USC-SIPI REPORT #412

Emotions in Engineering: Methods for the Interpretation of Ambiguous Emotional Content

by

Emily K. Mower

December 2010

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

Emotions in Engineering: Methods for the Interpretation of Ambiguous Emotional Content

by

Emily K. Mower

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

December 2010

Dedication

I would like to dedicate this Dissertation to my family and friends.

To my parents, sister and brother, thank you. You have listened to me talk about my work ad nauseum for years and have always managed to sound interested and excited. To my Damen, the amount of support that you have given me has meant the world to me. Thank you for making my thesis completion process wonderful, exciting, and full of promise for the future. To my friends, it was so much fun to go on this journey with you guys. I look forward to our many adventures and discoveries!

Acknowledgements

I would like to thank my advisors, Dr. Shrikanth Narayanan and Dr. Maja Matarić for their guidance and support during my time at the University of Southern California. They have both provided me with a fantastic environment in which to grow and learn as a researcher and I could not be more grateful. I would also like to thank Dr. Sungbok Lee for his feedback and suggestions. His commitment to the research ideal is inspirational and has changed the way I approach research problems.

Many thanks to my commitee members, Dr. C.-C. Jay Kuo and Dr. Fei Sha. Thank you so much for your research direction suggestions; they have been fascinating.

Thanks also go to Dr. Panayiotis Georgiou and Dr. May-chen Kuo for the template upon which this thesis was built.

Finally, I would also like to thank my labmates and classmates who have provided me with invaluable advice, stimulating discussions, and wonderful adventures.

Table of Contents

List Of Tables

List Of Figures

Abstract

Emotion has intrigued researchers for generations. This fascination has permeated the engineering community, motivating the development of affective computational models for classification. However, human emotion remains notoriously difficult to interpret both because of the mismatch between the emotional cue generation (the speaker) and perception (the observer) processes and because of the presence of complex emotions, emotions that contain shades of multiple affective classes. Proper representations of emotion would ameliorate this problem by introducing multidimensional characterizations of the data that permit the quantification and description of the varied affective components of each utterance. Currently, the mathematical representation of emotion is an area that is under explored.

Research in emotion expression and perception provides a complex and human-centered platform for the integration of machine learning techniques and multimodal signal processing towards the design of interpretable data representations. The focus of this dissertation is to provide a computational description of human emotion perception and combine this knowledge with the information gleaned from emotion classification experiments to develop a mathematical characterization capable of interpreting naturalistic expressions of emotion utilizing a data representation method called Emotion Profiles.

The analysis of human emotion perception provides an understanding of how humans integrate audio and video information during emotional presentations. The goals of this work are to determine how audio and video information interact during the human emotional evaluation process and a to identify a subset of the features that contribute to specific types of emotion perception. We identify perceptually-relevant feature modulations and multi-modal feature integration trends using statistical analyses over the evaluator reports.

The trends in evaluator reports are analyzed using emotion classification. We study evaluator performance using a combination of Hidden Markov Models (HMM) and Naïve Bayes (NB) classification. The HMM classification is used to predict individual evaluator emotional assessments. The NB classification provides an estimate of the consistency of the evaluator's mental model of emotion. We demonstrate that evaluator reports created by evaluators with higher levels of estimated consistency are more accurately predicted than evaluator reports from evaluators that are less consistent.

The insights gleaned from the emotion perception and classification studies are aggregated to develop a new emotional representation scheme, called Emotion Profiles (EP). The design of the EPs is predicated on the knowledge that naturalistic emotion expressions can be approximately described using one or more labels from a set of basic emotions. Emotion profiles (EPs) are a quantitative measure expressing the degree of the presence or absence of a set of basic emotions within an expression. They avoid the need for a hard-labeled assignment by instead providing a method for describing the shades of emotion present in an utterance. These profiles can be used to determine a most likely assignment for an utterance, to map out the evolution of the emotional tenor of an interaction, or

to interpret utterances that have multiple affective components. The Emotion-Profile technique is able to accurately identify the emotion of utterances with definable ground truths (emotions with an evaluator consensus) and is able to interpret the affective content of emotions with ambiguous emotional content (no evaluator consensus), emotions that are typically discarded during classification tasks.

The algorithms and statistical analyses presented in this work are tested using two databases. The first database is a combination of synthetic (facial information) and natural human (vocal information) cues. The affective content of the two modalities is either matched (congruent presentation) or mismatched (conflicting presentation). The congruent and conflicting presentations are used to assess the affective perceptual relevance of both individual modalities and the specific feature modulations of those modalities. The second database is an audio-visual + motion-capture database collected at the University of Southern California, the USC IEMOCAP database. This database is used to assess the efficacy of the EP technique for quantifying the emotional content of an utterance. The IEMOCAP database is also used in the classification studies to determine how well individual evaluators can be modeled and how accurately discrete emotional labels (e.g., angry, happy, sad, neutral) can be predicted given audio and motion-capture feature information.

The future directions of this work include the unification of the emotion perception, classification, and quantification studies. The classification framework will be extended to include evaluator-specific features (an extension of the emotion perception studies) and temporal features based on EP estimates. This unification will produce a classification

framework that is not only more effective than previous versions, but is also able to adapt to specific user emotion production and perception styles.

Chapter 1

Introduction

Interactive technologies are becoming increasingly prevalent in society. These technologies range from simple hand-held devices to fully embodied robotic agents. Each of these interfaces contains underlying protocols that dictate the interaction behaviors upon which these technologies rely. The protocols range from simple if-then loops to complicated emotion and drive-based internal representations of agent state. Users observe manifestations of agent state through the conveyed interactive behaviors.

Increasingly, these behaviors are modulated by emotional qualities [7,51,99,111]. The animators of Disney have long understood the importance of endowing their characters with appropriate emotional attributes. In "The Illusion of Life," Thomas and Johnston assert that, "From the earliest days, it has been the portrayal of emotions that has given the Disney characters the illusion of life [113]." According to Joseph Bates it is this "illusion of life" that creates a believable character, one for which the audience is willing to suspend its disbelief [7].

The creation of reliably recognized emotion expressions requires an understanding of underlying social expectations. Agents incapable of creating emotional expressions

that meet (or ideally exceed) the baseline social expectations may not be capable of maintaining long-term user interactions [52,96,110].

However, synthetic emotion is not expressed in a vacuum. Consequently, an agent that is able to reliably produce, but not recognize a human interaction partner's emotion state may still be unable to meet its interaction goals. Proper interpretation of user state hinges on more than task awareness, it also depends on the more subtle qualities of user expression [35,36,94]. In fact, there is evidence in the psychological community that even humans who lack proper emotional production and perception abilities cannot form or maintain relationships [40].

Human emotion perception is a complex and dynamic process. There is much debate within the psychology and neuroscience communities regarding the true definition of emotion and how it should be quantified and analyzed. This work does not seek to develop a new definition of emotion, rather it uses currently accepted definitions of emotion to understand how we can develop models capable of capturing the modulations present in emotional expressions. This work presents an engineering approach, focusing on the development of techniques to estimate high-level emotion labels from the low-level feature-level modulations of utterances. While motivated by psychological theories of emotion, this mapping does not necessarily seek to produce a true human-centric model of either the human emotion perception or production processes. Rather, it seeks to provide techniques that can inform the design of emotional audio-visual behavior.

Emotional models for expression and user state interpretation are currently used in diagnostics, interactions with, and interventions for children with autism [42,62,70,73]. In these scenarios the computer agent acts as either a tutor or a peer for the child. The

computer agent must be able to do more than convey the material, it also must be able to maintain the interest of the child, due to the sensitive nature of the domain. As a result, the agent needs to provide more than basic instructive information, it must also provide motivation, empathy, and encouragement while recognizing signs of child frustration, interest, and boredom. This necessitates the development of systems designed to both recognize human affect and to respond appropriately to human affect. Both of these systems require proper emotional models capable of operating within the scope of human social expectations.

1.1 Emotion: Definitions, Descriptions, and Quantification

In "Human Emotions," Carol Izard states that emotions are comprised of three components: conscious feelings or experiences of emotion; processes that occur in the nervous system or brain; and observable expressive patterns (e.g., facial expressions). There are several theories that describe how emotions are realized and produced in the body. The cognitive/appraisal theories state that emotions result from two sources: an individual's physiological reaction to a situation and an individual's cognitive appraisal of that situation. The physiological theories argue that this appraisal is an intuitive and automatic process [60]. The causal theories of emotion are outside of the scope of this document and will not be discussed further.

In this thesis, we focus on methods for quantifying the emotion content of an utterance. The two most commonly used methods are the dimensional and categorical

(basic/discrete) characterizations of emotion. Categorical descriptions describe the emotion content of an utterance in terms of semantic labels of emotion (e.g., angry, happy, neutral, sad). Dimensional descriptions of emotion seek to describe emotion in terms of its underlying properties. These dimensions often include valence, describing the positivity (positive vs. negative), activation, describing the level of arousal (calm vs. excited), and dominance, describing the level of aggression (passive vs. aggressive) of the utterance.

It is not within the scope of this work to enter into the long-raging debate between these two methods. Rather, the quantification methods are chosen on a per study basis, to maximize the knowledge gained from each of the presented studies.

1.1.1 Dimensional Characterization

The dimensional view of emotion is based on the idea that emotions exist on a continuum, captured by axes with specific semantic meaning. The dimensions of emotion most commonly utilized were introduced by Harold Schlosberg in 1954 [103]. These dimensions included pleasantness-unpleasantness, attention-rejection, and low-high activation. Robert Abelson and Vellow Sermat [1] suggested that a two-dimensional structure is capable of capturing most emotions. They determined that the axes of pleasant-unpleasant and tension-sleep are adequate to describe human emotions after applying a multidimensional scaling approach over combinations of 13 stimuli from the Lightfoot series, a series of facial expressions. These axes, or variations of these axes, have been used in countless works, most often taking on the labels valence and activation/arousal [28, 43, 64, 74, 100, 102].

1.1.2 Categorical Characterization

The theory of a discrete characterization of emotion is based on the assumption that there is a set of emotions that can be considered, "basic." A basic emotion is defined as an emotion that is differentiable from all other emotions. In Ekman's "Basic Emotions," the author elucidates the properties of emotions that allow for the differentiation between the basic emotions (Table 1.1).

1	"Distinctive universal signals
2	Distinctive physiology
3	Automatic appraisal
4	Distinctive universals in antecedent events
5	Distinctive appearance developmentally
6	Presence in other primates
7	Quick onset
8	Brief duration
9	Unbidden occurrence
10	Distinctive thoughts, memories, images
11	Distinctive subjective experience."

Table 1.1: Ekman's characteristics that provide differentiation among the basic emotions, From [40] pg. 53

The set of basic emotions can be thought of as a subset of the space of human emotion, forming a "basis" for the emotional space. More complex, or secondary, emotions can be created by blending combinations of the basic emotions. For example, the secondary emotion of jealousy can be thought of as the combination of the basic emotions of anger and sadness [124].

In his critique of the theory of basic emotions, Andrew Ortony asserted that the idea of basic emotions is attractive for three main reasons: there exists a set of emotions

that pervade cultural boundaries and exist even in higher animals (such as primates); they are universally recognized and associated with specific facial expressions; and they seem to provide survival advantages to either the species or individual. However, there are high levels of variation in the makeup of the lists of basic emotions among emotions researchers, which would seem to question the validity of such an assertion [91]. The size of these lists may range from two emotions [89] to fifteen [40]. The proposed basic emotion sets still differ even when semantic variability is mitigated.

However, even in the presence of this noise, there are often four emotions postulated as basic. This emotion list includes anger, happiness, sadness, and fear. The basic emotions utilized in this work are a subset of this basic emotion list and include: anger, happiness, sadness, and neutrality (the absence of discernible emotional content). The emotion of fear is not included in this thesis because it is not well represented in any of the included datasets.

1.1.3 Emotion Evaluation Structures

In all of the work presented in this Dissertation, the emotional labels are assigned based on evaluator reports. One criticism of this style is that it relies on a fully conscious approach to attributing the emotional content of an utterance while the process itself is both conscious and unconscious. However, this limitation is inherent in the domain as it is not yet possible to develop a true understanding of both an individual's conscious and unconscious emotional reaction to a stimulus, although physiological emotion recognition is an ever developing field (for a survey, please see [69]). Furthermore, there is knowledge

to be gained from an individual's opinion of an emotional stimulus, even if it differs slightly from his or her true underlying reaction.

