52	0.59
	UpFace-LowFace	0.66	0.40	0.30	0.35
	Up-Low-Audio	0.65	0.46	0.46	0.43
Emph	LowFace-Audio	0.62	0.79	0.63	0.59
	UpFace-Audio	0.53	0.68	0.32	0.52
	UpFace-LowFace	0.52	0.77	0.08	0.57
	Up-Low-Audio	0.56	0.75	0.34	0.56
Non-Em	LowFace-Audio	0.63	0.71	0.57	0.35
	UpFace-Audio	0.63	0.06	0.30	0.60
	UpFace-LowFace	0.60	0.17	0.13	0.35
	Up-Low-Audio	0.62	0.31	0.33	0.43

Table 6.7: *Correlation analysis between utterance-level profiles, aggregated over different types of segments: (i) all segments ('All'), (ii) emphasized segments ('Emph'), and (iii) non-emphasized segments ('Non-Em'), in the SAVEE dataset (averaged correlation over all four speakers)*

6.4.3.2 Correlation for Emphasized vs. Non-Emphasized Segments

The second and third rows ('Emph' and 'Non-Em') of Tables 6.6 and 6.7 compare how emphasized and non-emphasized segments correlate with each other. We also show that in general, the emphasized segments show similar or higher correlation than the non-emphasized segments for both the IEMOCAP and SAVEE datasets, except for few cases when both of the emphasized and non-emphasized segments reveal no correlation. Emphasis is another source of facial modulation and can be confused with facial movements for conveying emotion. Therefore, the higher correlation of emotion estimates for the emphasized segments may imply that humans try to strengthen emotional messages with the co-production of emphasis.

6.4.3.3 Emotional Expressions During Conversation

Previous psychological studies found that humans recognize happiness the easiest [60]. We show that happy emotion estimate achieve the highest correlation for each pair of modalities, which may indicate that the more correlated modalities are, the easier for humans to recognize the underlying emotion.

Angry emotion class shows consistent correlation within each of the IEMOCAP and SAVEE datasets across different pairs of signals. For the IEMOCAP dataset, the correlation was consistently 0.33-0.34. For the SAVEE dataset, it was consistently 0.65-0.68. This consistency is not observed in other emotion classes of happiness, neutrality, and sadness. The higher correlation of the SAVEE datasets between different signals may indicate that the expression of anger in read speech of an individual coordinates more between the lower face, upper face, and speech signals, compared to more spontaneous emotion expression during two-person conversations.

For the IEMOCAP dataset, the results demonstrate that the happy emotion class has the highest correlation between the upper and lower face profiles, achieving 0.76. We also found that the correlation between speech and lower face region is higher than the correlation between speech and upper face region, for anger, happiness, and sadness. The neutral emotion class shows relatively lower correlation than other emotion classes, partly due to the ambiguous nature of the neutral expressions.

For the SAVEE dataset, the results also show that the happy emotion class has the highest correlation, however the difference from IEMOCAP is that the correlation between lower face and audio profiles achieves the highest at 0.71. The happy and neutral emotion classes shows higher correlation between lower face and speech signals, compared to the correlation between the upper face and speech signals. The second highest correlation was in the sad emotion class, where the correlation between the upper face and speech signals achieves 0.60.

In addition to correlation of happiness, the correlation of anger is higher for the SAVEE dataset than that of the IEMOCAP dataset, consistently across pairs of lower face-audio (IEMOCAP: 0.34, SAVEE: 0.65), upper face-audio (IEMOCAP: 0.34, SAVEE: 0.68), upper face-lower face (IEMOCAP: 0.33, SAVEE: 0.66), and upper face-lower face-audio (IEMOCAP: 0.33, SAVEE: 0.66) signals. This correlation difference may be due to the nature of the two datasets: interaction setting of the IEMOCAP and individual read speech setting of the SAVEE dataset, where human subjects may regulate the expression of anger and tend to not show anger explicitly during social interactions more than individual posed expression [78].

The posed expressions of the SAVEE may also result in higher correlation in modalities. The highest correlations of angry, neutral, and sad emotion classes are higher for the SAVEE dataset than the IEMOCAP data. The correlation analysis of lower face, upper face, and speech signals for different emotion classes demonstrate how emotion estimates from different modalities associate with each other.

6.4.3.4 Correlation for Each Emotion Label

We explore whether the emotion class of utterances, when using all segments, emphasized segments, and non-emphasized segments, reveal different correlation trends (Table 6.8). For the IEMOCAP dataset, happy utterances have high correlation for the happy component in the upper and lower face regions, achieving 0.65 and 0.64, respectively. On the other hand, there is less correlation between these pairs for the angry and sad utterances. This correlation difference between happy utterances and angry/sad utterances indicates that the estimated presence of happiness is not correlated when the underlying emotion is negative or has low valence, i.e., anger and sadness. All of the angry, happy, neutral, and sad utterances did not show high correlation (< 0.5) for the SAVEE data, and hence not reported in this section.

Also, no matter what the ground truth emotion labels are, the lower and speech modalities show moderate to high correlation for the happy component (0.41-0.67). This means that when the presence of happiness is weak in the lower face, it is the same for the speech, no matter what emotion is underlying. This also means that when the presence of happiness is strong in the lower face, it is same for the speech.

Type	Pair	ang	hap	neu	sad
Ang	LowFace-Audio	0.30	0.10	0.18	-0.09
	UpFace-Audio	0.26	0.12	0.24	-0.02
	UpFace-LowFace	0.35	0.42	0.27	0.22
	Up-Low-Audio	0.30	0.21	0.23	0.04
Hap	LowFace-Audio	0.19	0.32	0.20	0.24
	UpFace-Audio	0.23	0.29	0.24	0.14
	UpFace-LowFace	0.21	**0.67**	0.32	0.40
	Up-Low-Audio	0.21	0.43	0.25	0.26
Neu	LowFace-Audio	0.24	0.18	0.06	0.19
	UpFace-Audio	0.14	0.13	0.07	0.04
	UpFace-LowFace	0.04	0.67	0.16	0.30
	Up-Low-Audio	0.14	0.33	0.09	0.18
Sad	LowFace-Audio	0.09	0.20	0.05	0.16
	UpFace-Audio	0.08	0.10	0.10	0.00
	UpFace-LowFace	0.09	0.41	0.19	0.09
	Up-Low-Audio	0.08	0.24	0.12	0.09
Em-Ang	LowFace-Audio	0.17	0.08	0.02	-0.10
	UpFace-Audio	0.14	0.19	0.07	0.06
	UpFace-LowFace	0.13	0.20	0.10	0.08
	Up-Low-Audio	0.15	0.16	0.06	0.01
Em-Hap	LowFace-Audio	0.21	0.34	0.26	0.27
	UpFace-Audio	0.05	0.33	0.21	0.14
	UpFace-LowFace	0.29	**0.65**	0.17	0.38
	Up-Low-Audio	0.18	0.44	0.21	0.26
Em-Neu	LowFace-Audio	0.42	0.21	-0.01	0.29
	UpFace-Audio	0.18	0.04	-0.01	-0.07
	UpFace-LowFace	0.15	**0.57**	0.10	0.24
	Up-Low-Audio	0.25	0.27	0.03	0.15
Em-Sad	LowFace-Audio	0.03	0.17	0.04	0.12
	UpFace-Audio	-0.02	0.03	-0.05	-0.04
	UpFace-LowFace	0.00	0.25	0.11	0.14
	Up-Low-Audio	0.00	0.15	0.03	0.07
NEm-Ang	LowFace-Audio	0.28	0.10	0.18	-0.03
	UpFace-Audio	0.04	0.11	0.06	-0.01
	UpFace-LowFace	0.06	0.43	0.15	0.11
	Up-Low-Audio	0.12	0.22	0.13	0.02
NEm-Hap	LowFace-Audio	0.20	0.31	0.16	0.22
	UpFace-Audio	0.07	0.25	0.06	0.13
	UpFace-LowFace	0.16	**0.64**	0.27	0.37
	Up-Low-Audio	0.14	0.40	0.17	0.24
NEm-Neu	LowFace-Audio	0.23	0.20	0.06	0.20
	UpFace-Audio	0.10	0.13	0.02	0.07
	UpFace-LowFace	0.03	**0.64**	0.13	0.27
	Up-Low-Audio	0.12	0.32	0.07	0.18
NEm-Sad	LowFace-Audio	0.08	0.19	0.04	0.14
	UpFace-Audio	0.03	0.04	0.07	-0.02
	UpFace-LowFace	0.03	0.39	0.16	0.06
	Up-Low-Audio	0.05	0.21	0.09	0.06

Table 6.8: *Correlation analysis between utterance-level profiles, for utterances with each emotion label: anger ('Ang'), happiness ('Hap'), neutrality ('Neu'), and sadness ('Sad'). Also, profiles aggregated over different types of segments: (i) emphasized ('Em') and (ii) non-emphasized ('NEm') segments, in the IEMOCAP data.*

6.5 Conclusions

We present ISLA, a framework for automatic emotion recognition when a person is speaking. This framework considers two sources of facial modulations due to speech: speech emphasis and production. The segmentation and labeling steps of ISLA are informed by speech signals, hypothesized to alter the temporal characteristics of individual face regions. We also explore how to combine the outputs of the ISLA framework with emotion estimates from speech, for the design of audio-visual emotion recognition systems. We identify the relative contributions of the lower face, upper face, and speech modalities in emotion recognition, and design an audio-visual classification system that utilizes these relative contributions in emotion recognition.

We show that the upper face region, particularly the eyebrow region, is highly associated with emphasis signal, which is estimated by increased pitch from speech. We show that the proposed ISLA framework that utilizes the pitch signal as the ISLA signal, and segments and labels the upper face movements using this signal significantly outperforms the previous state-of-the-art [114]. This generalizes Chapter 5 where we demonstrate that the lower face region, particularly the mouth region, highly coordinates with the phoneme signal.

We further investigate how emotion estimates from the upper face, lower face, and speech modalities correlate with each other. The correlation analysis demonstrates that the expression of happiness is highly correlated between lower and upper face regions, regardless of the underlying emotion label. We also demonstrate that the correlation in anger shows consistency within both the IEMOCAP and SAVEE datasets, however the correlation was higher for the SAVEE data. We hypothesize that this may be due to the nature of IEMOCAP dataset, which is dyadic conversation between two people– people tend to not show anger overtly during interactions [224]. The experimental results using two emotion datasets, IEMOCAP and SAVEE, show the highest accuracy of 67.22% for the IEMOCAP and 86.01% for the SAVEE datasets, both outperform the baseline models (significantly for the IEMOCAP).

The findings of this chapter provide insight into how to effectively capture temporal characteristics of the upper and lower face movements during speech. The novelty of ISLA is that it offers a framework that includes informed segmentation and labeling to control for sources of modulation inherent in facial movements. In addition, the interactions and relative contributions of the lower face, upper face, and speech modalities in emotion classification inform the design of audio-visual emotion recognition systems.

6.6 Acknowledgement

The work is funded by IBM. The authors gratefully acknowledge the support.

Part III: Localization of Salient Events

CHAPTER 7

Emotion Spotting

7.1 Introduction

This chapter aims to discover consistent patterns in time regions of emotion evidence in the lower face, upper face, and speech modalities. Previous studies have found that humans require different amounts of information over time to accurately perceive emotion expressions. This varies as a function of emotion classes. For example, recognition of happiness requires longer stimuli than recognition of anger. We develop a data-driven system that captures emotion evidence at different timings and durations, for different emotion classes and different modalities. We use a combination of four binary emotion classifiers to estimate short-time emotion, and explore patterns (timings and durations) of emotion evidence. Our results demonstrate similar patterns for each emotion class across different subject-independent training folds of the IEMOCAP corpus. In addition, we show that the our proposed method that only uses a portion of the data (59%) can achieve comparable accuracy to a system that uses all of the data within each utterance. Our

The work presented in this chapter has been submitted in the following article: **Yelin Kim** and Emily Mower Provost. "Emotion Spotting: Discovering Regions of Evidence in Audio-Visual Emotion Expressions." ACM International Conference on Multimodal Interaction (ICMI). Tokyo, Japan. November, 2016. **Submitted**.

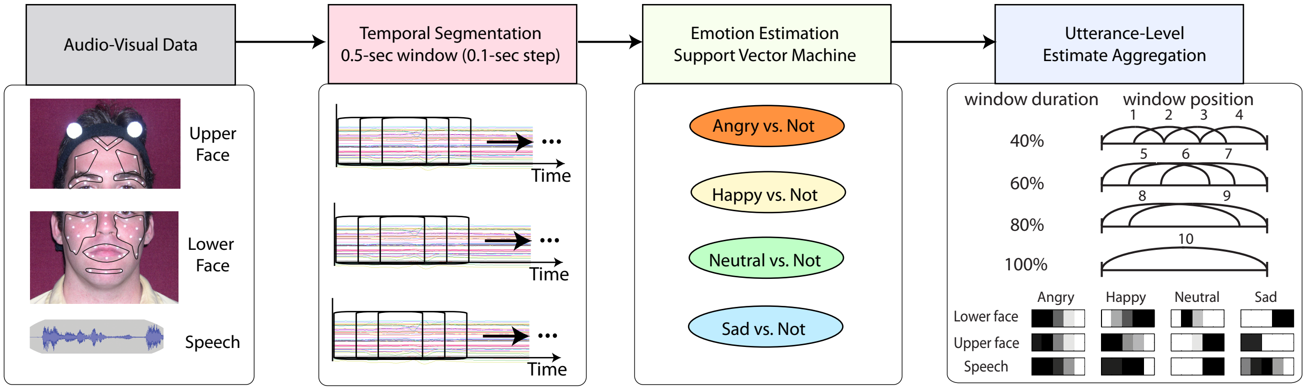


Figure 7.1: *Overview of the proposed classification method. We first segment lower face, upper face, and speech modalities using fixed-length windows and calculate segment-level emotion estimates using SVMs. We then aggregate the segment-level emotion estimates with different temporal window configurations (Index 1–10).*

data-driven method has a higher accuracy compared to a baseline method that randomly chooses a portion of the data. We show that the performance gain of the method is mostly from prototypical emotion expressions (defined as expressions with rater consensus). The novelty of our study is in its understanding of how multimodal cues reveal emotion over time.

7.2 Motivation

Audio-visual emotion recognition systems play a pivotal role in natural and human-centered interactive technology [47, 86, 98, 176, 184]. These systems use audio-visual inputs of users, such as their facial movements and vocal changes, to infer their emotions. For instance, Pepper, a personal robot with emotional capabilities, uses its camera sensors and microphones to gauge changes in facial and vocal expressions during interactions with users [1]. Studies on audio-visual emotion recognition have been growing rapidly within the field of multimodal interaction, often with a focus on how to combine the emotion information from audio and visual modalities. However, there has been less investigation of how these multiple modalities unfold emotion over time. Partial information may be sufficient for inferring human emotion [66, 81, 185]. For example, a smile may make interaction partners perceive happiness even though the person shows neutral expressions most of the time [97]. Likewise, sudden bursts of anger can be significant indicators of a person's angry emotion [66].

A basic assumption behind previous emotion recognition systems is that human emotions are expressed simultaneously with the same duration in multiple modalities [49, 112, 151, 164]. Previous systems have often overlooked the modality-specific temporal characteristics. A proper understanding of these characteristics may allow us to process only the relevant subsets of each modality, rather than all presented information. In this chapter, we explore regions within an utterance that contain emotion evidence, varying for the lower face, upper face, and speech modalities. We focus on timings and durations of these regions, which we call 'temporal patterns' throughout this chapter. We aim to investigate three important research questions.

The first research question pertains to generalizability and subject-independency in the temporal patterns of emotion. Previous studies have explored relationship between multiple modalities, however they either neglected temporal patterns [239] or generalizability across multiple human subjects [24]. Human perception studies found that there exist different durations required to correctly recognize emotion for individual emotion classes [180]. This indicates the need to answer the important research question, **(Q1)** Are there consistent temporal patterns of emotion expressions across subjects in the lower face, upper face, and speech modalities?

We evaluate the efficacy of these temporal patterns in audio-visual emotion recognition systems. We are interested in the following two research questions: **(Q2)** Can we achieve similar accuracy to the all-mean method, but using less data, and higher accuracy than the baseline method with randomly selected windows? and **(Q3)** What types of emotion expressions are associated with consistent emotion patterns?

In this chapter, we address the three research questions by proposing a data-driven approach to find consistent temporal patterns of emotion in the lower face, upper face, and speech modalities, varying for four emotion classes (anger, happiness, neutrality, sadness). Our approach identifies temporal regions within an utterance that lead to the highest emotion recognition rate, varying for different modalities and for different emotion classes. Figure 7.1 shows the overview of our system. Our method first segments the lower face, upper face, and speech modalities using fixed-length windows. We estimate the emotion content in each of the segments. We create sets of emotion evidence, defined as contiguous segments in time. Each set has a different position or timing within the utterance (e.g., the beginning vs. the middle) and duration (e.g., 40% vs. 60%). We classify the utterance based on the emotional evidence within each set. Our goal is to identify the optimal parameters (timing and duration). We compare our proposed methods to two baseline methods: the first baseline uses all the data within an utterance as in traditional systems ('all-mean method') and the second one uses partial data within an utterance, but these regions are randomly selected rather than data-driven.

The key novelty of this follow-up study is our investigation on the three research questions. The experimental results demonstrate that there exists consistent temporal patterns of the timing and duration. These temporal patterns show similarity over speakers within the same emotion class and modality. Our proposed system achieves similar accuracy to a traditional system that uses all the data, while using only 40–80% of the data for emotion inference. It also significantly outperforms the baseline method that uses random temporal regions within an utterance. The findings of our work provide insight into how lower face, upper face, and speech modalities reveal emotional evidence at different timings and time scales.

7.3 Proposed System

Our system is composed of four main modules: feature extraction, segment-level emotion estimation, window-based averaging, and final emotion classification at the utterance level. We first extract audio-visual features from the lower face, upper face, and speech modalities, using the features that have shown to be effective in previous emotion recognition studies [26, 158, 211, 221]. Next, we segment the audio-visual features into fixed-length windows and estimate segment-level emotions using Support Vector Machines (SVMs). We then apply various window configurations with different window timing and duration on the segment-level emotion estimates, to find the best window configuration for each modality and each emotion class. Finally, we use the emotion estimates that are aggregated by the identified windows to infer the utterance-level emotion class.

To train and test our proposed system, we employ leave-one-speaker-out cross-validation. Since the IEMOCAP data include ten speakers in total, we conduct ten-fold experiments. In each of the ten experiments, we use nine training speakers to train the system and a held-out speaker to test the emotion classification performance of the trained system. To choose the best-performing window configuration, we do leave-one-speaker-out cross-validation on each of the nine training speakers. This means that for each of the ten speakers, we compute the validation accuracy of nine training speakers, when eight speakers are used to train the system and a held-out training speaker is used for validation.

7.4 Feature Extraction

The feature utilized in this chapter are divided in to three modalities: (1) lower face, (2) upper face, and (3) speech.

The lower face includes three face regions (chin, mouth, and cheek) and the upper face includes forehead, eyebrow, and upper eyebrow. We extract the (x, y, z)-coordinates of the motion capture features to track the movements of the lower face and upper face. The origin is chosen as the nose tip, and the facial features are rotated to compensate for head rotation. We pre-process the markers by first translating to make the nose tip the origin and then rotating to compensate for the head movement. To reduce subject variations in facial configurations, we also mean-normalize each marker position by making the mean of each dimension of marker positions per each subject to be the global mean over all subjects, as in previous work [113, 114, 148]. We also exclude segments less than seven frames (0.058 seconds) due to insufficient temporal information, as in [113, 114].

The speech features contain spectral and prosodic features, that have been shown to be useful in emotion recognition [158]. This includes pitch and energy for prosodic features and mel-filterbank coefficients (MFBs) for spectral features, extracted using the Praat program[18]. This results in 29-dimensional speech features in total.

7.5 Segment-Level Emotion Estimation

7.5.1 Temporal Segmentation

Based on findings from previous work [114], we choose to use 0.5-second windows, moving with 0.1-second time steps. We use all the fixed-length windows and include segments at the end of an utterance that are shorter than 0.5 seconds. The segment-level features are computed using the mean, standard deviation, upper quantile, lower quantile, quantile range, and 3-degree polynomial regression coefficients within each segment. This results in 648 features for the lower face, 456 features for the upper face, and 232 features for the speech.

7.5.2 Emotion Estimation

We estimate segment-level emotion evidence using the Emotion Profile (EP) technique proposed by Mower et al. [155, 158]. We first train four binary emotion classifiers using the utterance-level data of the lower face, upper face, and speech. Each set of emotion classifiers consists of four binary classifiers for anger, happiness, neutrality, and sadness recognition. Each classifier is a radial basis function kernel Support Vector Machine (SVM), where the soft margin parameter c is chosen as 1 as in [28, 142]. We set the gamma in kernel function as a reverse of the number of input features, to be consistent with a default value suggested in [35].

For each test utterance, we use the segmented data of the utterance as an input to the trained emotion classifiers. The segment-level SVM outputs are used to compute EPs. As in previous work [155, 158], we use a distance from the SVM hyperplane as the confidence level of presence of each emotion component of EPs. Each component is a signed value, where a negative value means the absence of emotion (e.g., not angry), and a positive value means the presence of emotion (e.g., angry). Finally, we convert these SVM outputs into

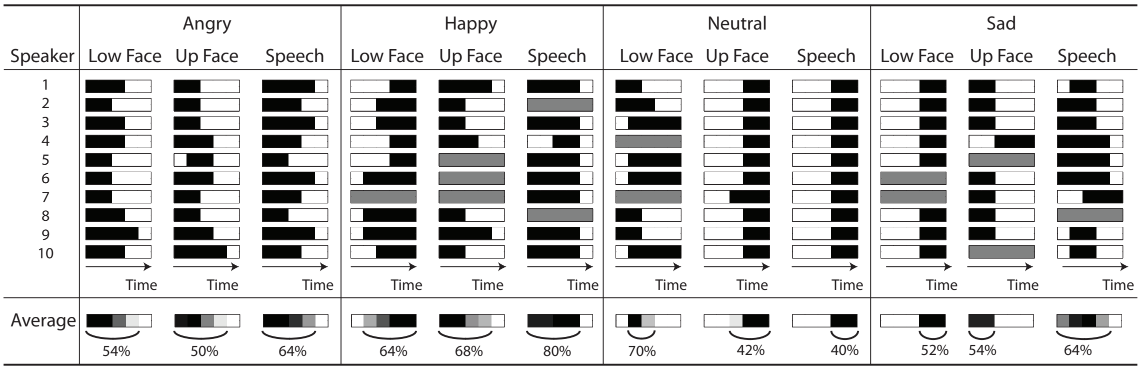


Figure 7.2: *Temporal window configurations for each speaker, for individual modalities (LowOpt: lower face, UpOpt: upper face, AudOpt: speech) and for each emotion component (Angry, Happy, Neutral, Sad). The last row is an averaged window configurations over ten speakers. For each speaker, black regions are the chosen regions used for emotion classification. The darker regions in the last row show overlapping windows from the ten speakers. We also show the average percentage of an utterance used over the ten speakers.*

probabilistic values, by first applying z-normalization and taking a sigmoid function. We then normalize the values of each emotion component so that the sum of the four emotion components becomes 1.

7.6 Window-Based Averaging

A traditional method to aggregate segment-level emotion estimates is to take the mean over all the segment outputs within an utterance [114]. Our proposed system instead take the mean of segment-level emotion estimates from a region within an utterance. We investigate which regions and durations of the EP from the upper face, lower face, and speech, are useful for final emotion inference. This allows us to explore timings and durations of emotion evidence within an utterance and to use smaller number of segments in emotion classification.

We explore two types of window configurations: timing and duration of windows. The last module of Figure 7.1 shows ten different configurations with different window durations and positions, each denoted as indices 1 to 10. We use cross-validation over each training fold to choose one of these ten window configurations. The indices are as follows:

- Index 1–4: we divide an utterance into four regions, each with 40% duration of an utterance
- Index 5–7: we divide an utterance into three regions, each with 60% duration of an utterance
- Index 8, 9 : we divide an utterance into two regions, each with 80% duration of an utterance
- Index 10 : 100% duration of an utterance (all data in an utterance)

We compute validation accuracy of per-angry, per-happy, per-neutral, and per-sad emotion classes, when one of the three modalities (lower face, upper face, and speech) is used in classification. For each emotion-modality pair (12 pairs in total for four emotion classes and three modalities), we choose the best window configuration over the training cross-validation.

7.7 Emotion Classification

We use the chosen window timing and duration to calculate utterance-level emotion estimates from segment-level emotion estimates. For each of the three modalities, lower face, upper face, and speech, we have four-dimensional emotion estimates of angry, happy, neutral, and sad classes. For each modality and emotion pair, we individually take the average of segment-level estimates within a region of an utterance based on the chosen window configuration. For instance, if for a test speaker the cross-validated parameters (Figure 7.1) are 5, 1, and 8 for angry classification using lower face, upper face, and speech, then we take a mean of the angry component using 60% of the beginning of an utterance, 40% of the beginning of an utterance, and 80% of the beginning of an utterance for the test speaker for the segment-level emotion estimates from the lower face, upper face, and speech modalities, respectively.

Once we apply different window configurations for each modality and each emotion component, we take an average over different modalities to get the four-dimensional EP at the utterance level. Each dimension of the EPs is the averaged emotion component of angry, happy, neutral, and sad emotion classes. As in [158], we choose the final emotion label that is a maximum component among the four emotion components. For instance, if the outputs of angry, happy, neutral, and sad binary classifiers are $[0.17, -0.63, -0.23, 0.80]$, then we choose sadness as the emotion label of the test data.

7.8 Results and Discussions

We design and perform experiments to address our three research questions (Q1-Q3).

To address Q1, we first investigate the chosen window timings and durations across ten training folds (Section 7.8.1). We show that there exist consistent patterns, varying for different emotion classes and individual modalities. We provide insight into how these temporal patterns match findings from human perceptual studies.

To address Q2, We evaluate performance of our proposed audio-visual emotion recognition systems by comparing with a baseline method. The baseline method selects temporal regions using a uniformly distributed randomization of the ten different window configurations. We randomly select the parameters for each emotion component and for each modality of the lower face, upper face, and speech. We run 50 runs and obtain the classification results of the 50 runs. We then take an average of these 50 runs for each speaker to compare our proposed method with the baseline.

We compare the performance of our system to a traditional system that utilizes all the information within an utterance, the all-mean method. This method takes a mean of the emotion evidence within an utterance to infer the final emotion.

For Q3, we compare the performance gain of our system based on the inter-rater agreement of each utterance. We divide the utterances into prototypical (defined as rater consensus) and non-prototypical (defined as no rater consensus, but the presence of majority vote) utterances.

The performance measure of our system is unweighted (UW) recall, to be consistent with previous work [142, 159]. We perform the paired t-test to test the significance of accuracy

	L.F+U.F+Aud	L.F	U.F	Aud	L.F+Aud	U.F+Aud	L.F+U.F
Proposed	65.60[*]	59.76[*]	52.85[*]	54.56	64.50	61.24	60.19[*]
all-mean	65.59	60.65	52.62	55.47	65.83	62.08	60.57
Baseline	63.70	57.92	51.21	53.41	62.79	60.02	58.78

Table 7.1: *Unweighted recall of unimodal and multimodal experiments for our proposed window method, the all-mean method, and the baseline method using randomized window configurations. The symbols "[*]" next to the accuracies indicate statistical significance levels ($p < 0.05$) between our proposed method and baseline. All the results between the proposed method and the all-mean method are statistically comparable to each other ($p > 0.05$).*

comparisons between our proposed method and the baseline and between our proposed method and the all-mean method, as described in Section 1.2.7.

7.8.1 Temporal Evidence Analysis

The chosen window configurations indicate that there is consistency across speakers with respect to the timing and duration of emotion. Figure 7.2 demonstrates our key findings on the timings and durations of windows. The first ten rows show the chosen window configurations from cross-validation over training speakers, for individual modalities, and for each emotion class, while the last row shows an average over the ten speakers. For each speaker, the chosen regions are represented as black, and for the averaged region, the darker regions represent overlapping regions from the ten speakers. The areas are consistent across multiple speakers for different modalities and emotion classes.