Ortony addresses this fundamental difficulty in his book, "The Cognitive Structure of Emotions". He states that:

> There is yet no known objective measure that can conclusively establish that a person is experiencing some particular emotion, just as there is no known way of establishing that a person is experiencing some particular color. In practice, however, this does not normally constitute a problem because we are willing to treat people's reports of their emotions as valid. Because emotions are subjective experiences, like the sensation of color or pain, people have direct access to them, so that if a person is experiencing fear, for example, that person cannot be mistaken about the fact that he or she is experiencing fear ([92], pg. 9)."

He thus asserts that the subjective reported emotion of the evaluators is acceptable evidence for an assignment of a ground truth.

1.1.4 Working Definitions

In this thesis, we cast the emotion perception process as a recognition problem, mapping between the observation and evaluator reported interpretation of a stimulus. This inherently conscious interpretation allows us to understand the emotional weight of a given stimulus.

We define the emotion label, or ground truth, of our emotional utterances in one of two ways. The first method defines an emotion's ground truth as the majority-voted opinion

of the set of evaluators. The second method, used in user-specific modeling experiments, is the individual's opinion for a given emotion.

1.2 Problem Statement and Methods

The goal of this work is to develop a mapping between emotion expression and reported emotion perception. This thesis presents two study genres, designed to forward the development of a new emotion quantification and interpretation framework. The first study genre is emotion recognition. In this thesis, we will demonstrate that emotion recognition is affected both by the inherent naturalness of human expression and variations in evaluator reporting styles. In the second study genre, emotion perception, this thesis will demonstrate that individuals have consistent methods for evaluating emotion expressions and specific features and modalities upon which they rely. The results of these studies are combined to forward a new emotion quantification and classification paradigm in which emotions are described based on a set of either semantic labels or data-driven components, called Emotion Profiles (EP). Please see Figure 1.1 for a pictorial description of the relationship between these studies.

1.2.1 Statistical Analyses of Perception

Evaluators do not use all information available during the emotion perception process. This thesis presents an analysis of the interaction between emotional audio (human voice) and video (simple animation) cues to further the understanding of how individuals integrate emotional information. The emotional relevance of the channels is analyzed with

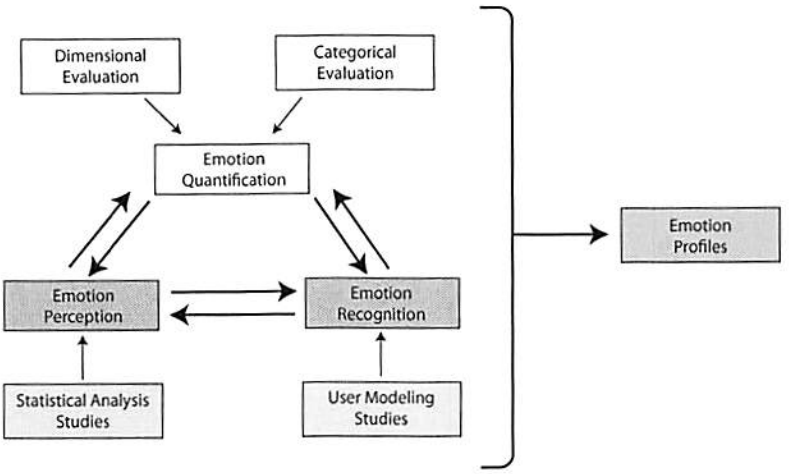


Figure 1.1: The relationship between the components of this thesis. The light yellow sections indicate work utilized, rather than work presented.

respect to their effect on human perception and through the study of the extracted audio-visual features that contribute most prominently to human perception. As a result of the unequal level of expressivity across the two channels, the audio biases the perception of the evaluators. However, even in the presence of a strong audio bias, the video data affect human perception. The feature selection results indicate that when presented with emotionally matched stimuli, users rely on both audio and video cues, but when presented with emotionally mismatched information, users rely solely on audio information. This result suggests that observers integrate natural audio cues and synthetic video cues only when the information expressed is in congruence. It is therefore important to properly design the presentation of audio-visual cues as incorrect design may cause observers to ignore the information conveyed in one of the channels.

1.2.2 Modeling Across Users

Evaluators provide a ground truth for emotion recognition studies. Unfortunately, these evaluations contain large amounts of variability both related and unrelated to the evaluated utterances. One approach to handling this variability is to model reported emotion perception at the individual level. However, the perceptions of specific users may not adequately capture the emotional acoustic properties of an utterance. This problem can be mitigated by the common technique of averaging evaluations from multiple users. We demonstrate that this averaging procedure improves classification performance when compared to classification results from models created using individual-specific evaluations. We also demonstrate that the performance increases are related to the consistency with which evaluators label data. These results suggest that the acoustic properties of emotional speech are better captured using models formed from averaged evaluations rather than from individual-specific evaluations.

1.2.3 Emotion Recognition via Emotion Profiling

Emotion recognition is complicated by more than feature modulation and user variability, it is often obfuscated by an uncertain ground truth. The ground truth is often masked and reporting processes may not capture true perception. Further, the features involved in the production of emotion are related across multiple time scales and are correlated with internal representations that are not observable. Engineering solutions are well situated to approximate this process. Although these models may not be able to capture and model the true link between emotion production and emotion perception, they can provide insight towards the relationship between these two processes.

Emotion expressions can be described by creating a representation in terms of the presence or absence of a subset of categorical emotional labels (e.g., angry, happy, sad) within the data being evaluated (e.g., a spoken utterance). This multiple labeling representation can be expressed using Emotion Profiles (EP). EPs provide a quantitative measure for expressing the degree of the presence or absence of a set of basic emotions within an expression. They avoid the need for a hard-labeled assignment by instead providing a method for describing the shades of emotion present in the data. These profiles can be used in turn to determine a most likely assignment for an utterance, to map out the evolution of the emotional tenor of an interaction, or to interpret utterances that have multiple affective components.

Profile-based techniques have been used within the community as a method for expressing the variability inherent in multi-evaluator expressions [108]. The profiles were used to represent the distribution of reported emotion labels from a set of evaluators for a given utterance. The authors compared the entropy of their automatic classification system to that present in human evaluations. In our previous work [88], EPs were described as a method for representing the phoneme-level classification output over an utterance. These profiles described the percentage of phonemes classified as one of five emotion classes.

The profiling method described in this Dissertation is derived from the combined output of a classification system composed of n binary classifiers and the confidences derived from each classification. EPs are created by weighting the output of the four classifiers by an estimate of the confidence of the assignment. EPs can be used to estimate a single emotion label by selecting the emotion class with the highest level of confidence,

represented by the EP or by further classifying the generated EPs. These EPs can also be used as a unit to describe emotions that cannot be captured by a single ground truth label. This technique provides a method for discriminately representing emotional utterances with ambiguous content.

1.2.4 Emotional Data Corpora

The presented work demonstrates the development of an emotion classification framework using two databases, one with artificially created emotionally matched and mismatched stimuli, the other with actors in a motion-capture setting. The first database ("Congruent-Conflicting") is used to study the effect of varying levels of expression on the reported human perception of emotion. This database is composed of synthetically generated audio-visual expressions composed of a computer generated face and a human voice. The expressions contain both congruent (emotionally matched audio and video information) and conflicting (emotionally mismatched audio and video information). It is used to analyze emotion expression both with respect to high-level label and with respect to features of importance.

The second database ("USC IEMOCAP") is used to study the relationship between feature modulations at the phoneme-level and utterance-level and evaluator emotional reports. The USC IEMOCAP database is an audio-visual plus motion-capture database recorded using a dyadic human interaction paradigm. It is used to study both the accuracy and trade-offs of user-specific modeling and to evaluate the efficacy of the emotion profiling technique for emotion recognition.

1.3 Related work and contributions to this topic

1.3.1 Related work in the Statistical Analysis of Perception

This work was principally motivated by the McGurk effect [77]. The McGurk effect occurs when mismatched audio and video syllables are presented to a human observer. Instead of perceiving either of the two presented syllables, McGurk and MacDonald found that observers perceived a third, distinct syllable. This finding has motivated many emotion research studies designed to determine if such an effect occurs within the emotion domain. Emotional McGurk studies are primarily conducted using either discrete emotion assignment (e.g., happy or angry) [29,30,32,45,58,76] or dimensional evaluation [83–85]. This effect has also been studied using fMRI [83] and EEG measurements [78]. In these studies, the emotion presentations have included congruent and conflicting information from the facial and vocal channels (e.g., [32]), facial channel and context (e.g., [83]), and facial and body postural/positional information (e.g., [78]).

In discrete choice evaluations, users are asked to rate the utterance by selecting the emotional label that best fits the data. Such evaluations allow researchers the point along an emotional continuum at which a given face or voice is of a sufficient emotional strength to bias the decision of the evaluators [29]. In a dimensional analysis, evaluators are asked to rate the presented stimuli according to the properties of those stimuli. Common properties (or dimensions) include valence (positive vs. negative), activation (calm vs. excited), and dominance (passive vs. aggressive). One common dimensional evaluation technique utilizes Self-Assessment Manikins (SAM) [9]. This evaluation methodology presents the dimensions of valence, activation, and dominance using a pictorial, text-free

display. This display method allows evaluators to ground their assessments using the provided end-points.

In [29], researchers combined still images with single spoken words in three experiments. The first experiment presented images morphed from two archetypal happy and sad emotional images into a visual emotional continuum. These images were accompanied by either vocally happy or sad human utterances. The evaluators were presented with an audio-only, video-only, or combined audio-visual presentation and were asked to assign the combined presentation into one of the discrete emotional categories of "happy" or "sad." The researchers found that the evaluators were able to correctly recognize the emotion in the audio-clip 100% of the time. In the combined audio-visual presentation, they found that the voice altered the probability that an evaluator would identify the presentation as "sad." The researchers then repeated the experiment, this time asking the users to judge the face and ignore the voice. They found that the emotion presented in the audio channel still had an effect on the discrete emotion assignment. However, in this experiment, they found that the effect was smaller than that seen previously. In the final experiment, the researchers created a vocal continuum ranging from happiness to fear (to allow for a more natural vocal continuum). They asked the users to attune to the voice and to ignore the face. They found that, as in the audio channel experiments, users were still affected by the visually presented emotion.

A similar study was presented in [58]. In that work, evaluators were asked to rate 144 stimuli composed of still images and emotional speech using happy, angry, and neutral emotions. The evaluators were asked to respond as quickly as possible after viewing a stimulus presentation. In the first experiment, the researchers asked the evaluators to

attune to the emotion presented in the facial channel. They found that the emotion presented in the vocal channel (the unattended emotion) affected the evaluators with respect to accuracy (discrete classification of facial emotion) and response time (faster for congruent presentations). When the evaluators were instead asked to attune to the facial channel, the researchers found that vocal channel mismatches decreased the facial emotion recognition accuracy and increased the response time. In the final experiment, users were presented with an emotional stimulus (a vocal utterance), a pause, and a second emotional stimulus (the facial emotion) used as a "go-signal." The researchers found that when the emotion presentations were separated by a delay, the channels no longer interacted in the emotion evaluation and the evaluators based their decisions on the vocal signal only.

Interactions between emotional channels have also been studied using emotional faces paired with contextual movies [83]. In that study, the evaluators were presented with four seconds of a movie and were then shown a static image. The contextual emotions included positive, negative, and neutral. The emotional faces included happy, fear, and neutral. These combined presentations were rated using SAMs. The researchers found that faces presented with a positive or negative context were rated significantly differently than faces presented in a neutral context. Furthermore, the fMRI data showed that pairings between faces and emotional movies resulted in enhanced BOLD responses in several brain regions.

The McGurk effect has also been studied with respect to body posture and facial analyses [78]. In this study, researchers paired emotional faces with emotional body positions (fear and anger for both) to analyze the interplay between facial and postural

information in emotion evaluation. They found that evaluators were able to assess the emotion state (using a discrete choice evaluation) of the stimulus most quickly and accurately when viewing congruent presentations. The results showed that the analysis time for faces-only was faster than for bodies-only. However, these results suggest that facial emotional assessment is biased by the emotion embedded in body posture.