As shown in the table and figure, the lower face region is chosen consistently across the speakers at the beginning of an utterance for anger and at the end for happiness, and generally, at the end for sadness. The neutral class is a mixture of different window configurations. This finding is in line with historical difficulty in defining and classifying neutrality [139, 157, 190]. The upper face is generally chosen at the beginning of an utterance for anger, happiness, and sadness. Neutrality uses information at the end of an utterance. As shown in the table, speech requires longer durations to identify emotion classes of angry, happy, and sad. The three emotion classes are chosen at the beginning of an utterance

Method	UW	A	H	N	S	W
Proposed	**65.60**	72.88	72.02	40.96	76.53	66.38
All-mean	**65.59**	71.05	73.99	38.15	79.16	66.92
Baseline	**63.70**	68.49	73.10	36.91	76.29	64.87

Table 7.2: *IEMOCAP experimental results on the proposed, all-mean, and baseline methods. The accuracies are unweighted recall ('UW'), per-class accuracy for angry ('A'), happy ('H'), neutral ('N'), and sad ('S') emotion classes, and weighted accuracy ('W').*

in cross-validation. The neutrality requires 40–60% of information within an utterance for speech data, and the region is chosen at the end of an utterance.

The window configurations also reveal similar findings from previous studies. In a recent speech emotion recognition study [180], the authors found that angry, sad, fearful, and neutral emotion expressions are more accurately recognized given shorter data, compared to happy emotion expressions. As shown in Figure 7.2, the percentage of an utterance used for recognizing happiness is 80% for speech, which is higher than other emotion classes: 64% for anger, 40% for neutral, and 64% for sad emotion classes. This finding may indicate that our proposed data-driven window configurations can provide insight into how humans perceive emotion expressed over time.

7.8.2 Evaluation of Emotion Recognition

We compare the performance to our baseline, where windows are chosen at random. We also compare our results to a method that uses all the segment-level evidence within an utterance, instead of partial information, as in our proposed method. The results in this section will help us to answer Q2. Overall, our proposed method significantly outperforms the baseline method, and it achieves comparable accuracy to the all-mean method.

Table 7.1 shows the UW recall of unimodal and multimodal experiments for our proposed window method, all-mean method, and the baseline method using randomized window configurations. We test different combinations of modalities, and each column represents all modalities that combine the lower face (LF), the upper face (UF) and speech (Aud), the lower face, the upper face, the audio, the lower face with audio, the upper face with audio, and the lower and upper face. Our proposed window-mean method achieves 65.60% UW and the all-mean method achieves 65.59% UW. The difference is not significant. Our proposed method is significantly higher than the average UW accuracy of the randomized window method, achieving 65.60% vs. 63.70%, respectively (1.90% higher, $p < 0.05$), when all modalities are used.

Our proposed method also outperforms the baseline for all types of modality combinations. For the lower face, our proposed method achieves 1.84% higher UW accuracy, significantly improving the baseline ($p < 0.05$). The upper face region also significantly outperforms the baseline, achieving 1.63% higher UW accuracy ($p < 0.05$). Speech is also higher when we use our proposed method, but not significantly (1.14% improvement, $p = 0.06$). Speech with the lower face and speech with the upper face both achieve higher UW recall than the baseline, but this difference is not significant (1.71% improvement with $p = 0.06$ and 1.23% improvement with $p = 0.08$, respectively). The lower and upper face combination achieves significantly higher performance, showing 1.42% improvement with $p < 0.05$.

The bottom row of Figure 7.2 shows how much of an utterance is used for each emotion class, averaged over ten speakers. The results indicate that the used data of each utterance is only 40% to 80% of an utterance. This highlights the benefit of our system as it is capable of spotting a region within an utterance, and reasoning only over that region, while achieving comparable accuracy with an experiment that uses the full information of an utterance. The results also demonstrate that, on average, speech requires more regions of an utterance than lower and upper face regions for emotion classes, i.e., angry, happy, and sad, while the lower face is used more for neutral recognition.

Table 7.2 shows the comparison of emotion classification results between our proposed window method, the all-mean method, and the baseline using randomized window configurations, when all modalities are used. As shown in Table 7.1, using all modalities shows the highest accuracy compared to pairs of modalities. Each column of the tables shows the UW recall, per-emotion class accuracy for angry, happy, neutral, and sad emotion classes and weighted accuracy.

Type	Method	UW	A	H	N	S
Prot	Proposed	77.49	83.50	81.65	62.73	80.66
	All-mean	75.85	80.77	83.47	53.26	83.26
Non-prot	Proposed	57.29	64.15	56.54	36.15	72.33
	All-mean	57.33	62.23	58.60	33.56	74.95

Table 7.3: *IEMOCAP dataset: Emotion classification unweighted recall (%) for prototypical and non-prototypical utterances. The accuracies are unweighted recall ('UW'); and per-class accuracy for angry ('A'), happy ('H'), neutral ('N'), and sad ('S') emotion classes.*

Finally, to address Q3, we investigate the performance gain of our proposed system. The gain from the all-mean method is mostly from the prototypical expressions (defined as rater consensus). The results are presented in Table 7.3. For prototypical utterances, our method gets 77.49% and all-mean gets 75.84% (1.65% difference. p-value=0.53). For non-prototypical utterances (defined as no rater consensus), our method gets 57.29% and all-mean method gets 57.33% (0.05% difference, p-value = 0.93). This may indicate that the temporal patterns of emotion and emotion spotting is more useful when the expression is more explicit and prototypical to human evaluators.

7.9 Conclusions

In this chapter, we explore whether a subset of an utterance can be used for emotion inference and how the subset varies by emotion classes and modalities. We propose a windowing method that identifies window configurations, window duration and timing, for aggregating segment-level information to an utterance-level emotion inference. The experimental results demonstrate that the identified temporal window configurations show consistent patterns across speakers, specific to different emotion classes and modalities.

We compare our proposed windowing method to a baseline method that randomly selects window configurations and a traditional all-mean method that uses the full information within an utterance. Our proposed method shows significantly higher performance in emotion recognition than the baseline method, achieving 65.60% UW accuracy, 1.90% higher than the baseline). Our method also achieves similar performance to the traditional all-mean method (65.59%, statistically insignificant difference), while our method only uses 40–80% of information within each utterance.

The identified windows also show consistency across speakers, varying by different emotion classes and modalities. For the angry emotion class, lower face, upper face, and speech uses 54%, 50%, and 64% information in the beginning of an utterance. For the happy emotion class, lower face uses 64% at the end of an utterance, and upper face and speech use 68% and 80% in the beginning of an utterance. For the neutral emotion class, the lower face, upper face, and speech use 70%, 42%, and 40% of an utterance, however the temporal window patterns are less consistent across speakers. For the sad emotion class, the lower face uses 52% in the end of an utterance, while the upper face and speech use 54% and 64% in the beginning of an utterance. This finding also matches with findings from the psychology literature, particularly for speech data, where happy emotion class requires more information than other emotion classes to be accurately recognized.

CHAPTER 8

Transition Patterns in Behavior

8.1 Introduction

The pervasive installations of large camera networks and widely availability of digital video cameras have created a gigantic volume of video data that need to be processed and analyzed to retrieve useful information. As many videos involve human activities and behaviors, a central task and main challenge in video analytics is to effectively and efficiently extract complex and highly varying human-centric events. A general purpose event recognition system entails two essential steps: the localization of temporal segments in a video containing salient events (*when something happened*) and the classification of localized events into relevant categories (*what happened*). The extracted events can be piped for further analysis, such as indexing and retrieval of video collections in multimedia applications and suspicious behavior recognition in video surveillance.

Most update-to-date video event analysis methods treat event localization and classification as separate problems (e.g. [128, 169]). It has been noticed that these two problems are interrelated and can mutually bootstrap each other [41, 89]. Better event localization improves subsequent classification performance, while reliable event classification can be used as a guide for more precise localization. Based on this intuition, recent efforts have emerged in unifying both the localization and classification problems. These methods fall into two main categories: (i) *generative* approaches based on dynamic Bayesian models, such as the hidden Markov model (HMM) [20] and switching linear dynamical systems (SLDS) [171]; and (ii) *discriminative* approaches, which use maximum margin classifiers as in [33, 41, 89].

Conventional event models used in most existing methods only consider monolithic or persistent events. For example, action recognition focuses on the identification of action

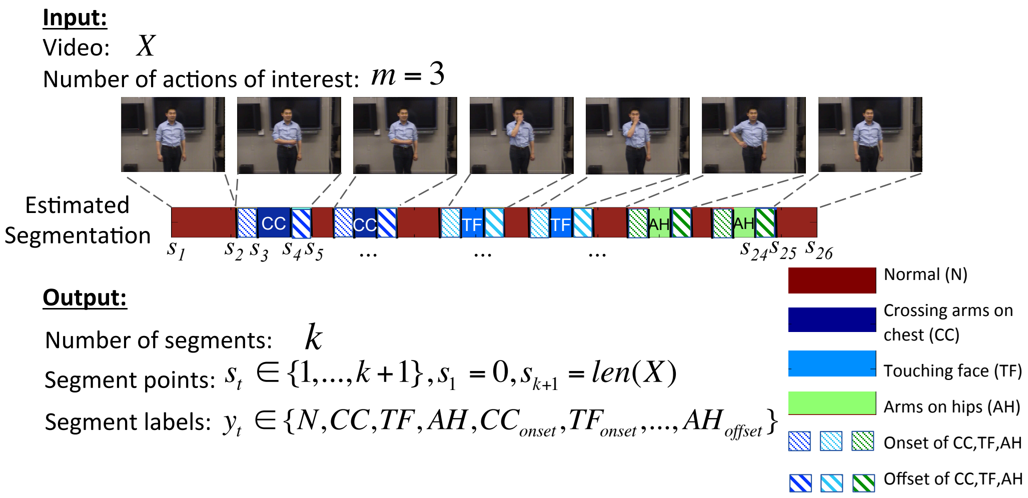

Input:
Video: X
Number of actions of interest: $m = 3$
Estimated Segmentation
CC
CC
TF
TF
AH
AH
s_1
s_2 s_3
s_4 s_5
...
...
...
s_{24} s_{25} s_{26}
Output:
Number of segments: k
Segment points: $s_t \in \{1,\ldots,k+1\}, s_1 = 0, s_{k+1} = len(X)$
Segment labels: $y_t \in \{N, CC, TF, AH, CC_{onset}, TF_{onset}, \ldots, AH_{offset}\}$
Normal (N)
Crossing arms on chest (CC)
Touching face (TF)
Arms on hips (AH)
Onset of CC,TF,AH
Offset of CC,TF,AH

Figure 8.1: Overview of the proposed video event localization and classification framework, where the event types are, e.g., *Crossing arms on Chest* (CC), *Touching Face* (TF), *Arms on Hip* (AH), and *Neutral* (N) (Section 8.3.1). The temporal onset and offset transitions between these events are optimally solved by efficient dynamic programming.

states such as walking or standing with arms folded. These methods ignore the regular transition patterns often exist between events of interest. To illustrate, consider a person with his/her arms down in a resting position who starts to raise his/her arm to touch his/her nose. A transition segment or event in which the arm moves upward governs the change between gesture states. Although a naive detection of such transition might be difficult (following the generative or discriminative approaches), the consecutive motion flow in between the transitions is indeed unique and recognizable. Explicitly incorporating *transition patterns* into the recognition framework will provide more reliable cues to localize and recognize persistent events.

In this chapter, we propose a new method that jointly analyzes video events with precise temporal localization and classification, by modeling arbitrary transition patterns between events. It improves event recognition rates by leveraging the clearly identified event boundaries. Our method combines two approaches together by explicit modeling of *event transition segments*: (i) large margin discriminative learning of distinct event patterns (also introduced in [41, 89]) and (ii) generative event-level transition probability models. The event location and classification can be found by an efficient dynamic programming (DP) inference. Our framework is general to any time series data that have transition patterns between events and is applicable to problems outside video analytics. For human action recognition in particular, the use of transition patterns can greatly improve performance. Since even the same action (e.g. touching face) can be highly varying in both spatial and temporal domains, their transition patterns are more important for robust systems. Explicit consideration of transition patterns increases robustness and can provide critical information for decision making [181, 204, 241].

We focus on the application of video-based human action recognition. Specifically, we extract per-frame human pose estimation cues (i.e. body joint coordinates) [199] as a time series signal. We compute variable-length segment-level features using statistical functionals and linear regression coefficients (slope) of the frame-level features for each segment. In the supervised training phase, we use labeled intervals of video events and their corresponding event types to train a discriminative model. This model is used in the testing phase, in which for a given test video, we infer the best segmentation start and end points with corresponding event labels, by searching for the highest pattern matching score and transition probability using efficient dynamic programming. Figure 8.1 provides an overview of our framework.

Our method has demonstrated significantly improved classification and localization performance on a newly collected video dataset and a public CMU-MAD [94] benchmark dataset, in comparison to a state-of-art work [89].

8.2 Proposed Method

Our method can be applied to general tasks of segmenting human actions with transition patterns. Our proposed algorithm (Equation 8.2) is generic to model arbitrary transitions between actions, and transitions between actions and neutral states (e.g., standing person with hands down). Any transition event model can be applied based on the transition characteristics that reflect the nature of the problem or the dataset. However, neutral states between events are prevalent in the datasets we performed experiments on, and thus it is important to model them effectively in our chosen transition event model. We describe our event transition model with segment transition probabilities in Section 8.2.1. We then describe our generic method for event finding, localization and classification: the training of a multi-class SVM using the peak and transition segments (Section 8.2.2.1) and the inference and labeling of each putative temporal segments using the SVM and dynamic programming (Section 8.2.2.2).

8.2.1 Transition Event Model

Event Peak and Transition Segments. Any transition event model can be used to describe the temporal characteristics present between events of interest. Since the two datasets we tested have prevalent neutral states between events, we explicitly models four types of segments: neutral, peak, onset, and offset. *Neutral* segments describe no significant visual cues of any event of interest. *Peak* segments describe salient and consistent visual cues of an event of interest. Both the definitions of neutral and peak can be application dependent (see Section 8.3). For each event type, we define two types of event transition segments based on the neutral and peak segments: *Onset* transition segments describe the transition from neutral to peak events, and *Offset* transition segments describe the transition from peak back to neutral.

In many video event analytic applications, segments of no particular utility or interest can be modeled as neutral events. Visual cues of onset transitions of the same peak event share commonalities (and the same for offset transitions). Thus a repeating sequence of "*neutral-onset-peak-offset-neutral*" can be found in many event types of interest. For instance, Figure 8.2 shows an example of neutral, onset, offset, and peak segments for the action event corresponding to "crossing arms on chest." We assume a simpler event model that does not consider direct transitions between events without going through the neutral event. This assumption effectively reduces the modeling of rarely occurred transitions, as supported by our experimental results.

Segment-level Transition Probability. We model the temporal patterns between

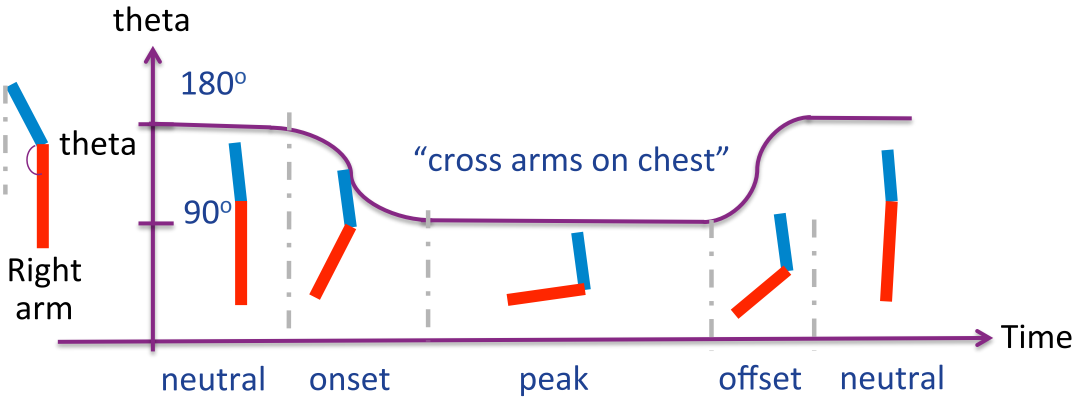


Figure 8.2: transition event model example: the *neutral-onset-peak-offset-neutral* model of *cross arms on chest*. For visualization purpose, the joint angle θ between the upper and lower arms is shown as a cue to segment out the "cross arms" and "arm-down" events.

neutral, peak, onset, and offset segments using a transition probability matrix. Following the *neutral-onset-peak-offset-neutral* observation from the training dataset, the transition probability from peak to offset, offset to normal, and onset to peak can be equally assigned to a default value based on the frequencies of event transitions. For the transition from neutral states, we model two cases: (i) the changing to one of the m types of possible events is modeled with a transition probability P, or (ii) the event remains unchanged, which is modeled with a *self-transition* probability γ. In this chapter, γ was chosen as 0.5 to maximize the randomness of repeating the same events.

8.2.2 SVM-based Event Localization and Classification

The input and output notations of our proposed system are described in Figure 8.1. We first train a multi (M)-class SVM using event peak and transition segments (vs. neutral segments). In testing, for a given video X without any segmentation information, we automatically find the optimal number of segments k, the temporal start and end points of each segment $s_t, t \in 1, ..., k+1$, where $s_1 = 0$ and $s_{k+1} =$

$len(X)$ the length of X, and segment labels $y_t, t \in 1, ..., k$. Our method keeps track of the highest sum of SVM scores and the log transition probability of all segments.

8.2.2.1 Training Segment-SVM with Max Margin Optimization

We learn discriminative patterns of each peak and transition segments using a multi-class SVM [48] similar to [89]. For each video sequence in the training data X^i, where $i \in \{1, 2, ..., n\}$, with known segments $t \in \{1, 2, ..., k_i\}$, where k_i is the number of segments of the i-th video sequence, we solve the following SVM and learn weights w_j for inference:

$$\min_{w_j, \xi_t^i \geq 0} \frac{1}{2M} \sum_{j=1}^{M} ||w_j||^2 + C \sum_{i=1}^{n} \sum_{t=1}^{k_i} \xi_t^i, \tag{8.1}$$

$$\text{s.t.} (w_{y_t^i} - w_y)^T \varphi(X^i_{(s_t^i, s^i_{(t+1)}]}) \geq 1 - \xi_t^i, \forall i, t, y \neq y_t^i,$$

where $\varphi(X^i_{(s_t^i, s^i_{(t+1)}]})$ is the segment-level feature of the segment $X^i_{(s_t^i, s^i_{(t+1)}]}$, consisting of frames from s_t^i to $s^i_{(t+1)}$. We describe the segment-level feature mapping in detail in Section 8.3.

8.2.2.2 Efficient Inference with Dynamic Programming

Transition-based Segmentation. For each test video sequence X with unknown segment points and labels, we segment and classify the sequence using the following optimization function that maximizes the sum of the total SVM scores and the log transition probability between consecutive segment pairs:

$$\max_{k,s_t,y_t} \sum_{t=1}^{k} w_{y_t}^T \varphi(x_t) + (1+\gamma)\log P(y_t|y_{t-1}), \text{ s.t.} \tag{8.2}$$

$$l_{min} \le s_{t+1} - s_t \le l_{max}, \forall t,$$

$$s_1 = 0, s_{k+1} = len(X),$$

The intuition is to maximize the sum of segment-specific scores for each segmentation configuration, i.e. determine the number of total segments k, segment points s_t, and segment labels y_t, where $t \in \{1, 2, ..., k+1\}$, as well as the probability of transition from one segment to another. l_{min} and l_{max} are the minimum and maximum length of segments in the training data.

The relationship between temporally adjacent segments $(1+\gamma)\log P(y_t|y_{t-1})$ is calculated based on our prior transition probabilities described in Section 8.2.1. Our novelty compared to Hoai et al. [89] is the $\log P(y_t|y_{t-1})$ term that explicitly considers event transitions in the optimization framework. Our work also differs from [89] in that non-maxima suppression based segmentation is performed (instead of a maximum SVM score based segmentation). Hoai et al. chooses the optimal segmentation that maximizes the difference of SVM scores between the best and the second best class, by filtering using the Hinge loss. We take a different approach by seeking the optimal segmentation that maximizes the sum of both (i) the SVM score of the segment class and (ii) the transition probability between consecutive segments.

Inference using DP. To solve Eq.(8.2) efficiently, we formulate the following function f to determine the best segmentation for the truncated time series $X_{(0,u]}$,

$$f(u, y_k) = \max_{k,s_t,y_t} \sum_{t=1}^{k} w_{y_t}^T \varphi(x_t) + (1+\gamma)\log P(y_t|y_{t-1}), \tag{8.3}$$

where k is the number of segments for the truncated $X_{(0,u]}$. u can be considered as the increasing "front" of the dynamic programming (DP) formulation. Since the transition probability depends on the last segment's label y_k of the truncated time series $X_{(0,u]}$, each f value depends on u as well as y_k. Therefore, for every tuple $u \in (0, len(X))$, $l \in [l_{min}, l_{max}]$ and class $y \in \{1, 2, ..., M\}$, we calculate $\eta(u, l, y) = w_y^T \varphi(X_{(u-l,u]})$ for inference, where η is the SVM score of the segment $X_{(u-l,u]}$. Dynamic programming computes $\max_{y_k} f(len(X), y_k)$ efficiently using Equation 8.4. Algorithm 1 lists the pseudo code, where w is a learned weight vector, *testX* and $len(X)$ are test video sequence and the number of frames of it, m_{tr} and std_tr are mean and standard deviation of each feature dimension in the training data for z-standardization, nCl is the number of classes, and $transMat$ is a transition matrix to calculate f.

$$\begin{aligned} f(u, y_k) = \max_{l, y_{k-1}} & f(u-l, y_{k-1}) + \eta(u, l, y_k) \\ & + (1+\gamma) \log P(y_k | y_{k-1}) \end{aligned} \tag{8.4}$$

Algorithm 1: DP with transition Event Model

```
Data: learned weight vector $w$, test video $X$, $m_{tr}$, $std_{tr}$, $l_{min}$, $l_{max}$, number of
      classes $nCl$
Result: $f$, $bestL$, $bestY_{k-1}$
for each frame $u = l_{min}$ : $len(X)$ do
    for each last segment label $y_k$ = 1:$nCl$ do
        for $l = l_{min}$:$min(l_{max}, u-1)$ do
            Calculate $\eta(u, l, y) = w_y^T \varphi(X_{(u-l,u]})$, where $\varphi(X_{(u-l,u]})$ is z-standardized
            using $m_{tr}$ and $std_t r$.
        end
        for each second last segment label $y_{k-1}$ = 1:$nCl$ do
            $f_{temp}(l, y_{k-1}) = f(u-l, y_{k-1}) + \eta(u, l, y_k) + \log P(y_k|y_{k-1})$
        end
        find $y_{k-1}^*$, $l^*$ that maximizes $f_{temp}(u, y_k)$. $f(u, y_k) = f_{temp}(l^*, y_{k-1}^*)$
        $bestL(u, y_k) = l^*$ $bestY_{k-1}(u, y_k) = y_{k-1}^*$
    end
end
Use $f, bestL, bestY_{k-1}$ for back-tracking
```

The complexity of our algorithm is $O(M^2(l_{max} - l_{min} + 1)(len(X) - l_{min} + 1))$.

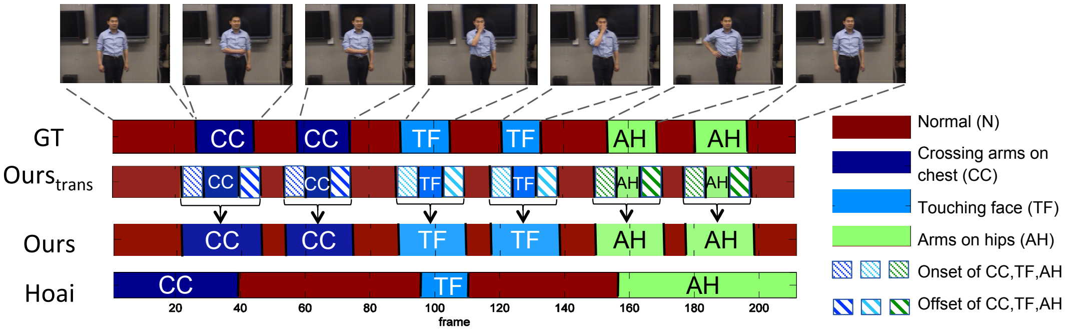


Figure 8.3: Evaluation results from our Smartroom (Clean) Dataset (video 1). The four rows of illustrations depict ground truth (first row), result of our method with transition segments (second row), result of our method with combined transition (onset and offset) segments into a single action segment to match the comparison of Hoai et al. (third row), and SVM+DP method output presented by Hoai et al.[89] (bottom row), respectively.

8.3 Experiments

We evaluate our method for joint segmentation and classification of video events on two datasets: (i) the Smartroom Dataset we collected for real-life suspicious behavior recognition and (ii) the public CMU-MAD human action dataset [94]. Both of the datasets contain large variability in human poses and actions.

We compare the performance of our algorithm to the SVM-DP algorithm of Hoai et al. [89]. For a fair comparison to the SVM-DP algorithm of Hoai et al., which does not consider the transition segments, we calculate the recognition rate after transferring the estimated M action classes with transition segments, where $M = \{m \text{ peak events}\} + \{1 \text{ neutral event}\} + \{m \text{ offset events}\} + \{m \text{ onset events}\}$, to m peak action classes, as shown in Figure 8.3. We combine the detected onset, offset, and peak segments of each action into one action. For instance in our Smartroom Dataset, after we finish back-tracking and get 10-class labels for each detected segment, we combine onset, offset, and peak segments into one action segment to match the 4-class ground-truth labels.

We report the performance of both algorithms in terms of frame-level and event-level recognition rates. (i) Frame-level recognition rate measures the ratio of frames that are correctly classified. We compute frame-level precision ('Prec'), recall ('Rec'), and F-measure ('F-mea'). The accuracy is calculated as $(TP+TN)/(TP+TN+FP+FN)$, where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative, respectively. (ii) The measure of event-level recognition rate is suggested in [94] to reflect the ratio of event segments that are correctly identified, by counting the number of correct frames that overlaps with 50% of a segment. We evaluate event-level precision, recall, and F-measure. Event-level precision ($prec$) computes the ratio between the number of correctly detected events and the number of detected events and event-level recall (rec) computes the ratio between the number of correctly detected events and the number of ground truth events. Event-level F-measure computes the balanced F-score using $2 * \frac{prec*rec}{prec+rec}$. In our datasets where there is at most 9 ground truth events, our event-level recognition rate is highly sensitive compared to frame-level recognition rates.

8.3.1 Smartroom Dataset

We collect and create a new Smartroom Dataset, described in Section 2.3.

We use the MODEC algorithm [199] to estimate per-frame body pose cues to serve as action features, and we employ a Kalman filter to produce a smooth pose time series. The pose estimation from the image is converted into body joint angles as shown in Figure 8.5. The performance of MODEC pose estimation varies for different clothing and illumination conditions. We evaluate the robustness of event recognition upon such variability in the input data. We divide the Smartroom dataset into two subsets and evaluate our system for each subset: (i) the ones with more accurate pose estimation ("Clean"), (ii) the remaining with large pose estimation noise due to appearance and clothing variations ("Noisy"). Comparisons of the MODEC pose

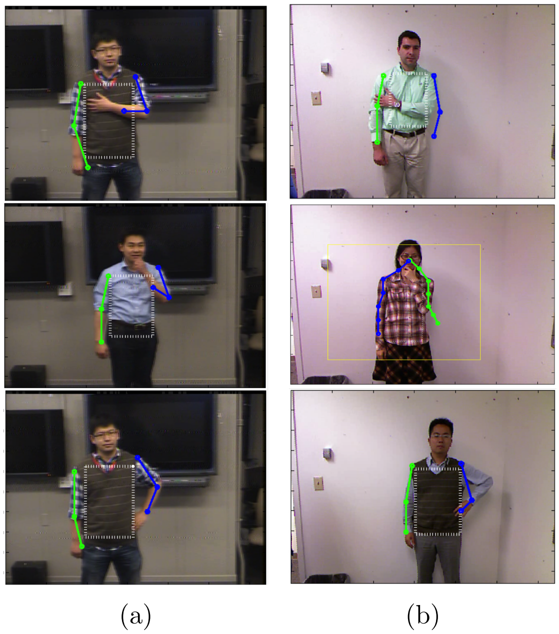


Figure 8.4: Pose estimation comparison between the Smartroom (a) Clean and (b) Noisy datasets for *Crossing arms on chest* (top), *Touching face* (center), and *Putting arms on hip* (bottom) actions. The performance of the MODEC algorithm [199] varies for different clothing and illumination conditions. The Smartroom (Clean) dataset shows more accurate pose estimation than the Smartroom (Noisy) dataset.