The expression of emotion has also been studied in a more localized manner. One such method utilizes a "Bubble" [50]. This method is designed to identify regions of interest that correspond to task-related performance by only permitting users to view certain areas of the stimulus. The stimulus is covered by an opaque mask. Regions are randomly shown to the evaluators by creating Gaussian "bubbles," which allow users to glimpse regions of the masked stimulus. Given an infinite number of trials, all window combinations will be explored. The "Bubble" method allows for a systematic evaluation of stimuli components, but produces results that are difficult to translate into system design suggestions.

These past studies suggest that the video and audio channels interact during human emotion processing when presented synchronously. However, due to the discrete nature of the evaluation frameworks, it is difficult to determine how the perception of the targeted emotions change in the presence of conflicting information. The work presented in this thesis uses the dimensional evaluation method reported in [9] to ascertain the nature of the audio-visual channel interactions. This work also differs from previous work in its use of video clips rather than static photographs. The inclusion of dynamic facial stimuli in this work makes the results more transferable to the interactive design domain.

1.3.2 Related Work in Evaluator-Specific Modeling

Emotion recognition has been studied extensively [20,25,55,126]. However, these studies do not provide analyses of inter-evaluator differences. In [8,21,115], the authors present an analysis of the differences between self-evaluations and the evaluations of others. In [108], the authors present a new emotion classification accuracy metric that considers common inter-evaluator emotion classification errors. The question of inter-evaluator averaging remains unexplored.

Human evaluators are as unique as snowflakes. Consequently, one would expect that Hidden Markov Models (HMM) trained on individual-specific data would better capture the variability inherent in the individual's evaluation style. However, we demonstrate that models trained on averaged data either outperform or perform comparably to those trained solely on the individual-specific data. The results also suggest that evaluations from individuals with a higher level of internal emotional consistency are more representative of the emotional acoustic properties of the clips than those of less consistent evaluators.

1.3.3 Related work in Emotion Recognition

Engineering models provide an important avenue through which to develop a greater understanding of human emotion. These techniques enable quantitative analysis of current theories, illuminating features that are common to specific types of emotion perception and the patterns that exist across the emotion classes. Such computational models can inform design of automatic emotion classification systems from speech, and other forms

of emotion-relevant data. Multimodal classification of emotion is widely used across the community [15,106,119]. For a survey of the field, see [25,61,125].

In natural human communication, emotions do not follow a static mold. They vary temporally with speech [17], are expressed and perceived over multiple modalities [31,112], may be inherently ambiguous [13,34,37], or may have emotional connotations resulting from other emotional utterances within a dialog [65]. A classification scheme designed to recognize only the subset of emotional utterances consisting of well-defined emotions will not be able to handle the natural variability in human emotional expression.

Conventionally, when training emotion recognition classifiers, researchers utilize emotional expressions that are rated consistently, by a set of human evaluators. These expressions are referred to as prototypical emotion expressions. This process ensures that the models capture the emotionally-relevant modulations. However, while analyzing natural human interactions, including in an online human-computer or human-robot interaction (HCI or HRI) application, one cannot expect that every human utterance will contain clear emotional content. Consequently, techniques must be developed to handle, model, and utilize these emotionally ambiguous, or non-prototypical, utterances within the context of HCI or HRI.

1.3.4 Related Work in Emotion Profiling

Ambiguity in emotion expression and perception is a natural part of human communication. This ambiguity can be clarified by designating an utterance as either a prototypical or non-prototypical emotional episode, terms described by Russell in [101]. These labels can be used to provide a coarse description of the ambiguity present in an utterance.

Prototypical emotional episodes occur when all of the following elements are present: there is a consciously accessible affective feeling (defined as "core affect"); there is an obvious expression of the correct behavior with respect to an object; attention is directed toward the object, there is an appraisal of the object, and attributions of the object are constructed; the individual is aware of the affective state; there is an alignment of the psychophsyiological processes [101]. Non-prototypical emotional episodes occur when one or more of these elements are missing. Non-prototypical utterances can be differentiated from prototypical utterances by their enhanced emotional ambiguity.

Emotional ambiguity may result from the blending of emotions, masking of emotions, a cause-and-effect conflict of expression, the inherent ambiguity in emotion expression, and an expression of emotions in a sequence. Blended emotion expressions occur when two emotions are expressed concurrently. Masking occurs when one emotion (e.g., happiness) is used to mask another (e.g., anger). Cause-and-effect may result in a perception of ambiguity when the expressions have a conflict between the positive and negative characteristics of the expression (e.g., weeping for joy). Inherent ambiguity may occur when two classes of emotion are not strongly differentiated (e.g., irritation and anger). Finally ambiguity may also occur when a sequence of emotions is expressed consecutively within the boundary of one utterance [34]. In all of these cases, the utterance cannot be well described by a single hard label.

The proper representation and classification of emotionally ambiguous utterances has recently received increased attention. At the Interspeech Conference in 2009 there was an Emotion Challenge special session to focus on the classification of emotionally ambiguous utterances [105]. Similarly, at the Affective Computing and Intelligent Interaction

(ACII) Conference in 2009 there was also a special session entitled, "Recognition of Non-Prototypical Emotion from Speech–The Final Frontier?" This session focused on the need to interpret non-prototypical, or ambiguous, emotional utterances. Emotional ambiguity has also been studied with respect to classification performance [55, 107] and synthesis [10, 66].

Emotional profiles (EP) can be used to interpret the emotion content of ambiguous utterances. EP-based methods have been used to describe the emotional content of an utterance with respect to evaluator reports [57, 107], classification output [88], and perception, as a combination of multiple emotions, resulting from one group's actions towards another group [23]. EPs can be thought of as a quantified description of the properties that exist in the emotion classes considered. In [4], Lisa Feldman Barrett discusses the inherent differences that exist between classes of emotions. Between any two emotion classes, there may exist properties of those classes held in common, while the overall patterns of the classes are distinct. For instance, Barrett suggests that anger has characteristic feature modulations that are distinct from those of other classes. Thus, emotions labeled as angry must be sufficiently similar to each other and sufficiently different from the emotions labeled as another emotion. This overview suggests that in natural expressions of emotion, although there exists an overlap between the properties of distinct emotion classes, the underlying properties of two classes are differentiable. This further recommends a soft-labeling EP-based quantification for emotionally non-disjoint utterance classes.

EPs can be used to capture the emotional class properties expressed via class-specific feature modulations. Using the example of anger presented above, an angry emotion

should contain feature properties that are strongly representative of the class of anger but may also contain feature properties that are weakly similar to the class of sadness. However, this similarity to sadness does not suggest an error in classification, but a property of natural speech. Consequently, an EP representation capable of conveying strong evidence for anger (the major emotion) and weak evidence for sadness (the minor emotion) is well positioned to interpret the content of natural human emotional speech, since the minor expressions of emotion may suggest how an individual will act given a major emotion state and an event [57].

The classification technique employed in this thesis, support vector machines (SVM), has been used previously in emotion classification tasks [104, 107, 119]. SVM is a discriminative classification approach that identifies a maximally separating hyperplane between two classes. This method can be used to effectively separate the classes present in the data. There are two feature selection methods utilized in this thesis. The first is Information Gain, which has also been used widely in the literature [104, 107] and Principal Feature Analysis (PFA), which has also recently received attention in emotion classification [79, 80]. Both PFA and Information Gain are used to estimate the importance of the features using a classifier independent method. The purpose of this work is not to demonstrate the efficacy of the SVM, PFA, or Information Gain approaches, but instead to demonstrate the benefit of considering emotion classification output in terms of soft-labeling via relative confidences, rather than solely as hard labels.

EPs are a representation of the emotional components of an utterance using the degree of presence or absence of the emotions of angry, happy, neutral, and sad. Emotions that can be described as combinations of "basic" emotions should be characterizable using the

EP representation framework. This thesis investigates the ability of EPs to represent out-of-domain secondary emotions using frustration as a case study. The results demonstrate that EPs can represent the unseen secondary emotion statistically significantly differently than the "basic" emotions of angry, happy, neutral, and sad. The EPs can then be used to classify between angry, happy, neutral, sad, and frustrated as accurately as EPs trained with a frustration component. These results suggest that EPs are a robust representation for secondary emotions.

The correct number of profile components is not obvious. Four-dimensional EPs can accurately classify between the four classes of anger, happiness, neutrality, and sadness. However, it is not clear that the four emotions must also be the four components of the profile. This question is analyzed using a Cluster Profile (CP) representation. In CPs the underlying components are not based on the categorical labels of emotion but are instead based on data-driven clusters via the unsupervised clustering method of Agglomerative Hierarchical Clustering (AHC). The results demonstrate that CPs are as accurate at the four-way emotion classification task as EPs. The benefit to using CPs is that they do not require labeled training data for profile generation. However, CPs have a much higher dimensionality (15-components vs. four-components in EPs). These results suggest that the semantic emotional class labels of angry, happy, neutral, and sad have meaning not only with respect to perception, but also with respect to the underlying feature properties.

The data utilized in this thesis is from the USC IEMOCAP database [13]. This database has been used for studies ranging from interaction modeling [65, 67] to classification studies. In [88], the audio utterances were classified using Hidden Markov Models (HMM) into one of five states: anger, happiness, neutrality, sadness, and frustration.

The accuracies ranged from 47.34% for the classification of emotionally well-defined, or prototypical utterances, to 35.06% for the classification of emotionally ambiguous, or non-prototypical, utterances. In [81], the authors performed a profiling-based multimodal classification experiment on the IEMOCAP database. The authors utilized Mel Filterbank Coefficients (MFB), head motion features, and facial features selected using an emotion-independent Principal Feature Analysis (PFA) [71]. The authors developed four independent classifiers: an upper-face GMM, a lower-face eight-state HMM, a vocal four-state HMM, and a head-motion GMM. Each classifier outputted a profile expressing the soft-decision at the utterance-level. The output profiles were fused at the decision level, using a Bayesian framework. The training and testing were completed using leave-one-speaker-out cross-validation. The overall unweighted accuracy (an average of the per-class accuracies) for this system was 62.42%.

The work presented in this thesis is novel in that it presents a classification system based on the creation of EPs and uses this technique to interpret emotionally ambiguous utterances. It extends the EP description of [88] to include a measure of the confidence with which an emotional assessment is made. This confidence can be used to disambiguate the emotional content of utterances in expressions that would not otherwise be classified as a single expression of emotion.

1.4 Open problems and limitations

There are many open problems in the emotion recognition field due to the complex and dynamic nature of emotion expression. The techniques that we utilize to study emotion

recognition and evaluator reports are used commonly within the field. The novelty is in the combination of the machine-learning techniques and the problems that we address.

Highly ambiguous emotional utterances are rarely handled in conventional emotion recognition frameworks. By definition, these utterances either have an unclear ground truth or no ground truth at all. Conventionally, these utterances are not often considered in the testing phase. However, in natural human communication, many such utterances are this form. In, "Basic Emotions," Ekman states that, "Emotions obviously do occur without any evident signal, because we can, to a very large extent, inhibit the appearance of a signal. Also, a threshold may need to be crossed to bring about an expressive signal, and that threshold may vary across individuals [40]." This suggests that an emotion recognition system, built to analyze natural human speech, must be able to identify emotions ranging from subtle presentations to clear displays.

Our future work will pursue the disambiguation of emotional content. We will discuss the open problems below:

1. The EP-based recognition systems showed promising performance for the utterances in the USC IEMOCAP database. However, the current system does not take context into account. In [47], Frijda presents the laws of emotion, three of which reinforce the importance of context:

   (a) The law of comparative feeling: "The intensity of emotion depends on the relationship between an event and some frame of reference against which the event is evaluated. (pg. 353)"

(b) The law of conservation of emotional momentum: "Emotional events retain their power to elicit emotions indefinitely, unless counteracted by repetitive exposures that permit extinction or habituation, to the extent that these are possible (pg. 354)."

(c) The law of hedonistic asymmetry: "Pleasure is always contingent upon change and disappears with continuous satisfaction. Pain may persist under persisting adverse conditions (pg. 353)."

These three laws underscore the importance of considering the context in which the emotional clips were evaluated when estimating a label or profile for an expression as the feature modulations themselves do not describe the entire emotional scene, as viewed by the evaluators.

2. Our studies have demonstrated that individual evaluators have different emotion reporting trends. We show that we can more accurately predict the emotional ground truth of an average evaluator as compared to a single evaluator. Furthermore, our findings suggest that the reports of evaluators with a more consistent internal representation of emotion tend to be more easily predicted than those of evaluators with a less consistent representation. However, our findings were mitigated by the small number of evaluators considered and by acoustic-only models of emotion. We plan to explore evaluator-specific modeling via EP-based techniques to strengthen the findings of our earlier study.

3. Human study: We plan to create congruent and conflicting emotional expressions composed of both human facial and human vocal components. We will assess the

ability of EPs to quantify these expressions. We will also assess the evaluation trends of observers to determine channel reliance in the presence of unambiguous and ambiguous emotional information.

4. Application: Previous work has demonstrated that children with autism have a difficult time identifying the emotional content of expressions [22, 49, 59, 72]. We seek to develop a better understanding of the mismatch between cue generation and reported perception by developing a series of emotionally interactive avatar-guided scenarios. This investigation will allow us to better quantify the emotion perception deficit of children with autism.

1.5 Outline of thesis

The remainder of the thesis is organized as follows. Chapter 2 describes our studies of channel bias in audio-visual congruent and conflicting emotion expression. Chapter 3 describes our feature analysis studies of the "congruent-conflicting" database. Chapter 4 describes our studies of evaluator emotion reporting strategies. Chapter 5 describes our EP-based emotion recognition system. Chapter 6 describes a case study in the analysis of the robustness of the EP-based representation. Chapter 7 describes our work in utilizing data-driven clusters (rather than semantic clusters) for a profile-based representation. Finally, Chapter 8 provides discussion, conclusions, and future work.

Chapter 2

Statistical Data Analysis

In human-machine interaction there is an implicit assumption that the utterances and expressions produced by the machine will be recognized in a specified manner by a set of users. Conventionally, this assumption is validated using extensive user testing. However, this validation process is costly and time consuming. The goal of the work presented both in this chapter and in Chapter 3 is to develop a better understand human audio-visual emotion perception using statistical analyses. Accurate computational descriptions of the perceptual process may one day facilitate the construction of emotionally targeted and relevant stimuli.

Audio-visual emotional stimuli are often multichannel expressions composed of facial and vocal affect. This chapter presents an analysis of multichannel human emotion perception in the presence of emotionally matched ("congruent") and mismatched ("conflicting") audio-visual information in an animated display. Congruent information refers to an expression of the same emotion class across both the video and audio channel (e.g., angry face, angry voice). Conflicting information refers to the expression of different emotions across the two channels (e.g. angry face, happy voice).

This work was motivated by the well-known work of McGurk and MacDonald [77] called the McGurk Effect. The McGurk Effect is a multichannel syllabic perceptual phenomenon. The authors found that when presented with two distinct syllables on each of the facial and vocal channels, listeners perceived a third, distinct, syllable. This study led emotion researchers to investigate if such an effect exists for multichannel emotion perception. An understanding of the so called "Emotional McGurk Effect" would provide researchers with crucial information regarding the process of emotion perception.

The most common method for the study of the Emotional McGurk Effect is the presentation of concurrent emotionally evocative still photographs and vocalizations [29, 30, 58, 76]. In these perceptual experiments, participants are presented with either single channel (audio or photograph) or multichannel (audio and photograph) expressions of emotion. The participants then rate the stimuli using the forced-choice paradigm (e.g., happy vs. sad). The results demonstrate that the facial emotion expression tends to more strongly bias the emotional perception of the user than the vocal emotion expression [29]. However, dynamic audio-visual stimuli have not been as thoroughly studied.

The dynamic stimuli discussed in this chapter are composed of emotional audio (vocal) and video (facial) information. The emotions are either expressions of the same state (congruent presentation) or are of differing emotion states (conflicting presentation). This type of experimental construct permits an analysis of the relative importance of audio-visual features across broader levels of combinations.

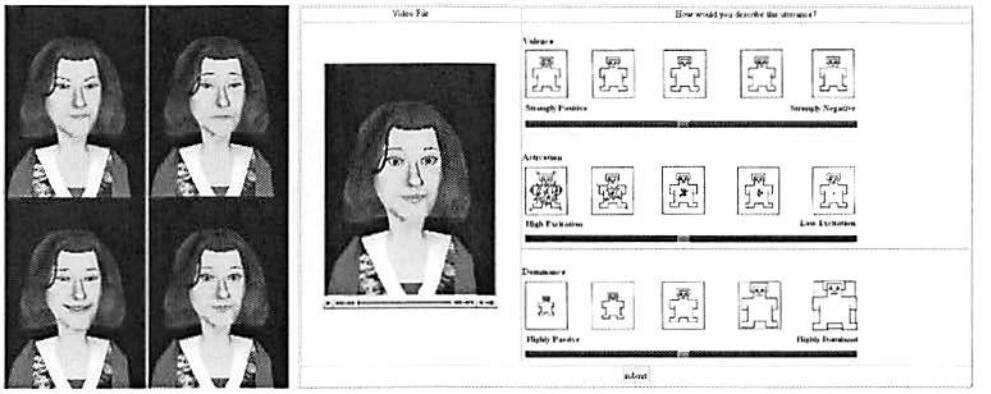
The stimuli are rated using self-report manikins [53]. These manikins present a pictorial description of the dimensional axes of emotion primitives. They include categories of valence (positive vs. negative), activation (calm vs. excited), and dominance (passive vs.

aggressive), which will be referred to as VAD. These dimensions allow for a continuous analysis of emotion perception rather than a discrete categorical analysis. To our knowledge, this is the first attempt to use the dimensional approach to analyze the combined perception of conflicting audio-visual stimuli in a continuous framework. This continuous environment allows for a more fine-grained understanding of how audio and video data interact in the human emotion perception process.

One of the challenges of such a study is creating stimuli that are free from artifacts. Purely human test data present challenges in that it may be difficult for actors to express an angry vocal signal with a happy facial expression. It would be undesirable to present stimuli to evaluators containing residual facial information resulting from unintended vocal emotional expressions and vice versa. As a result, we use an animated facial display. Despite its expressivity limitations, this interface allows for simple and artifact free synchronization between the audio and video streams.

2.1 Audio-Visual Stimuli

The vocal prompts utilized in this experiment were recorded from a female professional actress [122]. The actress recorded semantically neutral utterances across each of the following emotion states: happy, angry, sad, and neutral. The sentences were then rated by four evaluators using a forced-choice evaluation framework (happy, angry, sad, neutral, and other). Sentences that were rated uniformly by the four evaluators across all four emotion classes were used in the study. The resulting set was composed of nine distinct sentences recorded across all four emotions, for a total of 36 distinct vocal utterances.



(a) Clockwise from top left: angry, sad, neutral, happy.

(b) Dimensional evaluation tool.

Figure 2.1: The frames of the four emotional presentations (left) and online emotion evaluation interface (right) used in this study.

The video prompts created for this experiment were designed using the CSLU toolkit [109]. This toolkit allows a user to quickly and reliably create animations of targeted facial emotions that are synchronized with an input speech signal. The toolkit has sliders (representing the strength of emotion) for happy, angry, sad, and neutral emotions (for still stereotypical examples, see Figure 2.1(a)). Each vocal utterance (36 total) was combined with each of the four facial emotions (happy, angry, sad, and neutral) to create a total of 144 audio-visual clips.

2.2 Evaluation Procedure

The created stimuli were evaluated by 13 participants (ten male, three female) using a web interface (Figure 2.1(b)). The stimuli included audio-only, video-only, and audio-visual clips. These clips were randomly presented to the evaluators. There were a total of 139 audio-visual clips, 36 audio-only clips, and 35 video-only clips (one of the sad

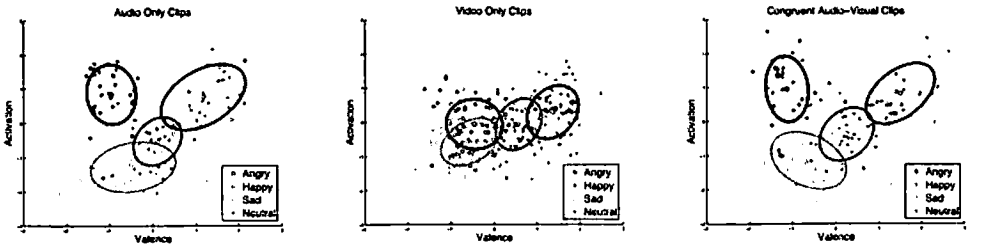
utterances was inadvertently, but inconsequentially, omitted due to a database error). Each participant evaluated 68 clips. The clip presentation order was randomized with respect to clip type (angry, happy, sad, neutral) and to clip content (audio, video, audio-visual). Each evaluator observed approximately 50% audio-visual clips, 25% audio clips and 25% video clips. The evaluators were allowed to stop and start the evaluation as many times as they desired.

The evaluation included both a flash video player and a rating scheme (Figure 2.1(b)). Each clip was rated from 0 – 100 along three dimensions, valence, activation, and dominance (VAD) using a slider bar underneath a pictorial display of the variation along the dimension. These scores were normalized using z-score normalization along all three dimensions for each evaluator. Z-score normalization was used to mitigate the effect of the various rating styles of the evaluators and thus make the evaluations more compatible.

2.3 General Results

The VAD ratings of the evaluators were plotted along the dimensions of valence and activation (Figure 2.2) to observe the effect of audio-visual information on emotion perception vs. that of either only video or audio information. The dominance dimension is not shown due to its high level of correlation with the activation plots. This visualization allows for a graphical depiction of the relationship between the emotion states and their VAD ratings.

The separation between the clusters was higher in the audio-only evaluation (Figure 2.2(a)) than in the video-only evaluation (Figure 2.2(b)). Discriminant analysis shows



(a) Audio-only evaluations. (b) Video-only evaluations. (c) Congruent audio-visual evaluations.

Figure 2.2: The valence (x-axis) and activation (y-axis) dimensions of the evaluations, the ellipses are 50% error ellipses.

that there exists a higher level of confusion in the video-only evaluation (classification rate: 71.3%) than in the audio-only evaluation (classification rate: 79.3%) (test of proportions, $\alpha \leq 0.1$, Table 2.1). This result suggests that the emotions presented in the audio data were more highly differentiable than in the video data, possibly due to the limited expression in the animated face used in this analysis.

A discriminant analysis of the congruent audio-visual data showed that the average classification accuracy non-significantly increased (test of proportions, $\alpha \leq 0.1$) to 80.8% (Table 2.1). The congruent angry and happy classification rates increased when compared to the video-only and audio-only classification rates. However, the neutral and sad classification rates decreased. This suggests that the audio and video data were providing emotionally confounding cues to the participant with respect to the sad and neutral emotion classes. The confusion between these two classes in the congruent audio-visual case was in between that of the audio-only (higher level of confusion) and video-only (lower level of confusion).

(a) Confusion matrix for audio-only (ave. = 79.3%).

	A	H	S	N
A	90.6	3.1	0	6.3
H	3.2	80.6	6.5	9.7
S	0	0	72.7	27.3
N	6.5	2	19.4	71.0

(b) Confusion matrix for video-only (ave. = 71.3%).

	A	H	S	N
A	70.0	1.4	11.4	17.1
H	0	76.7	1.4	21.9
S	11.3	2.8	71.8	14.1
N	5.9	20.6	7.4	66.2

(c) Confusion matrix for congruent audio-visual (ave. = 80.8%).

	A	H	S	N
A	95.0	0	2.5	2.5
H	3.6	89.3	0	7.1
S	7.7	0	69.2	23.1
N	9.7	9.7	16.1	64.5

Table 2.1: Discriminant analysis classification (A=angry, H=happy, S=sad, N=neutral) for (a) audio-only, (b) video-only evaluations, and (c) audio-visual.