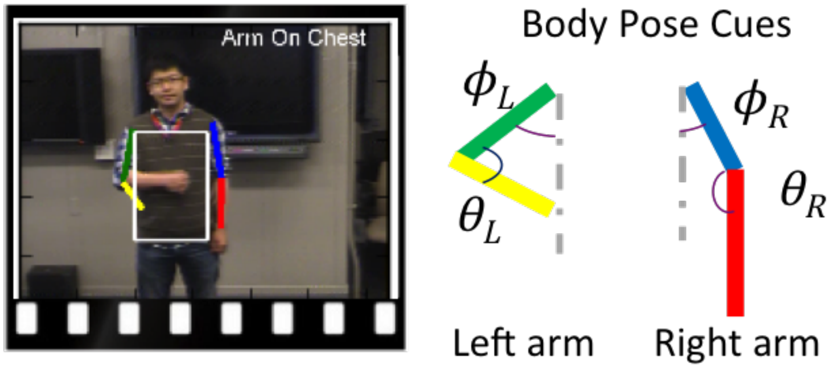


Figure 8.5: Estimated body pose cues of our Smartroom Dataset utilized for frame-level features. We estimate the four joint angles at the shoulders (between torso and upper arms: ϕ_L, ϕ_R) and the elbows (θ_L, θ_R).

estimations on the two subsets are shown in Figure 8.4. The Smartroom (Clean) dataset contains three videos, and the Smartroom (Noisy) dataset contains five videos.

Two types of segment-level features φ are extracted for each video segment: (1) the first and second-order statistics (mean and standard deviation) of the frame-level features, and (2) the linear regression coefficient (slope) across frames within each segment, which captures the dynamics of the changes of the frames within the segment. We perform z-standardization to normalize the segment-level features as follows: we first find the mean m_i and standard deviation st_i of each feature dimension i in the training data and normalize the training data (z-standardization) using the two statistics. Then, during the inference, we use the same mean m_i and standard deviation st_i of each feature dimension to normalize the test segments in the Dynamic Programming steps.

We perform leave-one-video-out cross validation, and take a subset (left-hand movements) of a video as a test sequence. We train our model using the remaining videos. Figure 8.3 shows the segmentation result comparison between the ground truth *(top)*, our algorithm *(center)*, and the algorithm presented by Hoai et al. ([89], "Hoai SVM+DP") *(bottom)*. Both methods determine the start and end points, as

	Frame-level						Event-level					
	Prec		Rec		F-mea		Prec		Rec		F-mea	
Method	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Ours	83.84	7.45	80.41	12.18	81.95	9.52	86.67	11.55	89.63	10.02	88.07	10.54
Hoai	56.19	5.32	60.50	7.98	58.15	5.74	71.11	7.70	67.41	12.24	68.32	3.86
Diff	**27.65**		**19.91**		**23.79**		**15.55**		**22.22**		**19.75**	

Table 8.1: *Recognition rate (%) of Smartroom (Clean) Dataset using our proposed algorithm and the Hoai et al. [89] at the frame and event level (see text). The last row ("Diff") shows the relative improvement of using our algorithm over the algorithm of Hoai et al.*

	Frame-level						Event-level					
	Prec		Rec		F-mea		Prec		Rec		F-mea	
Method	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Ours	44.41	18.85	40.38	18.20	41.33	17.09	25.36	16.36	54.45	15.91	33.51	17.93
Hoai	24.39	11.54	13.60	6.88	17.26	8.33	14.33	14.93	11.20	6.81	11.75	10.56
Diff	**20.02**		**26.78**		**24.07**		**11.03**		**43.24**		**21.76**	

Table 8.2: *Recognition rate (%) of Smartroom (Noisy) Dataset using our proposed algorithm and the Hoai et al. [89] at the frame and event level (see text). The last row ("Diff") shows the relative improvement of using our algorithm over the algorithm of Hoai et al.*

well as the label of each action event. Our method significantly outperforms the method of Hoai et al. in terms of both frame and event-level recognition rates.

Tables 8.1 and 8.2 show the comparisons between our algorithm and the algorithm of Hoai et al. for the Smartroom (clean) and Smartroom (noisy) datasets, respectively. For the Smartroom (clean) dataset, our algorithm has a frame-level precision of 83.84%, recall of 80.41%, and an F-measure of 81.95%. All of the frame-level recognition rates are higher than the SVM-DP method of Hoai et al. by 27.65%, 19.91%, 23.79% (relative improvements) in terms of precision, recall, and F-measure, respectively. Also, event-level precision, recall, and F-measure of our algorithm are 86.67% 89.63% 88.07%, respectively, 15.55%, 22.22%, and 19.75% higher than the method of Hoai et al. Our algorithm also demonstrates improvements in performance even when the pose estimation was noisy. For the Smartroom (noisy) dataset, our algorithm shows a frame-level precision of 44.41% , recall of 40.38%, and F-measure of 41.33%; relative improvement of 20.02%, 26.78%, and 24.07%, compared to the method of Hoai et al. The event-level recognition rates are also significantly improved when using our algorithm. The event-level precision of our system is 25.36 %, recall was 54.45%, and F-measure was 33.51%. These are 11.03%, 43.24%, and 21.76% relative improvement over the method of Hoai et al [89]. This demonstrates that with a presence of clear transitions between actions, our algorithm can robustly segment and classify each salient action.

8.3.2 CMU-MAD Dataset

We test our method on the CMU-MAD dataset [94], described in Section 2.4. We perform 5-fold cross validation over the 20 subjects and measure the event-level performance as suggested in [94]. Each fold contains videos of 4 subjects, each with 2 video sequences, in total 8 video sequences. We train our model using segments of the four folds and test our model for the held out. Due to the computational cost, we

	Frame-level						**Event-level**					
	Prec		Rec		F-mea		Prec		Rec		F-mea	
Method	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Ours	85.00	8.82	71.41	7.25	77.41	7.01	74.40	15.02	85.02	12.17	78.83	12.95
Hoai	73.79	9.62	70.57	9.96	71.87	8.70	73.45	15.84	83.88	13.06	77.85	14.23
Diff	**11.21**		**0.84**		**5.54**		**0.95**		**1.14**		**0.98**	

Table 8.3: *Recognition rate (%) comparison on the CMU-MAD dataset using our proposed algorithm ("Ours") and the Hoai et al. ("Hoai", [89]) at the frame and event level (see text). The last row ("Diff") shows the relative improvement.*

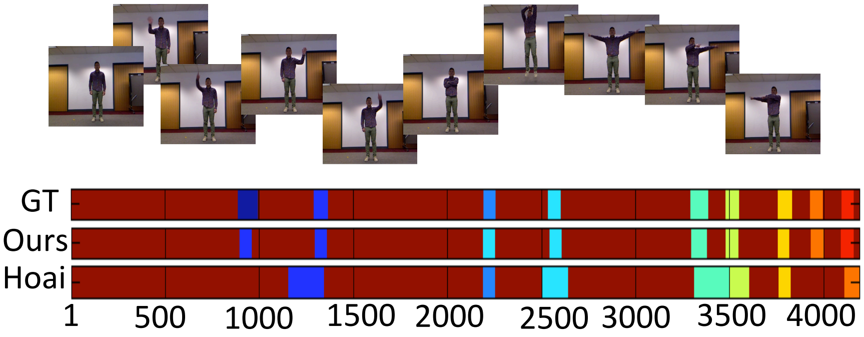


Figure 8.6: An example result (subject 20, sequence 20) from the CMU-MAD evaluation. Ground truth (top), our method (center), and SVM+DP presented by Hoai et al. [89] (bottom). Best viewed in color. The image is from the CMU-MAD dataset [94].

use DP over sliding windows of 500 frames (about 25% length of a video sequence) along the test time series as in [94], to solve for the optimal segment configuration that maximizes the sum of the SVM scores and the event transition probability.

Figure 8.6 shows the results of our algorithm (center) and Hoai's SVM+DP method (bottom), along with the ground truth segmentation (top). Table 8.3 summarizes the results. All of our frame-level recognition measures are higher than the SVM-DP method of Hoai et al. [89]. For event-level accuracy, our event-level precision (74.40%), recall (85.02%), and F-measure (78.3%) are higher than the SVM-DP method, by 0.95%, 1.14%, and 0.98%, respectively.

Our method improves the frame-level recognition rates compared to the previous work of Hoai et al. [89], achieving 85.00% (precision), 71.41% (recall), and 77.41% (F-measure), corresponding to relative improvement of 11.21%, 0.84%, and 5.54%, respectively. We achieve an event-level precision of 74.40%, recall of 85.02%, and F-measure of 78.83%, and all of these are slightly higher than that of Hoai et al. by 0.95%, 1.14%, and 0.98%, respectively. The improvement in both frame and event-level recognition rates using our algorithm over the previous method of Hoai et al. [89] demonstrates that for actions of interest with distinguishable transition patterns, our algorithm can localize and classify the action segments more effectively.

Regarding the difference between the Smartroom and CMU-MAD dataset results in performance gain, we raise two major points: (i) the transition segments were not explicitly labeled for the CMU-MAD dataset, therefore the segments were estimated during training. Since the major advantage of our method is a better modeling of the transition states, the improvement on CMU-MAD dataset is marginal. This also explains a greater performance gain in the frame-level compared to the event-level accuracy. In comparison, our Smartroom dataset includes clearer labeling in event transitions; hence the performance improves significantly due to better transition modeling. (ii) The visual features for the Smartroom dataset (i.e., pose estimation features from RGB cameras without depth information) are more difficult to estimate and thus are noisier than those of the CMU-MAD dataset (i.e., 3D pose estimation features using Kinect sensor). Therefore, a better transition model as a prior results in a greater performance gain on the Smartroom dataset, where the input features are noisier in nature.

8.4 Conclusions

In this chapter, we describe a new method combining discriminative large margin classification with generative modeling, where the explicit modeling of event transition segments improves the state-of-art performance on the joint localization and classification of video events. Our experimental results on two benchmark datasets shows promising recognition rates. An important future work we plan to pursue is the consideration of event transition probability with discriminative learning in finding an effective solution to model the full relationships between events.

Nevertheless, there is still room for improvement in the current work. In particular, though this chapter demonstrates that the modeling of onset and offset of event transitions can boost the localization and classification of video events, while effective solution to properly model the full relationships between pairwise events are yet to be explored. In future work, we will study automatic methods that can learn the transition probabilities of the full set of pairwise event transitions.

Conclusions and Future Directions

CHAPTER 9

Main Contributions

In this dissertation, we have studied how to computationally represent, model, and analyze complex and time-changing facial and vocal behaviors that co-occur with multiple sources of variation, with the aim of identifying emotion-specific patterns. Three important research questions have been answered to achieve this goal:

1. *Motivational Studies*: How can we fuse information from audio and visual expressions? How can we capture emotion expressed over time?
2. *Multiple factors in behavior*: How can we capture emotion expressed over time, and how can we handle multiple factors that modulate audio-visual behavior?
3. *Localization of salient events*: How can we detect salient events in audio-visual behavior?

The first part of dissertation (Chapter 3) explored facial and vocal behaviors during expressions of emotion. We first proposed methodologies using deep learning that capture complex non-linear interactions between audio and visual emotion expressions. This approach overcame the limitations of traditional methods that are only capable of capturing linear relationships between modalities, or, alternatively, require labeled data when extracting multimodal features. The proposed method showed improvement in emotion classification rates, particularly for ambiguous emotion content (defined as no rater consensus). We also investigated continuous changes of emotion in Chapter 4. We found that there exist structural patterns of emotion changes within an utterance, typical of each emotion class of anger, happiness, neutrality, and sadness. These structural patterns were shown to be effective in discriminating between different emotion classes.

The second part of this dissertation (Chapters 5 and 6) explored how emotion variations modulate facial movement when a person is speaking, a challenging situation in emotion recognition (e.g., recognition systems need to differentiate between a person smiling vs. saying ‘cheese’). We found that variable-length time units that capture the natural dynamics of facial movements are critical for emotion classification. We developed and proposed a new variable-length segmentation method that utilized the dynamics of individual face regions, showing significant improvement in the system accuracy.

The third part of this dissertation considers the problem of identifying salient regions in audio-visual affective behavior. We discovered that consistent patterns exist in the timings and durations of emotion evidence from the upper face, lower face, and speech modalities (Chapter 7). In Chapter 8, we further introduced an efficient inference method that can jointly segment and classify temporal data, with a focus on human action behavior. The novelty of this method is that it models transition patterns between event segments of interest, such as a person’s gesture changes when moving an arm upwards from a resting position to touch the nose. The method showed significant performance gain compared to traditional segmentation methods.

CHAPTER 10

Future Work

The overarching theme of this thesis is developing machine learning and signal processing frameworks to automatically identify emotion from audio-visual expressions. In future work, we will initially focus on the following research directions to explore the broader affective and social signals during human interactions.

Machine Adaptation for Personalized Technology

The first line of research aims to continuously adapt machine responses based on extracted information on the changing needs and affective states of individuals. Humans constantly sense and adapt their reactions based on the emotions and needs of their interlocutors. This research will seek to provide this adaptation capability to machines in the context of different domains, ranging from personal assistant systems in mobile phones to smart vehicles. For instance, speech recognition systems equipped in personal devices, such as Apple's Siri, can be trained based on a user's affective state. The systems that neglect affective states often have poorer speech recognition accuracy when a user's tone of voice changes due to emotion, but not speech itself [14]. The systems can also enhance user experience by naturally adapting to affective states, e.g., adapting the speed of games based on user's level of boredom [229] or engagement, adapting virtual reality systems for stress-coping training [45], or outputting apologies, such as 'we apologize for this inconvenience' when a user is frustrated [47]. This research will first explore methods to represent and track continuous changes in individuals' needs and emotive states. The automatic detection and representation of individuals' states will inform the design and development of personalized technology.

Human-Human Interaction Assistant System

The second line of research focuses on developing virtual assistant systems for human-human communication. It aims to design systems that sense, quantify, and track communication participants' emotion, engagement, and satisfaction level throughout the course of an interaction. The system outputs will provide users with objective feedback on their behavioral patterns during interactions, and will also suggest strategies to enhance the behaviors of the users, e.g., tone of voice or postures, so that they can achieve the desired outcomes from interactions. Example interactions include negotiation and collaboration between parent-child, teacher-student, and patient-caregiver. To this aim, this research will first explore hierarchical prediction models that infer overall interaction outcomes. The models will estimate temporal changes in affective and social cues using audio-visual features, and deploy these estimates as mid-level representation for the final inference of interaction outcomes, e.g., success or failure in negotiation. Our research will further design feedback mechanisms to the users based on the inference results.

Human-Centered Multimedia Content Analysis

The third line of research seeks to analyze and retrieve human-centered information from long, time-changing multimedia content, such as documentary interviews, surveillance videos, and presidential debates. This research can create a breakthrough in affective and social computing, since it can help acquire natural and authentic emotion expressions, which is rare in current research. Also, this research can greatly advance the retrieval of multimedia information relevant to its viewers. Such advancement in human-centered multimedia retrieval is critical in today's explosion of the number of multimedia data. For instance, security officers monitoring vast amounts of surveillance video footage are interested in detecting video segments of fights or conflicts for preventing potential crimes. Another example is presidential debates, where voters may be interested in retrieving moments of the candidates showing nonverbal messages. In these two examples, it will be beneficial for viewers if an automated system can retrieve a subset of the videos that are relevant to them. The purpose of this research is to identify salient time segments that contain application-specific affective and social events for multimedia retrieval and indexing. This research problem is complicated by highly-varying, unstructured human behaviors, and to tackle this problem, we will initially investigate time-series segmentation and analysis methods for uncovering structural patterns of events of interest.

Last, the research directions we pursue need innovation and active collaboration with researchers across multiple fields. We envision to apply the technique developed in this thesis to diverse areas, including psychology, sociology, behavioral sciences, psychiatry, engineering, computer science, and information science. The interdisciplinary collaboration with these fields will be critical to enable a true human-centered understanding. The interdisciplinary efforts will integrate knowledge and methodologies of these disciplines towards the unifying goal of automatic human behavior analysis.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] (Accessed: 2016-05-11), Pepper softbank official website, http://www.softbank.jp/en/robot/.

[2] Abelson, R. P., and V. Sermat (1962), Multidimensional scaling of facial expressions., *Journal of Experimental Psychology*, *63*(6), 546.

[3] Adolphs, R. (2002), Recognizing emotion from facial expressions: Psychological and neurological mechanisms, *Behavioral and cognitive neuroscience reviews*, *1*(1), 21–62.

[4] Anagnostopoulos, C., T. Iliou, and I. Giannoukos (2012), Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, *Artificial Intelligence Review*, pp. 1–23.

[5] Arons, B. (1994), Pitch-based emphasis detection for segmenting speech recordings., in *International Conference on Spoken Language Processing*, pp. 1931–1934.

[6] Atkinson, A. P., W. H. Dittrich, A. J. Gemmell, and A. W. Young (2004), Emotion perception from dynamic and static body expressions in point-light and full-light displays, *Perception*, *33*(6), 717–746.

[7] Attabi, Y., and P. Dumouchel (2013), Anchor models for emotion recognition from speech, *IEEE Transactions on Affective Computing*, *4*(3), 280–290.

[8] Barod, J. C., E. Koff, M. P. Lorch, and M. Nicholas (1986), The expression and perception of facial emotion in brain-damaged patients, *Neuropsychologia*, *24*(2), 169–180.

[9] Bassili, J. N. (1979), Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face., *Journal of personality and social psychology*, *37*(11), 2049–2058, doi:10.1037/0022-3514.37.11.2049.

[10] Bassili, J. N. (1979), Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face., *Journal of personality and social psychology*, *37*(11), 2049.

[11] Bates, D., M. Maechler, and B. Bolker (2007), lme4: Linear mixed-effects models using s4 classes (r package version 0.9975-11) [computer software].

[12] Batliner, A., D. Seppi, S. Steidl, and B. Schuller (2010), Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach, *Advances in Human Computer Interaction, Special Issue on Emotion-Aware Natural Interaction.*

[13] Bengio, Y. (2009), Learning deep architectures for AI, *Foundations and Trends in Machine Learning*, *2*(1), 1–127.

[14] Benzeghiba, M., et al. (2007), Automatic speech recognition and speech variability: A review, *Speech Communication*, *49*(10), 763–786.

[15] Bevacqua, E., and C. Pelachaud (2004), Expressive audio-visual speech, *Computer Animation and Virtual Worlds*, *15*(3-4), 297–304.

[16] Bhattacharya, S., B. Nojavanasghari, T. Chen, D. Liu, S.-F. Chang, and M. Shah (2013), Towards a comprehensive computational model foraesthetic assessment of videos, in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 361–364, ACM.

[17] Bigot, B., I. Ferrane, and Z. Ibrahim (2008), Towards the detection and the characterization of conversational speech zones in audiovisual documents, in *International Workshop on Content-Based Multimedia Indexing, 2008. CBMI 2008.*, pp. 162–169, IEEE.

[18] Boersma, P., and D. Weenink (), Praat: doing phonetics by computer (version 6.0.17)[computer program]. retrieved 21 april 2016 from http://www.praat.org/.

[19] Boston, M., J. Hale, R. Kliegl, U. Patil, and S. Vasishth (2008), Parsing costs as predictors of reading difficulty: An evaluation using the potsdam sentence corpus, *The Mind Research Repository (beta)*, (1).

[20] Brand, M., and V. Kettnaker (2000), Discovery and segmentation of activities in video, *IEEE PAMI*, *22*(8), 844–851.

[21] Browne, M., and J. Nesselroade (2005), Representing psychological processes with dynamic factor models: Some promising uses and extensions of arma time series models, *Psychometrics: A festschrift to Roderick P. McDonald*, pp. 415–452.

[22] Brückner, R., and B. Schuller (2012), Likability Classification – A not so Deep Neural Network Approach, in *Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, ISCA, ISCA, Portland, OR, 4 pages (acceptance rate: 52 %, IF* 1.05 (2010)).

[23] Bulteel, K., E. Ceulemans, R. J. Thompson, C. E. Waugh, I. H. Gotlib, F. Tuerlinckx, and P. Kuppens (2014), Decon: A tool to detect emotional concordance in multivariate time series data of emotional responding, *Biological psychology*, *98*, 29–42.

[24] Busso, C., and S. S. Narayanan (2007), Interrelation between speech and facial gestures in emotional utterances: a single subject study, *Audio, Speech, and Language Processing, IEEE Transactions on*, *15*(8), 2331–2347.

[25] Busso, C., Z. Deng, S. Yildirim, M. Bulut, C.-M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan (2004), Analysis of emotion recognition using facial expressions, speech and multimodal information, in *international conference on Multimodal interfaces*, pp. 205–211, ACM.

[26] Busso, C., S. Lee, and S. Narayanan (2007), Using neutral speech models for emotional speech analysis, *Proceedings of Interspeech*, pp. 2225–2228.

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

[28] Busso, C., S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost (2015), MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception, *IEEE Transactions on Affective Computing.*

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

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

[31] Cambria, E., B. Schuller, Y. Xia, and C. Havasi (2013), New avenues in opinion mining and sentiment analysis, *IEEE Intelligent Systems*, p. 1.

[32] Cattell, R. (1963), The structuring of change by p-technique and incremental r-technique, *Problems in measuring change*, pp. 167–198.

[33] Chan-Hon-Tong, A., C. Achard, and L. Lucat (2013), Deeply optimized hough transform: Application to action segmentation, in *ICIAP*, pp. 51–60, Springer.

[34] Chandrasekaran, C., A. Trubanova, S. Stillittano, A. Caplier, and A. A. Ghazanfar (2009), The natural statistics of audiovisual speech, *PLoS computational biology*, *5*(7), e1000,436.

[35] Chang, C., and C. Lin (2011), Libsvm: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST)*, *2*(3), 27.

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

[37] Chen, J., M. K. Leung, and Y. Gao (2003), Noisy logo recognition using line segment hausdorff distance, *Pattern recognition*, *36*(4), 943–955.

[38] Chen, L. S., and T. S. Huang (2000), Emotional expressions in audiovisual human computer interaction, in *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, vol. 1, pp. 423–426, IEEE.

[39] Chen, S., and Q. Jin (2015), Multi-modal dimensional emotion recognition using recurrent neural networks, in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pp. 49–56, ACM.

[40] Chen, T., F. X. Yu, J. Chen, Y. Cui, Y.-Y. Chen, and S.-F. Chang (2014), Object-based visual sentiment concept analysis and application, in *Proceedings of the ACM International Conference on Multimedia*, pp. 367–376, ACM.

[41] Cheng, Y., Q. Fan, S. Pankanti, and A. Choudhary (2014), Temporal sequence modeling for video event detection, in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pp. 2235–2242, IEEE.

[42] Chin, Y. H., S. H. Lin, C. H. Lin, E. Siahaan, A. Frisky, and J. C. Wang (2014), Emotion profile-based music recommendation, in *2014 7th International Conference on Ubi-Media Computing and Workshops*.

[43] Cohen, I., A. Garg, and T. S. Huang (2000), Emotion recognition from facial expressions using multilevel hmm, in *Neural Information Processing Systems (NIPS)*, Denver, CO.

[44] Cohn, J. F., and K. L. Schmidt (2004), The timing of facial motion in posed and spontaneous smiles, *IJWMIP*, *2*(02), 121–132.

[45] Ćosić, K., et al. (), Virtual reality adaptive stimulation in stress resistance training.

[46] Cotsaces, C., N. Nikolaidis, and I. Pitas (2006), Video shot detection and condensed representation. a review, *Signal Processing Magazine, IEEE*, *23*(2), 28–37.

[47] Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor (2001), Emotion recognition in human-computer interaction, *Signal Processing Magazine, IEEE*, *18*(1), 32–80.

[48] Crammer, K., and Y. Singer (2002), On the learnability and design of output codes for multiclass problems, *Machine Learning*, *47*(2-3), 201–233.

[49] Cvejic, E., J. Kim, and C. Davis (2011), Temporal relationship between auditory and visual prosodic cues, in *Twelfth Annual Conference of the International Speech Communication Association*.

[50] Datcu, D., and L. J. Rothkrantz (2014), Semantic audio-visual data fusion for automatic emotion recognition, *Emotion Recognition: A Pattern Analysis Approach*, pp. 411–435.

[51] De Gelder, B., and J. Vroomen (2000), The perception of emotions by ear and by eye, *Cognition & Emotion*, *14*(3), 289–311.

[52] Dietterich, T. G. (1998), Approximate statistical tests for comparing supervised classification learning algorithms, *Neural Computation*, *10*(7), 1895–1923.

[53] Ding, X., W.-S. Chu, F. D. L. Torre, J. F. Cohn, and Q. Wang (2013), Facial action unit event detection by cascade of tasks, in *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 2400–2407, IEEE.

[54] Ducci, L. (1981), Reaction times in the recognition of facial expressions of emotion., *Italian Journal of Psychology*.

[55] Duch, W., J. Biesiada, T. Winiarski, K. Grudzinski, and K. Grabczewski (2003), Feature selection based on information theory filters, in *Neural Networks and Soft Computing: Proceedings of the Sixth International Conference on Neural Networks and Soft Computing, Zakopane, Poland, June 11-15, 2002*, vol. 1, p. 173, Physica Verlag.

[56] Duchenne, G.-B., and R. A. Cuthbertson (1990), *The mechanism of human facial expression*, Cambridge university press.

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

[58] Ekman, P. (1999), Basic emotions, *Handbook of cognition and emotion*, *98*, 45–60.

[59] Ekman, P., and W. V. Friesen (1977), Facial action coding system.

[60] Ekman, P., W. V. Friesen, and P. Ellsworth (1972), *Emotion in the human face: Guide-lines for research and an integration of findings: guidelines for research and an integration of findings*, Pergamon.

[61] Ekman, P., R. J. Davidson, and W. V. Friesen (1990), The duchenne smile: Emotional expression and brain physiology: Ii., *Journal of personality and social psychology*, *58*(2), 342.

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

[63] Ellsworth, P. C., and C. A. Smith (1988), Shades of joy: Patterns of appraisal differentiating pleasant emotions, *Cognition & Emotion*, *2*(4), 301–331.

[64] Eyben, F., M. Wöllmer, and B. Schuller (2010), Opensmile: the munich versatile and fast open-source audio feature extractor, in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, ACM.

[65] Felty, T. (2005), Dynamic time warping [online], in *MATLAB Central File Exchange*.

[66] Freitas-Magalhães, A. (2012), Microexpression and macroexpression, *Encyclopedia of human behavior*, *2*, 173–183.

[67] Frijda, N. H., P. Kuipers, and E. Ter Schure (1989), Relations among emotion, appraisal, and emotional action readiness., *Journal of personality and social psychology*, *57*(2), 212.

[68] Gabrielsson, A. (2002), Emotion perceived and emotion felt: Same or different?, *Musicae Scientiae*, *5*(1 suppl), 123–147.

[69] Galata, A., N. Johnson, and D. Hogg (2001), Learning variable-length markov models of behavior, *CVIU*, *81*(3), 398–413, doi:http://dx.doi.org/10.1006/cviu.2000.0894.

[70] Garg, V., H. Kumar, and R. Sinha (2013), Speech based emotion recognition based on hierarchical decision tree with svm, blg and svr classifiers, in *Communications (NCC), 2013 National Conference on*, pp. 1–5, IEEE.

[71] Gaulin, S., and D. McBurney (2001), *Psychology: An evolutionary approach*, Prentice Hall.

[72] Gharavian, D., M. Sheikhan, A. Nazerieh, and S. Garoucy (2012), Speech emotion recognition using fcbf feature selection method and ga-optimized fuzzy artmap neural network, *Neural Computing and Applications*, *21*(8), 2115–2126.

[73] Gold, B., N. Morgan, and D. Ellis (2011), *Speech and audio signal processing: processing and perception of speech and music*, John Wiley & Sons.

[74] Grandjean, D., D. Sander, and K. Scherer (2008), Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization, *Consciousness and cognition*, *17*(2), 484.

[75] Granström, B., D. House, and M. Lundeberg (1999), Prosodic cues in multimodal speech perception, in *Proceedings of the International Congress of Phonetic Sciences (ICPhS99)*, pp. 655–658.

[76] Graves, A. (2012), *Supervised sequence labelling with recurrent neural networks*, vol. 385, Springer.

[77] Grimm, M., K. Kroschel, and S. Narayanan (2007), Support vector regression for automatic recognition of spontaneous emotions in speech, in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–1085, IEEE.

[78] Gross, J. J. (2002), Emotion regulation: Affective, cognitive, and social consequences, *Psychophysiology*, *39*(3), 281–291.

[79] Gunes, H., B. Schuller, M. Pantic, and R. Cowie (2011), Emotion representation, analysis and synthesis in continuous space: A survey, in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pp. 827–834, IEEE.

[80] Hadar, U., T. Steiner, and F. C. Rose (1983), Involvement of head movement in speech production and its implications for language pathology., *Advances in neurology*, *42*, 247–261.