2.4 Perception of Presentation Types

When designing an audio-visual interactive agent, it is important to determine if the emotion perception resulting from an audio-visual presentation of emotion will differ significantly from that of an audio-only or video-only presentation. We investigate the perceptual differences of the presentation conditions by analyzing the VAD ratings of the three presentation conditions (audio-only, video-only, audio-visual). The differences between the presentation conditions were analyzed by comparing the emotion-specific cluster means of the congruent audio-visual presentation to those of the audio-only and video-only presentations using a one-way ANOVA analysis. The independent variables were the z-normalized VAD ratings. The dependent variables were presentation class. The ANOVA analysis indicated that the group means for the presentation conditions were significantly different for angry across all three VAD dimensions ($F(2; 130) > 25.273$; $p < 0.001$), for happy across the activation dimension ($F(2; 129) = 7.84$; $p = 0.001$), for

(a) Audio-Only

	A	H	S	N
A	-	VD	AD	VAD
H	VD	-	VAD	VA
S	AD	VAD	-	AD
N	VAD	VA	AD	-

(b) Video-Only

	A	H	S	N
A	-	VAD	AD	VD
H	VAD	-	VAD	VA
S	AD	VAD	-	VAD
N	VD	VA	VAD	-

(c) Congruent Audio-Visual

	A	H	S	N
A	-	VD	AD	VAD
H	VD	-	VAD	VA
S	AD	VAD	-	VAD
N	VAD	VA	VAD	-

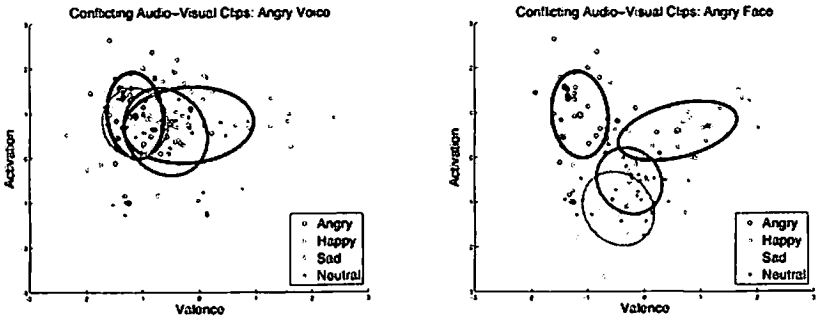
Table 2.2: ANOVA post hoc analysis (A - angry, H - happy, S - sad, N - neutral) of the three presentation conditions. The letters VAD indicate that the cluster means are significantly different ($\alpha = 0.01$) in the valence, activation, and dominance dimensions.

sad across the activation and dominance dimensions ($F(2; 116) > 5.769$; $p = 0.004$), and for neutral across the valence and activation dimensions ($F(2; 127) > 6.453$; $p = 0.002$). This results suggested that the clusters were distinct in the three presentation conditions

The ANOVA analysis was repeated with the dependent variable as emotion class (the same independent variables were used) to determine whether or not distinct emotion classes existed in the audio-visual space. The four clusters are distinct in at least two dimensions at the $\alpha = 0.01$ level of significance in all three presentation conditions (ANOVA post-hoc analysis, Table 2.2).

2.5 Biases in Evaluation

The design of audio-visual emotional interfaces requires both a knowledge of how observers interpret specific facial and vocal features and how observers weight the audio and video channels during the perceptual process. This weighting process is dependent on the



(a) "Angry" vocal emotion held constant, facial emotion varied.

(b) "Angry" facial emotion held constant, vocal emotion varied.

Figure 2.3: Comparison between the emotion perceptions resulting from conflicting audio-visual presentation.

relevance of the emotional information contained in the channels, and on the affective bandwidth of the channels. The affective bandwidth of a channel is defined as, "... how much affective information a channel lets through [96]." The bandwidth of the channel is a function of the physical limitations (e.g., number of degrees of freedom) and the emotional relevance of the channel (e.g., the voice is the primary source for activation differentiation but alone cannot sufficiently convey valence [56]). An understanding of the audio-visual perceptual process would allow designers to tailor the information presented to maximize the emotional information conveyed to, and recognized by, observers.

2.6 Analysis of VAD Ratings for Emotional Clusters

Within this experiment, the natural audio channel dominated the perception of the users. This bias can be observed graphically (Figures 2.3(a) and 2.3(b)). The figures depict the reported valence-activation perception of the conflicting audio-visual presentations. In Figure 2.3(a), all of the conflicting audio-visual presentations have an angry voice (and

(a) Confusion matrix for angry voice held constant (ave. = 40.5%).

	A	H	S	N
A	55.0	5.0	12.5	27.5
H	14.0	48.8	16.3	20.9
S	36.4	9.1	39.4	15.2
N	28.1	46.9	12.5	12.5

(b) Confusion matrix for angry face held constant (ave. = 70.5%).

	A	H	S	N
A	87.5	2.5	7.5	2.5
H	10.3	75.9	3.4	10.3
S	8.0	0	72.0	20.0
N	11.4	8.6	34.3	45.7

Table 2.3: Classification accuracy in the presence of conflicting audio-visual information, (a) angry voice held constant, (b) angry face held constant.

angry, happy, neutral, or sad faces) while in Figure 2.3(b), all of the conflicting audio-visual presentations have an angry face (and angry, happy, neutral, or sad voices). These figures demonstrate that the perception is more strongly influenced by the vocal emotion than by the facial emotion. In Figure 2.3(a) the perception resulting from the varied facial emotions are more similar to the congruent angry presentation than the perceptions resulting from the varied vocal emotions of Figure 2.3(b).

The presence of the audio bias can also be verified using discriminant analysis (Table 2.3). In this investigation, the conflicting audio-visual presentations were again grouped by: 1) the emotion expressed over the vocal channel (i.e., angry voice with angry, happy, neutral, and sad faces) and 2) the emotion expressed over the facial channel (i.e., angry face with angry, happy, neutral, and sad voices). The classification goal is to recognize the four emotion classes, where a class is defined as a combination of a facial and vocal emotion, e.g., angry voice – neutral face. In this analysis, a higher level of accuracy suggests that the four emotion classes are more separable (and thus more distinct) in the channel with the higher level of accuracy. The classification accuracy of the vocal emotion group was 40.5% while the classification accuracy of the facial emotion group was 70.5%. The results demonstrate that the four emotion classes are less differentiable

when the audio emotion is held constant than when the video emotion is held constant (Table 2.3), providing further evidence of an audio bias.

The audio bias can also be assessed by comparing the means of the four emotion classes in the vocal and facial emotion groups using an ANOVA post hoc analysis. In the vocal emotion group, the class means were distinct only across the valence and dominance dimensions ($F(3; 144) > 5.152$; $p = 0.002$). In the facial emotion group, the cluster means were significantly different across all three dimensions ($F(3; 125) > 34.239$; $p < 0.001$). These three pieces of evidence indicate that the vocal information of the clip provided a stronger influence on the emotion perception of the participant than did the facial information. The presence of an audio bias suggests that when evaluators were presented with ambiguous or conflicting emotional information, they used the natural vocal channel to a larger degree than the synthetic facial channel to determine the emotion state.

2.7 Analysis of Video Contribution to Perception

Although the audio biased the evaluations, the video information did provide emotionally salient information. The contribution of the audio and video information to user audio-visual emotion perception can be seen by evaluating the audio-visual cluster shifts of the conflicting presentations. In this analysis, the cluster center of the audio-visual presentation (e.g. angry voice – happy face) was compared to the audio-only and video-only cluster centers (e.g., angry voice and happy face) using paired t-tests implemented in Matlab. All reported results refer to a significance of $\alpha \leq 0.05$.

The audio biased the audio-visual perception of activation. In 10 of the 12 conflicting audio-visual presentation types, the cluster means of the audio-visual presentation were significantly different than the video-only cluster mean presentations. These same presentations were significantly different from the audio-only presentation in only one of the 12 conflicting presentations (Table 2.4). This result suggests that the audio information biased the evaluations of the users in the activation dimension.

The valence dimension was not as strongly biased by the audio information as the activation dimension. In the valence dimension, 10 of the 12 conflicting audio-visual presentation clusters had means significantly different than those of the video-only presentations. In the activation dimension, 8 of the 12 audio-visual clusters had means significantly different than those of the audio-only presentations (Table 2.4). This suggests that in the valence dimension, the evaluators integrated both the audio and video information when making emotional assessments.

2.8 Conclusions

This chapter provided evidence supporting the joint processing of audio and visual cues in emotion perception. This was most stridently recognized when comparing the cluster results from the audio-only and video-only data to the congruent audio-visual clusters.

This chapter also provided evidence suggesting that the combination of audio and visual cues does not always result in a combined emotional rating between the ratings of the individual channels. It would seem that the integration of these cues results in a

Audio	Video	ΔV_{audio}	ΔA_{audio}	ΔD_{audio}	ΔV_{video}	ΔA_{video}	ΔD_{video}
Angry	Happy	**0.771***	-0.097	**-0.465**	**-1.558***	**0.459***	**0.967***
	Sad	-0.174	-0.045	- 0.207	**-0.568***	**1.390***	**1.693***
	Neutral	**0.364**	-0.250	-0.203	**-1.165***	0.661	**1.293***
Happy	Angry	**-0.582**	-0.165	0.174	**1.053***	**0.669***	**-0.243**
	Sad	**-1.391***	-0.187	- 0.370	**0.349**	**1.167***	**0.312**
	Neutral	0.061	0.045	0.194	**0.666***	**0.874***	**0.473**
Sad	Angry	-0.001	0.137	**0.922**	0.005	**-1.08***	**-0.640**
	Happy	**0.717**	**0.399**	**0.429**	**-1.108***	**-1.173***	**-0.501**
	Neutral	0.464	0.193	**0.641**	**-0.562***	**-1.024***	- 0.225
Neutral	Angry	**-0.366***	0.006	0.278	**0.200**	**-0.452**	**-0.435**
	Happy	**0.681***	0.253	0.053	**-0.583***	**-0.564***	-0.028
	Sad	**-0.624***	-0.243	-0.295	0.0462	-0.182	0.092

Table 2.4: Cluster shift analysis with respect to the VAD dimensions (where ΔV_{audio} represents the shift in valence mean from the audio-only evaluation to the audio-visual evaluation). Entries in bold designate evaluations of the audio-visual presentations that are significantly different, with $\alpha \leq 0.05$, from that of either the video-only or audio-only presentations (paired t-test). Entries with a star (*) designate evaluations that are significantly different with $\alpha \leq 0.001$.

different experience than observing the cues individually. This has been shown previously in [29,58] regarding facial, but not vocal prominence.

One of the limitations of this work was the limited level of expression inherent in the animated face. Users tuned to the audio more predominantly than the video when making their emotional assessments. We believe that this is due in part to the highly expressive vocal information. Since the two channels did not have a similar level of expression this may have led to the perceived importance of the audio signal. In previous studies [29] it was found that the facial information strongly influenced the perception of the audio information when photographs of human faces were used.

In future work, we will use a more expressive animated face to analyze the interplay between the facial and vocal channel with an enhanced level of facial expression. The use of continuous domain analysis provides a novel tool for understanding the relationship

between the level of expression and the relative strength of the emotional bias. Our further work will also analyze a synthetic voice combined with the current animation to determine if a combination of two channels with similar levels of expression will result in facial information having a more prominent role in the evaluation of the emotional display.

2.9 Work Published

The work presented in this chapter was published in the following articles:

1. **Emily Mower**, Maja J Matarić and Shrikanth S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information." *IEEE Transactions on Multimedia*, 11:5(843-855). August 2009.
2. **Emily Mower**, Sungbok Lee, Maja J Matarić, Shrikanth Narayanan. "Joint-processing of audio-visual signals in human perception of conflicting synthetic character emotions." In Proceedings of *IEEE International Conference on Multimedia & Expo (ICME)*, Hannover, Germany, June 2008.
3. **Emily Mower**, Sungbok Lee, Maja J Matarić, Shrikanth Narayanan. "Human perception of synthetic character emotions in the presence of conflicting and congruent vocal and facial expressions." In Proceedings of *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Las Vegas, Nevada, March-April 2008.

Chapter 3

Emotionally Salient Features

The proper expression of robotic and computer animated character emotions have the potential to influence consumer willingness to adopt technology. As technology continues to develop, robots and simulated avatars ("synthetic characters") will likely take on roles as caregiver, guide, and tutor for populations ranging from the elderly to children with autism. In these roles, it is important that robots and synthetic characters have interpretable and reliably recognized emotional expressions, which will allow target populations to more easily accept the involvement of synthetic characters in their day to day lives.

Reliable and interpretable synthetic emotion expression requires a detailed understanding of how users process synthetic character emotional displays. This chapter presents a quantitative analysis of the importance of specific audio-visual features with respect to emotion perception. Armed with this knowledge, designers may be able to control the number of feature combinations that they explore. Instead of implementing and testing broad combinations of features, designers may be able to concentrate on those features upon which observers rely when when making synthetic affective assessments.

As described in the previous chapter, the work of McGurk and MacDonald [77] has provided a framework commonly employed for the study of human emotional perception. The McGurk experimental paradigm is often employed in emotion perception research. One common evaluation method [29,30,32,58,76] is to create an emotional continuum, anchoring the ends with two archetypal emotional images and presenting these images with emotional vocalizations. Subjects then identify the emotion presented from a discrete set (e.g., angry vs. happy). This presentation framework allows the researchers to model the perceptual influence of the two modalities. However, discrete emotional evaluation frameworks do not fully capture the interplay between the two channels. The complexities of the two channels may be better modeled using a continuous framework (e.g., valence, activation, dominance, "VAD") [9, 55, 84, 85] rather than a discrete framework (e.g., angry, happy, sad, neutral). This framework allows users to express the complexity of an emotional presentation using the properties of the emotion, rather than the lexical description. Continuous frameworks have also been used to analyze the interplay between facial actions and personality perception [2].

In this chapter, emotionally relevant features are identified using the conflicting and congruent presentation framework, discussed in the previous chapter. In a conflicting presentation, a presentation in which the emotions expressed in the facial and vocal channels do not match, the evaluators must make an assessment using mismatched emotional cues. This presentation style is an important research tool because it provides combinations of features that would not, under ordinary circumstances, be viewed concurrently, allowing for a greater exploration of the feature space. Features that are selected across both congruent and conflicting presentations are features that provide emotionally discriminative

power both in the presence and absence of emotional ambiguity. The feature selection method employed in this chapter is Information Gain, which has been used previously to identify emotionally salient features [93]. The explanatory power of the resulting reduced feature set is validated using Support Vector Machine classification [117].

The results suggest that the pitch range and spectral components of the speech signal are perceptually relevant features. Design-relevant prior knowledge statistics (e.g., the average valence rating for angry speech) are also perceptually relevant. However, these prior knowledge features contribute supplementary, rather than complementary information. The results suggest that the perceptually relevant valence features are: pitch and energy ranges and facial expression features (eye shape, eyebrow angle, and lip position), the perceptually relevant activation features are: are energy and spectral features, and the dominance dimension features are: energy, spectral, and pitch range features. This novelty of this work is in its analysis of dynamic audio-visual features and their contribution to dimensional emotional evaluation.

3.1 Feature Sets

The data used in the analyses presented in this chapter are described in Chapter 2.1. The data are dynamic presentations of emotion across an animated face and a natural human voice. The facial and vocal emotions are combined to produce congruent presentations (in which the emotions expressed in the face and voice match) and conflicting presentations (in which the emotions expressed in the face and voice do not match). In the previous

chapter we demonstrated that the audio biases the perception of the evaluators, in this chapter we will investigate the contribution of specific feature types.

3.1.1 Audio Features

The audio features utilized in this experiment included 20 prosodic features and 26 spectral features averaged over an utterance. The prosodic features included pitch, energy, and timing statistics. The spectral features included the mean and standard deviation of the first 13 MFCCs, also used in [54]. These features are summarized in Table 3.1. It is also important to note that the selected audio features represent relevant design parameters that can be used to modulate synthetic speech [11, 90].

3.1.2 Video Features: FACS

The Facial Action Coding System (FACS) was developed by Ekman and Friesen as a method to catalogue the muscle movements of the human facial structure [41]. These features allow for a design-centered analysis of a video sequence through the use of actuated facial units (a subset of the facial muscles acting to achieve a visually perceivable facial movement).

This method of video analysis is important for design centered user modeling. Since the facial features described by action units are physically realizable motion, any facial feature identified as important could, given sufficient actuation, be implemented on a synthetic character. Consequently, the method identifies salient facial motions from a set of available facial actions.

The features used in this study represent a simplified subset of the FACS action units due to the simplicity of the input video stream. The video features employed in this study are summarized in Table 3.1. These features include eyebrow (movements, types, and angles), eye shape, and lip corner position features. Other areas of the face were not analyzed because they were static with respect to emotion presentation for these data. These features were manually coded by the author.

Stream	Feature Class	Measures
Audio	Pitch	mean, standard deviation, median, min, max, range, upper quartile, lower quartile, quartile range
	Volume	mean, standard deviation, max, upper quartile, lower quartile, quartile range
	Rate	pause to speech ratio, speech duration mean, standard deviation, pause duration mean, standard deviation
	MFCC	1 - 13, mean and standard deviation
	Prior Knowledge: Binary Emotion	angry voice, happy voice, sad voice, neutral voice
	Prior Knowledge: Mean Statistics	valence, activation, dominance of each emotion class
Video	Eyebrow Movement	none, downward, upward, downward upward, upward downward, downward upward downward, upward downward upward
	Eyebrow Movement Type	none, once, twice, thrice
	Eyebrow Angle	flat, inner raised, inner lowered, outer raised, outer lowered
	Lip Corner Position	neutral, raised, lowered
	Eye Shape	eyes wide, top soft, top sharp, bottom soft, bottom sharp
	Prior Knowledge: Binary Emotion	angry face, happy face, sad face, neutral face
	Prior Knowledge: Mean Statistics	valence, activation, dominance of each emotion class

Table 3.1: A summary of the audio and video features used in this study.

3.1.3 Prior Knowledge Features

In this study there were prior knowledge features included in both the audio and the video features sets. These prior knowledge audio-visual features included average value statistics for the individual audio or video channel (e.g., the average VAD ratings of the audio-only and video-only components of the clip) and indicator variables that encode the presence or absence of an emotion in the audio and video channels (e.g., angry video- y/n, happy audio- y/n). From a design perspective these features describe the relevance of general emotion descriptors with respect to subsequent emotion perception. Although single semantic labels (e.g., "happy") do not fully describe the properties of the clip, the knowledge of this label may provide insight into the resulting perception of the evaluator.

3.2 Method

3.2.1 Class Definition

This study was designed to identify the audio-visual features that contribute most to the emotional perception of the users. The contribution of the features was assessed using Information Gain and the feature set was reduced using a minimum gain threshold. The discriminative ability of the reduced feature set was validated using Support Vector Machine (SVM) classification, a classification tool developed by Vapnik [117]. The classification performances of the reduced feature sets were compared to the classification performances of the full feature sets using SVM. SVM is a classification algorithm that finds a

maximally separating hyperplane by optionally transforming the input data into a higher-dimensional space. SVM has been employed for emotion classification tasks [5,15,68,98]. SVM is implemented here using Weka, a Java-based data mining software package [120].

The evaluations of the presented audio-visual, audio-only, or video-only clips were rated dimensionally (valence, activation, and dominance), on a scale from 0-100 (Chapter 2, Figure 2.1(b)). These evaluations were preprocessed using z-normalization across all three VAD dimensions, to allow for inter-evaluator comparisons. The emotional VAD space was also preprocessed using a binary discretization based on the neutral VAD centroids. This binarization was used to account for the simplicity of the video information; the perception of this channel did not vary widely across emotion class. After discretization, each evaluator rating was composed of a 3-dimensional binary vector representing the VAD rating with respect to the neutral centroid (e.g., valence: positive vs. negative).

3.2.2 Feature selection

The goal of the feature selection analysis is to determine which audio-visual features contribute to the explanation of variance within the audio-visual perceptual evaluations. The data were prepared for feature selection techniques by separating, the data into three groups, evaluations of congruent data ("congruent database"), evaluations of conflicting data ("conflicting database"), and evaluations of congruent and conflicting data ("combined dataset"). The feature selection techniques were applied to the three data subsets separately and the results were compared. The feature set was reduced using the Information Gain Attribute Selection algorithm, an algorithm implemented in Weka. Information gain feature selection techniques have been used previously in salient emotional feature

selection [93]. Information gain describes the decrease in the entropy of set X, $H(X)$ (e.g., valence), given the conditional entropy between X and attribute Y, $H(X|Y)$ (e.g., valence given the presence of a lowered eyebrow) is known (Equation 3.1) [82]. Features were retained if they contributed a gain of at least 0.1 with respect to the target class (discretized valence, activation, and dominance).

$$Gain(S, A) \equiv H(X) - H(X|Y) \tag{3.1}$$

3.3 Feature Selection Results

3.3.1 Feature Selection Results for the Combined Congruent – Conflicting Dataset

The features selected for the combined audio-visual congruent-conflicting presentations can be viewed in Table 3.2. Features in italics were observed over two VAD dimensions, features in bold italics were observed across all three dimensions. The features that contribute most to the explanation of evaluator variance, as suggested by Information Gain feature selection, are the prior knowledge audio statistics (e.g., average valence rating), the high energy band spectral components, and the lower pitch quartile. The prior knowledge audio statistics include the average ratings for valence, activation, and dominance. These statistics represent the centroid of the dimensional ratings of the audio and video components of the audio-visual clip.

Feature selection was extended to the audio-only and video-only presentations to determine if the features selected as important to audio-visual perception also explain the

Dim	Relevant Features
Val	***ave_audio_val*** (0.159), ***ave_audio_dom***, ***ave_audio_act***, ***mfcc12_mean***, ***vol_quartlow***, ***f0_quartlow***, *mfcc03_mean*, ave_video_dom, eyebrow_angle, lip_corner_position, happy_voice, ave_video_val, eyebrow_angle_flat, eye_shape_bottom_sharp, *f0_quartup* (0.1)
Act	*vol_quartup* (0.472), *vol_quartrange*, *vol_std*, *vol_mean*, *mfcc01_std*, mfcc07_std,*vol_max*, *mfcc01_mean*,*mfcc08_std*, mfcc10_mean, *mfcc12_std*, mfcc13_mean, ***ave_audio_val***, ***ave_audio_dom***, ***ave_audio_act***, speech_duration_std, *mfcc03_mean*, *mfcc05_std*, pause_to_speech_ratio, ***mfcc08_mean***,*f0_quartrange*, *mfcc11_std*, f0_mean, mfcc10_std, ***mfcc12_mean***, *mfcc06_std*,*mfcc13_std*, mfcc02_mean, ***f0_quartlow***, f0_std, mfcc11_mean, f0_range, *f0_quartup*, *mfcc09_mean*, f0_max, mfcc07_mean, mfcc09_std, *mfcc03_std*, mfcc04_mean, mfcc05_mean, mfcc04_std, mfcc06_mean, f0_min, f0_median, pause_dur_mean, sad_voice, *vol_quartlow*, mfcc02_std, pause_duration_std, speech_duration_mean, *angry_voice* (0.165)
Dom	***ave_audio_dom*** (0.204), ***ave_audio_val***, *vol_mean*, ***ave_audio_act***, *angry_voice*, ***mfcc12_mean***,*mfcc06_std*, ***mfcc08_mean***, *mfcc11_std*, *vol_max*, *f0_quartrange*, *mfcc08_std*, *mfcc05_std*, *mfcc12_std*, *mfcc09_mean*, *vol_quartup*, *vol_quartrange*, ***f0_quartlow***, *mfcc01_mean*, *mfcc01_std*, *vol_std*, *mfcc13_std*, *mfcc03_std* (0.103)

Table 3.2: A summary of the features used in the audio-visual analysis of this study. The order of the feature (left - right) indicates their relative importance. The numbers in parentheses represent the highest and lowest mean information gain above the threshold. Bold italic fonts represent features selected across all three dimensions, italic fonts represent features selected across two dimensions.

variance inherent in the audio-only and video-only evaluations. The feature representation across the presentation conditions will be discussed using two abbreviations: VAD_{Audio} and VAD_{Video}. These abbreviations indicate the features that were selected for the valence, activation, and dominance in *both* the audio and audio-visual presentations or the video and audio-visual presentations. Each feature occurs in the $VAD_{Audio|Video}$ sets a maximum of six times, corresponding to the valence, activation, and dominance for the audio/video-only presentations and the valence, activation, and dominance for the audio-visual presentations. The only features selected in six cases (across all audio-visual and audio-only presentations- VAD^{+}_{Audio}) were the three prior knowledge audio statistics (the emotion specific centroids for valence, activation, and dominance, e.g., ave_audio_val),

the quartile pitch range, and a high frequency MFCC feature. The binary variable representing the presence of an angry voice was selected in five of the six dimensions.