[81] Haggard, E. A., and K. S. Isaacs (1966), Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy, in *Methods of research in psychotherapy*, pp. 154–165, Springer.

[82] Hakim, A., S. Marsland, and H. W. Guesgen (2012), A robust joint face model for human emotion recognition, in *Proceedings of the 27th Conference on Image and Vision Computing New Zealand*, pp. 352–357, ACM.

[83] Hamaker, E., E. Ceulemans, R. Grasman, and F. Tuerlinckx (2015), Modeling affect dynamics: State of the art and future challenges, *Emotion Review*, *7*(4), 316–322.

[84] Haq, S., and P. J. Jackson (2010), *Machine Audition: Principles, Algorithms and Systems*, chap. Multimodal Emotion Recognition, pp. 398–423, IGI Global, Hershey PA.

[85] Heilman, K. M., D. Bowers, and E. Valenstein (1993), Emotional disorders associated with neurological diseases, *Clinical neuropsychology*, *3*, 461–97.

[86] Helander, M. G. (2014), *Handbook of human-computer interaction*, Elsevier.

[87] Hinton, G. (2002), Training products of experts by minimizing contrastive divergence, *Neural computation*, *14*(8), 1771–1800.

[88] Hinton, G., S. Osindero, and Y.-W. Teh (2006), A fast learning algorithm for deep belief nets, *Neural Computation*, *18*(7), 1527–1554.

[89] Hoai, M., Z.-Z. Lan, and F. De la Torre (2011), Joint segmentation and classification of human actions in video, in *IEEE CVPR*, pp. 3265–3272, IEEE.

[90] Hoey, J. (2001), Hierarchical unsupervised learning of facial expression categories, in *IEEE Workshop on Detection and Recognition of Events in Video*, pp. 99–106, Vancouver, BC.

[91] Hollenstein, T., and D. Lanteigne (2014), Models and methods of emotional concordance, *Biological psychology*, *98*, 1–5.

[92] Holt, G., M. Reinders, and E. Hendriks (2007), Multi-dimensional dynamic time warping for gesture recognition, in *Thirteenth annual conference of the Advanced School for Computing and Imaging*.

[93] Hornak, J., E. Rolls, and D. Wade (1996), Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage, *Neuropsychologia*, *34*(4), 247–261.

[94] Huang, D., Y. Wang, S. Yao, and F. De la Torre (2014), Sequential max-margin event detectors, in *ECCV*.

[95] Hugenberg, K. (2005), Social categorization and the perception of facial affect: target race moderates the response latency advantage for happy faces., *Emotion*, *5*(3), 267.

[96] Hussain, M. S., S. K. D'Mello, and R. A. Calvo (2014), 25 research and development tools in affective computing, *The Oxford Handbook of Affective Computing*, p. 349.

[97] Ito, A., X. Wang, M. Suzuki, and S. Makino (2005), Smile and laughter recognition using speech processing and face recognition from conversation video, in *Cyberworlds. International Conference on*, pp. 8–pp, IEEE.

[98] Jacko, J. A. (2012), *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*, CRC press.

[99] James, W. (1884), What is an emotion?, *Mind*, *9*, 17.

[100] Jennings, P. A., and M. T. Greenberg (2009), The prosocial classroom: Teacher social and emotional competence in relation to student and classroom outcomes, *Review of educational research*, *79*(1), 491–525.

[101] Jeon, J. H., R. Xia, and Y. Liu (2011), Sentence level emotion recognition based on decisions from subsentence segments, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4940–4943, Prague, Czech Republic.

[102] Jou, B., S. Bhattacharya, and S.-F. Chang (2014), Predicting viewer perceived emotions in animated gifs, in *Proceedings of the ACM International Conference on Multimedia*, pp. 213–216, ACM.

[103] Kächele, M., M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker (2014), Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression, *depression*, *1*, 1.

[104] Kalberer, G., P. Müller, and L. Van Gool (2002), Speech animation using viseme space., in *Vision, Modeling, and Visualization (VMV)*, pp. 463–470, Erlangen, Germany.

[105] Kalinli, O. (2012), Automatic phoneme segmentation using auditory attention features., in *INTERSPEECH*.

[106] Kaplan, S. H., S. Greenfield, and J. E. Ware Jr (1989), Assessing the effects of physician-patient interactions on the outcomes of chronic disease., *Medical care*, *27*(3), S110–S127.

[107] Keltner, D., and P. Ekman (2000), Emotion: an overview, *Encyclopedia of psychology*, *3*, 162–167.

[108] Kennedy, L. S., and D. P. Ellis (2003), Pitch-based emphasis detection for characterization of meeting recordings, in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pp. 243–248, IEEE.

[109] Kerr, S. L., and J. M. Neale (1993), Emotion perception in schizophrenia: specific deficit or further evidence of generalized poor performance?, *Journal of abnormal psychology*, *102*(2), 312.

[110] Keshet, J., S. Shalev-Shwartz, and Y. Singer (2005), Phoneme alignment based on discriminative learning.

[111] Kim, K. H., S. Bang, and S. Kim (2004), Emotion recognition system using short-term monitoring of physiological signals, *Medical and biological engineering and computing*, *42*(3), 419–427.

[112] Kim, Y., and E. Mower Provost (2013), Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3677–3681, IEEE.

[113] Kim, Y., and E. Mower Provost (2014), Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition, in *Proceedings of the ACM International Conference on Multimedia (ACM MM'14)*.

[114] Kim, Y., and E. Mower Provost (2015), Emotion recognition during speech using dynamics of multiple regions of the face, *ACM Trans. Multimedia Comput. Commun. Appl.*, *12*(1s), 25:1–25:23, doi:10.1145/2808204.

[115] Kim, Y., and E. Mower Provost (2016, In Submission), Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions, *ACM International Conference on Multimodal Interaction (ICMI)*.

[116] Kim, Y., and E. Mower Provost (2016, In Submission), Isla: A framework for controlling sources of modulation in audio-visual affective behavior, *IEEE Transactions on Affective Computing (IEEE TAC)*.

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

[118] Kim, Y., J. Chen, M.-C. Chang, X. Wang, E. Mower Provost, and S. Lyu (2015), Modeling transition patterns between events for temporal human action segmentation and classification, in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*, Ljubljana, Slovenia.

[119] Kipp, M., and J.-C. Martin (2009), Gesture and emotion: Can basic gestural form features discriminate emotions?, in *3rd International Conference onAffective Computing and Intelligent Interaction and Workshops. ACII 2009.*, pp. 1–8, IEEE.

[120] Kirouac, G., and F. Y. Doré (1983), Accuracy and latency of judgment of facial expressions of emotions, *Perceptual and motor skills*, *57*(3), 683–686.

[121] Kleinsmith, A., and N. Bianchi-Berthouze (2013), Affective body expression perception and recognition: A survey, *Affective Computing, IEEE Transactions on*, *4*(1), 15–33.

[122] Koelstra, S., M. Pantic, and I. Patras (2010), A dynamic texture-based approach to recognition of facial actions and their temporal models, *IEEE PAMI*, *32*(11), 1940–1954.

[123] Kohler, C. G., J. B. Walker, E. A. Martin, K. M. Healey, and P. J. Moberg (2009), Facial emotion perception in schizophrenia: a meta-analytic review, *Schizophrenia bulletin*, p. sbn192.

[124] Koolagudi, S. G., N. Kumar, and K. S. Rao (2011), Speech emotion recognition using segmental level prosodic analysis, in *Devices and Communications (ICDeCom), 2011 International Conference on*, pp. 1–5, IEEE.

[125] Koprinska, I., and S. Carrato (2001), Temporal video segmentation: A survey, *Signal processing: Image communication*, *16*(5), 477–500.

[126] Krizhevsky, A., I. Sutskever, and G. Hinton (2012), Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems 25*, pp. 1106–1114.

[127] Lang, A., K. Dhillon, and Q. Dong (1995), The effects of emotional arousal and valence on television viewers' cognitive capacity and memory, *Journal of Broadcasting & Electronic Media*, *39*(3), 313–327.

[128] Laptev, I., M. Marszalek, C. Schmid, and B. Rozenfeld (2008), Learning realistic human actions from movies, in *IEEE CVPR*, pp. 1–8, IEEE.

[129] Lawrence, I., and K. Lin (1989), A concordance correlation coefficient to evaluate reproducibility, *Biometrics*, pp. 255–268.

[130] Lee, C.-C., E. Mower, C. Busso, S. Lee, and S. Narayanan (2009), Emotion recognition using a hierarchical binary decision tree approach, in *Interspeech*, pp. 320–323, Jeju Island, South Korea.

[131] Lee, C.-C., E. Mower, C. Busso, S. Lee, and S. Narayanan (2011), Emotion recognition using a hierarchical binary decision tree approach, *Speech Communication*, *53*(9), 1162–1171.

[132] Lee, C. M., and S. S. Narayanan (2005), Toward detecting emotions in spoken dialogs, *Speech and Audio Processing, IEEE Transactions on*, *13*(2), 293–303.

[133] Lee, C. M., S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan (2004), Emotion recognition based on phoneme classes., in *INTERSPEECH*, pp. 205–211.

[134] Lee, H., C. Ekanadham, and A. Ng (2008), Sparse deep belief net model for visual area v2, *Advances in Neural Information Processing Systems(NIPS)*, *20*, 873–880.

[135] Lee, H., R. Grosse, R. Ranganath, and A. Y. Ng (2011), Unsupervised learning of hierarchical representations with convolutional deep belief networks, *Communications of the ACM*, *54*(10), 95–103.

[136] Lee, J.-G., J. Han, and K.-Y. Whang (2007), Trajectory clustering: a partition-and-group framework, in *ACM SIGMOD International Conference on Management of Data*, pp. 593–604, Beijing, China.

[137] Lu, Y., I. Cohen, X. Zhou, and Q. Tian (2007), Feature selection using principal feature analysis, in *ACM International Conference on Multimedia*, pp. 301–304, Augsburg, Germany.

[138] Lucey, P., T. Martin, and S. Sridharan (2004), Confusability of phonemes grouped according to their viseme classes in noisy environments, in *Australian International Conference on Speech Science & Tech*, pp. 265–270.

[139] Lugger, M., M.-E. Janoir, and B. Yang (2009), Combining classifiers with diverse feature sets for robust speaker independent emotion recognition, in *Signal Processing Conference, 2009 17th European*, pp. 1225–1229, IEEE.

[140] Mansoorizadeh, M., and N. M. Charkari (2010), Multimodal information fusion application to human emotion recognition from face and speech, *Multimedia Tools and Applications*, *49*(2), 277–297.

[141] Mariooryad, S., and C. Busso (2012), Factorizing speaker, lexical and emotional variabilities observed in facial expressions, in *IEEE International Conference on Image Processing (ICIP 2012)*, pp. 2605–2608, Orlando, FL, USA.

[142] Mariooryad, S., and C. Busso (2013), Feature and model level compensation of lexical content for facial emotion recognition, in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013)*, Shanghai, China, doi: 10.1109/FG.2013.6553752.

[143] Mariooryad, S., and C. Busso (2015), Facial expression recognition in the presence of speech using blind lexical compensation, *IEEE Transactions on Affective Computing*, *PP*(99), 1–1, doi:10.1109/TAFFC.2015.2490070.

[144] Meeren, H. K., C. C. van Heijnsbergen, and B. de Gelder (2005), Rapid perceptual integration of facial expression and emotional body language, *Proceedings of the National Academy of Sciences of the United States of America*, *102*(45), 16,518–16,523.

[145] Mehrabian, A. (1981), Silent messages: Implicit communication of emotion and attitude, *Belmont, CA: Wadsworth*.

[146] Meng, H., and N. Bianchi-Berthouze (2011), Naturalistic affective expression classification by a multi-stage approach based on hidden markov models, in *Affective Computing and Intelligent Interaction*, pp. 378–387, Springer.

[147] Messinger, D. S., A. Fogel, and K. L. Dickson (2001), All smiles are positive, but some smiles are more positive than others., *Developmental Psychology*, *37*(5), 642.

[148] Metallinou, A., C. Busso, S. Lee, and S. Narayanan (2010), Visual emotion recognition using compact facial representations and viseme information, in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 2474–2477, IEEE.

[149] Metallinou, A., C. Busso, S. Lee, and S. Narayanan (2010), Visual emotion recognition using compact facial representations and viseme information, in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 2474–2477, IEEE.

[150] Metallinou, A., S. Lee, and S. Narayanan (2010), Decision level combination of multiple modalities for recognition and analysis of emotional expression, in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 2462–2465, Dallas, TX.

[151] Metallinou, A., M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan (2012), Context-sensitive learning for enhanced audiovisual emotion classification, *Affective Computing, IEEE Transactions on*, *3*(2), 184–198.

[152] Metallinou, A., A. Katsamanis, and S. Narayanan (2013), Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information, *Image and Vision Computing*, *31*(2), 137–152.

[153] Mohamed, A., G. Dahl, and G. Hinton (2012), Acoustic modeling using deep belief networks, *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 14–22.

[154] Morgan, N. (2012), Deep and wide: Multiple layers in automatic speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 7–13.

[155] Mower, E., and S. Narayanan (2011), A hierarchical static-dynamic framework for emotion classification, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2372–2375, Prague, Czech Republic.

[156] Mower, E., M. J. Mataric, and S. Narayanan (2009), Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information, *Multimedia, IEEE Transactions on*, *11*(5), 843–855.

[157] Mower, E., A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan (2009), Interpreting ambiguous emotional expressions, in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, pp. 1–8, IEEE.

[158] Mower, E., M. J. Mataric, and S. Narayanan (2011), A framework for automatic human emotion classification using emotion profiles, *Audio, Speech, and Language Processing, IEEE Transactions on*, *19*(5), 1057–1070.

[159] Mower Provost, E. (2013), Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3682–3686, Vancouver, BC.

[160] Mower Provost, E., and S. Narayanan (2012), Simplifying emotion classification through emotion distillation, in *Proceedings of APSIPA Annual Summit and Conference.*

[161] Munhall, K. G., J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson (2004), Visual prosody and speech intelligibility head movement improves auditory speech perception, *Psychological science*, *15*(2), 133–137.

[162] Myers, C., L. Rabiner, and A. Rosenberg (1980), Performance tradeoffs in dynamic time warping algorithms for isolated word recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, *28*(6), 623 – 635, doi: 10.1109/TASSP.1980.1163491.

[163] Narayanan, S., and P. G. Georgiou (2013), Behavioral signal processing: Deriving human behavioral informatics from speech and language, *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, *101*(5), 1203.