The most highly represented video features were the mean video statistics describing the emotion specific activation and valence evaluations. These two features were represented in four of the VAD^{+}_{Video} components. The mean video dominance statistic was represented across only three of the VAD^{+}_{Video} components.

There were several features represented once (out of a possible six times) in the VAD^{+}_{Audio} or VAD^{+}_{Video}. This set of features includes three of the FACS-inspired features (a binary feature addressing eyebrow angle, a feature addressing eyebrow movement direction, and an eye shape feature). All three of the features were utilized in the video valence classification problem. This result suggests that these features provide specialized dimensional differentiation. This feature set of singularly represented features also includes two of the binary channel features (happy voice and sad voice indicators). These features were utilized in the audio-visual valence and activation classification tasks respectively. This suggests that these features, while not applicable to channel dependent classifications (i.e. video-only), do provide additional information with respect to multimodal discretization and disambiguation.

3.3.2 Feature Selection Results for the Congruent and Conflicting Datasets

The combined congruent-conflicting dataset was separated into two datasets: congruent presentations and conflicting presentations. Feature selection was applied to these separated datasets to analyze the features that contribute to the perception of emotionally matched and mismatched emotional expressions. In the congruent presentations, both

Dim	Relevant Features
Val	PRIOR KNOWLEDGE: angry (face, voice), happy (face, voice), neutral (face, voice) AVERAGE CHANNEL RATINGS: audio **VAD**, video **VAD** VIDEO: eye shape (**specific**, bottom sharp, bottom soft, wide eyes), eyebrow angle (general, flat, inner lowered, outer raised), eyebrow mvmt., **eyebrow mvmt. timing**, lip position PITCH: quartile (low, high) VOLUME: ***max***, std, quartile (low, ***high***, ***range***) RATE: speech duration std MFCC: mean (**1**, 4, 6, 7, 10, **11**, ***12***), std (**4**, **9**, 10)
Act	PRIOR KNOWLEDGE: happy (face, voice), sad (face, voice) AVERAGE CHANNEL RATINGS: audio **VA**, video **VAD** VIDEO: eye shape (**specific**, bottom soft, top sharp, top soft), eyebrow angle (general, inner raised, outer lowered), **eyebrow mvmt. timing**, lip position PITCH: mean, median, max, range, std, quartile (low, high, range) VOLUME: mean, ***max***, quartile (***high***, ***range***) RATE: pause duration mean, pause to speech ratio MFCC: mean (1, 2, 3, 5, 6, 8, 10, **11**, ***12***, 13),std (1, 3, **4**, 5, 6, 8, **9**, 10, 12, 13)
Dom	PRIOR KNOWLEDGE: angry (face, voice), sad (face, voice) AVERAGE CHANNEL RATINGS: audio **VAD**, video **VAD** VIDEO: eye shape (**specific**, top sharp, top soft), eyebrow angle (inner lowered, inner raised, outer lowered, outer raised), eyebrow mvmt., **eyebrow mvmt. timing** PITCH: max, std, quartile range VOLUME: mean, ***max***, std, quartile (***high***, ***range***) MFCC: mean (**1**, 3, 5, 8, **11**, ***12***), std (**4**, 5, 6, 7, 8, **9**, 11, 13)

Table 3.3: The audio-visual features selected in the **congruent database**. Features in bold are features that were selected across the valence, activation, and dominance dimensions of the *Congruent* database. Features in bold-italics are features that were selected in the $Congruent_{VAD}$ and $Conflicting_{AD}$ databases.

audio and video features were selected across all three dimensions (Table 3.3). In the combined congruent-conflicting dataset no video features were selected for the dimensions of either activation or dominance. This prominence of audio features (and corresponding paucity of video features) can still be seen by observing the features selected for the conflicting database (Table 3.4).

In the congruent database, there were a total of fifteen features selected across the three dimensions. The video features included an eye shape feature (describing the emotional shape of the eye), an eyebrow timing feature (describing if movement occurred

Dim	Relevant Features
Val	none over the threshold of 0.1
Act	PRIOR KNOWLEDGE: angry voice, sad voice AVERAGE CHANNEL RATINGS: audio ***VAD*** PITCH: mean, median, max, min, range, std, quartile (**low**, high, range) VOLUME: **mean**, ***max***, std, quartile (low, ***high range***) RATE: pause duration (mean, std), speech duration MFCC: mean (**1**, 3, 4, 5, 6, 7, **8**, **9**, 10, 11, ***12***), std (1, 2, 3, **6**, 8, 9, 10, 11, 13)
Dom	AVERAGE CHANNEL RATINGS: audio ***AD*** PITCH: **quartile low** VOLUME: **mean**, ***max***, quartile (***high***, ***range***) MFCC: mean (**8**, **9**, ***12***), std (5, **6**)

Table 3.4: The audio-visual features selected in the **conflicting database**. Features in bold are features that were selected across the valence, activation, and dominance dimensions of the *Congruent* database. Features in bold-italics are features that were selected in the $Congruent_{VAD}$ and $Conflicting_{AD}$ databases.

within the first, second, third, or multiple thirds of the utterance). The audio features included volume features (including the utterance length maximum and quantile, representing 25%–75% of the energy, max and range), MFCC mean and standard deviation features, and prior knowledge average statistics (describing the mean valence and action of the audio clip and the mean valence, activation, and dominance of the video clip). These features are presented in Table 3.3 in bold font.

The features selected in the conflicting database included only audio features. There were a total of eleven features selected across both the activation and dominance dimensions. There were no features selected in the valence dimension (all of the features presented an information gain under the threshold of 0.1). The features selected across the activation and dominance domain included a pitch feature (lower quantile, 25% of the full range), volume features (mean, max, upper and range quantile features), MFCC mean and standard deviation features, and prior knowledge average statistics (describing

Figure 3.1: Comparison between the emotion perceptions resulting from conflicting audio-visual presentation.

the mean activation and dominance of the audio clip). These features are presented in Table 3.4 in bold font.

There were a total of five features represented over the three dimensions of the congruent database and the activation and dominance of the conflicting database. These features included volume features (max, upper and range quantile), an MFCC mean feature, and a prior knowledge average statistic (describing the mean activation of the audio clip). These features are presented in Tables 3.3 and 3.4 in bold italic font. This feature set is an audio-only feature set. The common features can also be visualized in Figure 3.1.

3.4 Validation: SVM Classification

3.4.1 Validation of the Combined Congruent – Conflicting Feature Sets

The results from the SVM classification across the three presentation conditions (audio-visual, audio-only, video-only) are presented in Table 3.5. The valence was most accurately classified in the video-only presentation (84.45%), followed by the audio-visual presentation (76.62%), and the audio-only presentation (73.28%). The activation was

Presentation	Dimension	With Priors (%)		Without Priors (%)	
		Full	Reduced	Full	Reduced
Audio-Visual	Valence	76.618	77.453	75.1566	75.1566
	Activation	85.8038	85.595	84.9687	85.595
	Dominance	72.0251	69.3111	73.0689	68.2672
Audio	Valence	73.2759	80.1724	79.3103	80.1724
	Activation	86.2069	87.069	87.069	87.069
	Dominance	75	73.2759	75.8621	70.6897
Video	Valence	84.4523	84.4523	84.4523	85.5124
	Activation	55.477	61.1307	56.5371	61.1307
	Dominance	57.9505	65.7244	57.9505	65.7244

Table 3.5: This table presents the classification results (SVM) over the three presentation conditions (audio-only, video-only, audio-visual) and three dimensions (valence, activation, dominance). "Full" refers to classification performed with the original feature set. "Reduced" refers to classification performed with the feature set resulting from Information Gain feature selection.

most accurately classified in the audio-only presentation (86.21%), followed by the audio-visual presentation (85.80%), and the video-only presentation (55.48%). The dominance was most accurately recognized in the audio-only presentation (75.00%), followed by the audio-visual presentation (72.03%), and the video-only presentation (57.95%). These results support the channel bias results from the previous chapter, which asserted that audio biased the perception of activation while both the video and audio information contributed to the perception of valence (Chapter 2.7, Table 2.4).

The classification accuracies were tested using a difference of proportions test to determine if the classification accuracy changed when either the feature set was reduced, or when the prior information was removed. None of these accuracies differed significantly at the $\alpha = 0.05$ level across feature set size (full vs. reduced) or based on prior knowledge (present vs. absent). This result suggests that the reduced feature sets without prior knowledge statistics can explain the variance in the user evaluations with a similar of accuracy to that of the full feature set.

Presentation	Dimension	With Priors (%)		Without Priors (%)		Baseline
		Full	Reduced	Full	Reduced	
Congruent	Valence	88.71	89.52	88.71	89.52	54.84
	Activation	84.68	86.29	84.68	86.29	58.87
	Dominance	78.23	78.23	77.42	78.23	58.87
Conflicting	Valence	71.23	-	71.27	-	58.03
	Activation	84.79	84.51	83.67	84.23	52.96
	Dominance	64.79	67.32	63.66	67.89	55.21

Table 3.6: The SVM classification accuracies (percentages) over the two database divisions (congruent, conflicting) and three dimensions (valence, activation, dominance) using feature sets reduced with the Information Gain criterion discussed in Section 3.2.2. The columns marked "Full" refer to the full feature set. The columns marked "Reduced" refer to the reduced feature set.

3.4.2 Validation of the Congruent and Conflicting Feature Sets

The performance of the SVM classifier was affected by presentation type. In general, the congruent presentations were more accurately classified than the conflicting presentations (Table 3.6). Using the full feature set, valence was more accurately classified in the congruent presentations than in the conflicting presentation (88.71% vs. 71.23%, respectively). The activation of the congruent and conflicting presentations was recognized very similarly (84.68% vs. 84.79%, respectively). The dominance of the congruent presentations was more accurately recognized than the dominance of the conflicting presentations (78.23% vs. 64.79%, respectively).

The differences in classification accuracy were analyzed using the difference of proportions test. The classification accuracy (full feature set) for the valence dimension was significantly lower in the conflicting presentation than in the congruent presentation ($\alpha \leq 0.05$). The classification accuracy for the dominance dimension was also significantly lower in the conflicting presentations than in the congruent presentations for both

the full and the reduced feature sets (dominance, reduced feature set: $\alpha \leq 0.05$). The difference in the classification accuracy of the activation dimension was not significant across either presentation type or feature set size.

The classification accuracies were not affected by prior knowledge across either dimension or presentation type (congruent vs. conflicting). This suggests that the information contained within the prior knowledge features (semantic labels) is also contained within the audio and video features. The classification performance was also not affected by feature set size (full vs. reduced) in any dimension. In all conditions (except for conflicting valence and both full feature sets for conflicting dominance), the SVM performance beat the baseline with a significance of $\alpha \leq 0.01$.

3.5 Discussion

The results of the SVM analysis support the feature selection results regarding the reliance upon audio and video information demonstrated in the previous chapter. The SVM classification results for the activation domain do not change significantly between the congruent and conflicting presentations (Tables 3.5 and 3.6). This is expected since humans tend to rely upon audio for activation detection, and since when presented with conflicting information, evaluators were shown to rely primarily upon the audio channel. Therefore, when asked to determine the activation, the users were well prepared in either presentation condition and the variance in the evaluations did not increase enough to decrease the SVM classification results.