[164] Nefian, A. V., L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy (2002), A coupled hmm for audio-visual speech recognition, in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 2, pp. II–2013, IEEE.

[165] Nesselroade, J., and P. Molenaar (2003), Quantitative models for developmental processes, *Handbook of developmental psychology*, pp. 622–639.

[166] Ngiam, J., A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng (2011), Multimodal deep learning, in *International Conference on Machine Learning (ICML)*, pp. 689–696.

[167] Nguyen, M. H., L. Torresani, F. De la Torre, and C. Rother (2009), Weakly supervised discriminative localization and classification: a joint learning process, in *IEEE ICCV*, pp. 1925–1932, IEEE.

[168] Nicolle, J., V. Rapp, K. Bailly, L. Prevost, and M. Chetouani (2012), Robust continuous prediction of human emotions using multiscale dynamic cues, in *ACM international conference on Multimodal interaction*, pp. 501–508, ACM.

[169] Niebles, J. C., C.-W. Chen, and L. Fei-Fei (2010), Modeling temporal structure of decomposable motion segments for activity classification, in *ECCV*, pp. 392–405, Springer.

[170] Oatley, K., D. Keltner, and J. M. Jenkins (2006), *Understanding emotions* ., Blackwell publishing.

[171] Oh, S. M., J. M. Rehg, T. Balch, and F. Dellaert (2008), Learning and inferring motion patterns using parametric segmental switching linear dynamic systems, *IJCV*, *77*(1-3), 103–124.

[172] Ortony, A., G. L. Clore, and A. Collins (1990), *The cognitive structure of emotions*, Cambridge university press.

[173] Ozonoff, S., B. F. Pennington, and S. J. Rogers (1990), Are there emotion perception deficits in young autistic children?, *Journal of Child Psychology and Psychiatry*, *31*(3), 343–361.

[174] Pang, B., and L. Lee (2008), Opinion mining and sentiment analysis, *Foundations and trends in information retrieval*, *2*(1-2), 1–135.

[175] Pantic, M., and M. S. Bartlett (2007), Machine analysis of facial expressions.

[176] Pantic, M., and L. J. Rothkrantz (2003), Toward an affect-sensitive multimodal human-computer interaction, *Proceedings of the IEEE*, *91*(9), 1370–1390.

[177] Pantic, M., G. Caridakis, E. André, J. Kim, K. Karpouzis, and S. Kollias (2011), Multimodal emotion recognition from low-level cues, *Emotion-Oriented Systems*, pp. 115–132.

[178] Paul, D. (1990), Speech recognition using hidden markov models, *The Lincoln Laboratory Journal*, *3*(1), 41–62.

[179] Pell, M. D., and S. A. Kotz (2011), On the time course of vocal emotion recognition, *PLoS ONE*, *6*(11), doi:10.1371/journal.pone.0027256.

[180] Pell, M. D., and S. A. Kotz (2011), On the time course of vocal emotion recognition, *PLoS One*, *6*(11), e27,256.

[181] Pentland, A., and A. Liu (1999), Modeling and prediction of human behavior, *Neural computation*, *11*(1), 229–242.

[182] Petridis, S., H. Gunes, S. Kaltwang, and M. Pantic (2009), Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities, in *international conference on Multimodal interfaces*, pp. 23–30, ACM.

[183] Phillips, M., C. Ladouceur, and W. Drevets (2008), A neural model of voluntary and automatic emotion regulation: implications for understanding the pathophysiology and neurodevelopment of bipolar disorder, *Molecular psychiatry*, *13*(9), 833–857.

[184] Picard, R. W. (1999), Affective computing for hci., in *HCI (1)*, pp. 829–833.

[185] Polikovsky, S., Y. Kameda, and Y. Ohta (2009), Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor, in *International Conference on Crime Detection and Prevention (ICDP)*, pp. 1–6, IET.

[186] Polzehl, T., S. Sundaram, H. Ketabdar, M. Wagner, and F. Metze (2009), Emotion classification in children's speech using fusion of acoustic and linguistic features, *Proceedings of INTERSPEECH-2009, Brighton, UK*, pp. 340–343.

[187] Poppe, R. (2010), A survey on vision-based human action recognition, *Image and vision computing*, *28*(6), 976–990.

[188] Qiao, Y., N. Shimomura, and N. Minematsu (2008), Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons, in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 3989–3992, IEEE.

[189] Rani, P., N. Sarkar, and C. Liu (2005), Maintaining optimal challenge in computer games through real-time physiological feedback, in *Proceedings of the 11th international conference on human computer interaction*, vol. 58.

[190] Rao, K. S., and S. G. Koolagudi (2013), *Robust Emotion Recognition using Spectral and Prosodic Features*, Springer Science & Business Media.

[191] Ringeval, F., and M. Chetouani (2008), Exploiting a vowel based approach for acted emotion recognition, in *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction*, pp. 243–254, Springer.

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

[193] Roseman, I. J., and C. A. Smith (2001), Appraisal theory: Overview, assumptions, varieties, controversies.

[194] Rui, Y., A. Gupta, and A. Acero (2000), Automatically extracting highlights for tv baseball programs, in *Proceedings of the eighth ACM international conference on Multimedia*, pp. 105–115, ACM.

[195] Russell, J. A., and A. Mehrabian (1977), Evidence for a three-factor theory of emotions, *Journal of research in Personality*, *11*(3), 273–294.

[196] Sakoe, H., and S. Chiba (1978), Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, *26*(1), 43–49.

[197] Sánchez-Lozano, E., P. Lopez-Otero, L. Docio-Fernandez, E. Argones-Rúa, and J. L. Alba-Castro (2013), Audiovisual three-level fusion for continuous estimation of russell's emotion circumplex, in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pp. 31–40, ACM.

[198] Sandbach, G., S. Zafeiriou, M. Pantic, and D. Rueckert (2011), A dynamic approach to the recognition of 3d facial expressions and their temporal models, in *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, pp. 406–413, Santa Barbara, CA.

[199] Sapp, B., and B. Taskar (2013), Multimodal decomposable models for human pose estimation, in *IEEE CVPR*, pp. 3674–3681.

[200] Savran, A., H. Cao, M. Shah, A. Nenkova, and R. Verma (2012), Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering, in *ACM International Conference on Multimodal Interaction*, pp. 485–492, Santa Monica, CA.

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

[202] Scherer, K. R., A. Schorr, and T. Johnstone (2001), *Appraisal processes in emotion: Theory, methods, research*, Oxford University Press.

[203] Schlosberg, H. (1954), Three dimensions of emotion., *Psychological review*, *61*(2), 81.

[204] Schöner, G., H. Haken, and J. Kelso (1986), A stochastic theory of phase transitions in human hand movement, *Biological cybernetics*, *53*(4), 247–257.

[205] Schuller, B., and G. Rigoll (2006), Timing levels in segment-based speech emotion recognition., in *INTERSPEECH*, pp. 1818–1821, Pittsburgh, Pennsylvania.

[206] Schuller, B., G. Rigoll, and M. Lang (2003), Hidden markov model-based speech emotion recognition, in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong.

[207] Schuller, B., R. Müller, M. Lang, and G. Rigoll (2005), Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles, in *Proc. Interspeech*, pp. 805–808.

[208] Schuller, B., S. Steidl, and A. Batliner (2009), The interspeech 2009 emotion challenge, in *Proc. Interspeech*, pp. 312–315.

[209] Schuller, B., A. Batliner, S. Steidl, and D. Seppi (2011), Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge, *Speech Communication*, *53*(9), 1062–1087.

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

[211] Schuller, B., et al. (2007), The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals., in *INTERSPEECH*, pp. 2253–2256, Citeseer.

[212] Schuller, B., et al. (2012), The INTERSPEECH 2012 Speaker Trait Challenge, in *Proceedings INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association*, ISCA, ISCA, Portland, OR, 4 pages (acceptance rate: 52 %, IF* 1.05 (2010), 67 citations).

[213] Schuller, B., et al. (2013), The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism.

[214] Sebe, N., I. Cohen, T. Gevers, and T. S. Huang (2006), Emotion recognition based on joint visual and audio cues, in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, pp. 1136–1139, IEEE.

[215] Shah, M., D. G. Cooper, H. Cao, R. C. Gur, A. Nenkova, and R. Verma (2013), Action unit models of facial expression of emotion in the presence of speech, in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pp. 49–54, IEEE.

[216] Shimodaira, H., K.-i. Noma, M. Nakai, and S. Sagayama (2001), Dynamic time-alignment kernel in support vector machine, in *Neural Information Processing Systems (NIPS)*, vol. 14, pp. 921–928, Vancouver, BC.

[217] Sivaram, G., and H. Hermansky (2012), Sparse multilayer perceptron for phoneme recognition, *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(1), 23–29.

[218] Smolensky, P. (1986), Information processing in dynamical systems: Foundations of harmony theory, *Parallel distributed processing: Explorations in the microstructure of cognition*, *1*, 194–281.

[219] Sohn, K., D. Jung, H. Lee, and A. Hero (2011), Efficient learning of sparse, distributed, convolutional feature representations for object recognition, in *IEEE International Conference on Computer Vision (ICCV),*, pp. 2643–2650.

[220] Song, H., and E. Ferrer (2009), State-space modeling of dynamic psychological processes via the kalman smoother algorithm: Rationale, finite sample properties, and applications, *Structural Equation Modeling*, *16*(2), 338–363.

[221] Steidl, S., M. Levit, A. Batliner, E. Nöth, and H. Niemann (2005), " of all things the measure is man": Automatic classification of emotions and interlabeler consistency., in *ICASSP (1)*, pp. 317–320, Citeseer.

[222] Steunebrink, B., et al. (2010), The logical structure of emotions.

[223] Stuhlsatz, A., C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller (2011), Deep neural networks for acoustic emotion recognition: raising the benchmarks, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5688–5691, Prague, Czech Republic.

[224] Tamir, M., C. Mitchell, and J. J. Gross (2008), Hedonic and instrumental motives in anger regulation, *Psychological Science*, *19*(4), 324–328.

[225] Tang, K., L. Fei-Fei, and D. Koller (2012), Learning latent temporal structure for complex event detection, in *IEEE CVPR*, pp. 1250–1257, IEEE.

[226] Tang, Y., and C. Eliasmith (2010), Deep networks for robust visual recognition, in *International Conference on Machine Learning*, vol. 28, Citeseer.

[227] Tappert, C., C. Suen, and T. Wakahara (1990), The state of the art in online handwriting recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(8), 787 –808, doi:10.1109/34.57669.

[228] Taylor, G. W., G. E. Hinton, and S. T. Roweis (2006), Modeling human motion using binary latent variables, in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1345–1352.

[229] Tijs, T., D. Brokken, and W. IJsselsteijn (2008), Creating an emotionally adaptive game, in *International Conference on Entertainment Computing*, pp. 122–133, Springer.

[230] Toledano, D. T., L. A. H. Gómez, and L. V. Grande (2003), Automatic phonetic segmentation, *Speech and Audio Processing, IEEE Transactions on*, *11*(6), 617–625.

[231] Tracy, J. L., and R. W. Robins (2008), The automaticity of emotion recognition., *Emotion (Washington, D.C.)*, *8*(1), 81–95, doi:10.4092/jsre.18.135.

[232] Tronick, E. Z. (1989), Emotions and emotional communication in infants., *American psychologist*, *44*(2), 112.

[233] Turaga, P., R. Chellappa, V. S. Subrahmanian, and O. Udrea (2008), Machine recognition of human activities: A survey, *Circuits and Systems for Video Technology, IEEE Transactions on*, *18*(11), 1473–1488.

[234] Valstar, M., and M. Pantic (2012), Fully automatic recognition of the temporal phases of facial actions, *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, *42*(1), 28–43, doi:10.1109/TSMCB.2011.2163710.

[235] Valstar, M. F., and M. Pantic (2007), Combined support vector machines and hidden markov models for modeling facial action temporal dynamics, in *Human–Computer Interaction*, pp. 118–127, Springer.

[236] Ververidis, D., and C. Kotropoulos (2008), Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition, *Signal Processing*, *88*(12), 2956–2970.

[237] Vogt, T., and E. André (2005), Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition, in *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 474–477.

[238] Wagner, J., J. Kim, and E. André (2005), From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification, in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pp. 940–943, IEEE.

[239] Wagner, J., E. Andre, F. Lingenfelser, and J. Kim (2011), Exploring fusion methods for multimodal emotion recognition with missing data, *IEEE Transactions on Affective Computing*, *2*(4), 206–218.

[240] Wan, S., and J. Aggarwal (2014), Spontaneous facial expression recognition: A robust metric learning approach, *Pattern Recognition*, *47*(5), 1859–1868.

[241] Warren, W. H. (2006), The dynamics of perception and action., *Psychological review*, *113*(2), 358.

[242] Weinland, D., R. Ronfard, and E. Boyer (2011), A survey of vision-based methods for action representation, segmentation and recognition, *CVIU*, *115*(2), 224–241.

[243] Wilson, T., J. Wiebe, and P. Hoffmann (2005), Recognizing contextual polarity in phrase-level sentiment analysis, in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 347–354, Association for Computational Linguistics.

[244] Wimmer, M., B. Schuller, D. Arsic, G. Rigoll, and B. Radig (2008), Low-level fusion of audio and video feature for multi-modal emotion recognition, in *3rd International Conference on Computer Vision Theory and Applications. VISAPP*, vol. 2, pp. 145–151.

[245] Wöllmer, M., F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie (2008), Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies., in *INTERSPEECH*, vol. 2008, pp. 597–600.

[246] Wöllmer, M., A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan (2010), Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling, in *Interspeech*, pp. 2362–2365.

[247] Wollmer, M., F. Eyben, B. Schuller, and G. Rigoll (2011), A multi-stream asr framework for blstm modeling of conversational speech, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4860–4863, Prague, Czech Republic.

[248] Wood, F., C. Archambeau, J. Gasthaus, L. James, and Y. W. Teh (2009), A stochastic memoizer for sequence data, in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1129–1136, ACM.

[249] Wu, S., T. H. Falk, and W.-Y. Chan (2011), Automatic speech emotion recognition using modulation spectral features, *Speech Communication*, *53*(5), 768–785.

[250] Zhou, F., F. De la Torre, and J. F. Cohn (2010), Unsupervised discovery of facial events, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2581, San Francisco, CA.

[251] Zhou, F., F. De la Torre, and J. K. Hodgins (2013), Hierarchical aligned cluster analysis for temporal clustering of human motion, *IEEE PAMI*, *35*(3), 582–596.