In the valence dimension, humans tend to utilize facial information. In the audio-visual domain, the evaluators were able to integrate and utilize both the audio and the video information when making their valence assessment. However, as previously stated, when observing conflicting emotional presentations, evaluators tended to rely on the audio information. Therefore, when the evaluators attempted to analyze the valence dimension based primarily on the audio signal, the variance in the evaluations increased and the performance of the SVM classification using the full feature set decreased.

The VAD ratings of the dominance dimension are affected by both the audio and video channels. It is therefore expected that the results of the SVM classification would lie in between that of the activation and valence dimensions with respect to performance decrease. The SVM classification results are in accordance with the VAD shift analysis and also support the hypothesis that the variance in the evaluation of dominance is affected by both the audio and video channels.

3.6 Conclusion

This chapter presented a channel-level analysis of an animated character emotional presentation. This work identified the video and audio features that are utilized during emotional evaluations of both congruent and conflicting presentations.

Classification tasks have perviously been used for perceptual experiments. In [12], speech synthesis parameters were selected using classification techniques. The features that were selected in this process were used to synthesize and modify speech. The creation of this feature subset allowed the researchers to minimize the utterances to be rated

by evaluators. In future studies, this same technique will be applied to determine which combinations of facial and vocal channel capabilities should be utilized for emotion recognition applications. This presentation framework will allow for the study of how various audio-visual combinations affect human emotional perception.

The results of the SVM classification on the full and reduced feature sets suggest that it is possible to identify a reduced feature set with emotional explanatory power in the congruent presentations and in the activation and dominance dimensions of the conflicting presentations. SVM classification performance on this reduced feature set indicated that there was not a significant decline in performance. These feature sets also support the finding that users rely upon audio information to detect activation information [46]. However, the audio channel does not provide sufficient valence differentiation and observers must thus rely upon other modes of affective communication (video, context, etc.) [56].

In [18], the authors presented an analysis of audio-visual emotion modulation. The data were segmented by utterances and utterances were compared across four emotion classes (angry, happy, sad, neutral). The utterances were further segmented by phoneme boundaries. The phonemes of the emotional utterances were compared to a neutral baseline. The data suggested that phonemes that contained little emotion modulation (the feature values of the emotional utterance were not significantly different than those of the emotional utterance) were accompanied by more facial movement than those phonemes that were more strongly emotionally modulated. This recasts the emotion production problem as an emotional bit allocation problem in which the information is transmitted across the two channels based on the channel bandwidth available. In the work presented in this chapter, such emotional subtleties were not included due to the limited nature of

the video channel and the nonlinearities inherent in a fusion between audio and video channels. Future experiments will utilize audio-visual data augmented with motion capture recording [13]. This will allow for a closer study of the channel modulation, fusion, and perceptual integration resulting from the natural expression of audio-visual emotional utterances.

This work was limited by the expressivity constraints on the video channel. Given the expressivity inequalities and the single instantiation of the emotional interface, it is difficult to generalize these results broadly without follow-up investigations. Future studies will include a four-by-four factorial design including both synthetic and human faces and voices. These four combinations will begin to illustrate the interaction between audio and video information across varying levels of emotional expressivity. The human data for these future studies will be created using dynamic time warping. This method can be used to align the phoneme duration of a target sentence (e.g., angry) given the phoneme duration of another sentence (e.g., neutral). This manipulation permits a combination of audio and video data streams with the same lexical content, but different emotional realizations, allowing for a comparison across purely human audio and video features.

This chapter presented a quantitative analysis of the features important to emotional audio-visual perception. A reduced feature set was created and validated using SVM classification. However, more analyses are required to determine the impact of these features on emotion perception, rather than on emotion classification. Future work will explore this avenue using analysis by synthesis techniques to compare the emotional salience of features as a function of their identified relevance.

3.7 Work Published

The work presented in this chapter was published in the following articles:

1. **Emily Mower**, Maja J Matarić and Shrikanth S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information." *IEEE Transactions on Multimedia*, 11:5(843-855). August 2009.
2. **Emily Mower**, Maja J Matarić, Shrikanth Narayanan. "Selection of Emotionally Salient Audio-Visual Features for Modeling Human Evaluations of Synthetic Character Emotion Displays." In Proceedings of *IEEE International Symposium on Multimedia (ISM)*. Berkeley, California, December 2008.
3. **Emily Mower**, Sungbok Lee, Maja J Matarić, Shrikanth Narayanan. "Joint-processing of audio-visual signals in human perception of conflicting synthetic character emotions." In Proceedings of *IEEE International Conference on Multimedia & Expo (ICME)*, Hannover, Germany, June 2008.
4. **Emily Mower**, Sungbok Lee, Maja J Matarić, Shrikanth Narayanan. "Human perception of synthetic character emotions in the presence of conflicting and congruent vocal and facial expressions." In Proceedings of *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Las Vegas, Nevada, March-April 2008.

Chapter 4

Evaluators as Individuals

Quantitative models of user perception have the potential to facilitate the design of synthetic emotional expressions. These models could lead to computer agents and robots that more naturally and functionally blend into human society [27, 95]. User specific emotion modeling and synthesis requires an understanding of human emotion perception, often measured using stimuli presentation experiments. Unfortunately, the evaluation process is non-stationary. Subjective emotion appraisals of evaluators change as the evaluators tire and as they are exposed to increasing numbers of emotional utterances. It is common practice to estimate emotional ground truth by averaging evaluations from multiple evaluators. The question remains as to whether this averaging between evaluators with different internal representations of emotion sacrifices important individual information.

As discussed in the introduction, the field of emotion classification has been studied extensively [20, 25, 55, 126]. However, it is the view of the author that it is important to have a stronger understanding of how evaluators differ in their emotion reporting style and how these differences affect consequent classification accuracies. The differences between the way an individual perceives his/her portrayal of emotion and the way other evaluators

perceive these same displays has been studied [8,21,115] and the results suggest that there is a difference between how these groups view the affective content of stimuli. However, the effect on classification accuracy of individual evaluation styles has not been sufficiently addressed.

This chapter presents an analysis of human emotion evaluation. The foci of this chapter are: the measurement of the consistency between evaluators and the evaluation of evaluators based on automatic emotion recognition. Firstly, we study the consistency between the categorical emotion labels (e.g., angry, happy, sad, neutral) of the utterances and the evaluators' internal representation of valence (positive vs. negative) and activation (calm vs. excited), using Naïve Bayes classification. This classification framework is used to estimate the categorical emotion of an utterance given individual evaluators' ratings of valence and activation. It is hypothesized that higher performance will be observed for evaluators with internally consistent representations of the dimensional emotional space. It is possible to map from dimensional properties of emotion to categorical labels [101]. Therefore, it is hypothesized that those with an easily modeled mapping will be more accurately classified. Secondly, we model the relationship between the temporal acoustic properties of a clip and the subjective valence or activation rating using Hidden Markov Models.

The conventional approach in emotion recognition is to use subjective evaluations to measure the performance of the system. We propose the opposite approach: to use the results of an emotion recognition system to measure the accuracy of subjective evaluations. We hypothesize that the performance of the automatic system will increase if the emotion labels are accurate. These models are compared across three conditions: a)

training and testing on ***individual*** data; b) training and testing on ***averaged*** data; and c) training on ***averaged*** data and testing on ***individual*** data.

As stated in the introduction, humans and their methods of reporting the emotion content of affective stimuli are unique. Therefore, we hypothesized that models trained on the evaluations of individuals would result in higher classification results than models trained on an average evaluator or models trained on average evaluator and tested on individuals. However, we found that models trained on averaged data and tested on either individual data or averaged data outperformed the individual-specific train-test scenarios. This study also suggested that individuals who are more consistent in their appraisal of emotion are more accurately modeled than those individuals who are less consistent.

4.1 Data

4.1.1 IEMOCAP Data

The database utilized in this study is the USC IEMOCAP database, collected at the University of Southern California [13]. The USC IEMOCAP database is an audio-visual database, augmented with motion capture recording. It contains approximately 12 hours of data recorded from five male-female pairs of actors (ten actors total). The goal of the data collection was to elicit natural emotion expressions within a controlled setting. The benefit of the acted dyadic emotion elicitation strategy is that it permits the collection of a wide range of varied emotion expressions. The actors were asked to perform from (memorized) emotionally evocative scripts and to improvise upon given emotional targets.

The emotional freedom provided to the actors allowed for the collection of a wide range of emotional interpretations. The benefits to utilizing acted data are discussed more fully in [3, 19, 44].

The data were evaluated using two evaluation structures: categorical evaluation and dimensional evaluation. In both evaluation structures, the evaluators observed the audio-visual clips in order temporally (with context). In the categorical evaluations, evaluators were asked to rate the categorical emotion present from the set of: angry, happy, neutral, sad, frustrated, excited, disgusted, fearful, surprised, and other. The evaluators could tag an utterance with as many categorical labels as they deemed appropriate. There were a total of six categorical evaluators who evaluated overlapping subsets of the database. Each emotion was labeled by at least three categorical evaluators. In the dimensional evaluations, the evaluators were asked to rate the clip according to its valence, activation, and dominance properties. Valence describes the positive vs. negative aspect of the emotion [1 = most negative, 5 = most positive]. Activation describes the calm vs. excited aspect of the emotion [1 = most calm, 5 = most excited]. The dimensional evaluation task was completed by a separate set of six evaluators, again evaluating overlapping subsets of the data. Each emotional utterance within the database was labelled by at least two dimensional evaluators [13].

In each evaluation task, the disparate evaluators were combined into a single rating to determine an overall ground truth. The categorical ground truth was established using majority voting over all of the reported categorical labels. The dimensional ground truth was established by averaging (without rounding) over the dimensional evaluators [13].

4.1.2 Data Selection

In this chapter we present dimensional classification analyses. We use the audio files from the five female actresses in the IEMOCAP data. We consider only clips labeled (using majority voting over the categorical evaluators) as angry, happy, sad, or neutral. We present results using dimensional evaluations from the two evaluators who evaluated the largest quantity of data (evaluators one and two). Evaluator one analyzed 1,773 clips and evaluator two analyzed 1,682 clips from the angry, happy, sad, neutral set. The clips evaluated by evaluator two are a subset of those evaluated by evaluator one. Please see [13] for more database details.

4.1.3 Audio Features

We extracted 13 filterbanks of Mel Filterbanks (MFB), their delta, and acceleration from the audio files. MFBs model the human auditory system by creating filterbanks of increasing width as the frequency increases. This structure approximates the increasing de-sensitivity to deviations in frequency as the frequency content of the signal increases in human hearing. Mel Frequency Cepstral Coefficients (MFCC), commonly used in automatic speech recognition, are calculated by taking the Discrete Cosine Transform (DCT) of the MFBs. MFBs have been shown to contain more emotional information than MFCCs [20].

4.1.4 Treatment of Evaluations

This chapter presents two types of evaluator studies: ***individual*** and ***averaged***. The experiments based upon *individual* evaluations study the dimensional evaluator behaviors

of the two evaluators, evaluator one and evaluator two, separately. These evaluations are neither averaged nor normalized. The *averaged* evaluations are the averages of the valence and activation ratings of evaluators one and two. The averaged rating is always rounded up to the nearest integer.

4.2 Approach

There are two points of interest that arise when considering evaluator performance: evaluator consistency (how similarly evaluators rate clips of the same semantic label) and evaluator reliability (how representative the labels are of the acoustic properties of the utterance). To answer these questions we consider two probabilistic modeling techniques: Naïve Bayes and Hidden Markov Models, respectively. Previous work has demonstrated the efficacy of utilizing Naïve Bayes to recognize the emotional content of speech [118] and Hidden Markov Models to capture its underlying temporal properties [20].

4.2.1 Naïve Bayes

Research has shown that categorical emotions can be depicted as occupying specific portions of a dimensional space defined by valence and activation [24,55,86,101]. For example, archetypal angry emotions lie within an area defined by negative valence and high activation, while archetypal happy emotions lie within an area defined by positive valence and mid to high activation. This suggests that given only an evaluator's subjective evaluation of the valence and activation of an utterance, it should be possible to estimate the categorical emotion label [55].

One measure of evaluator consistency is to determine how well a simple classification algorithm predicts the categorical emotion label of a clip given the subject evaluation of valence and activation. We use Naïve Bayes classification for this analysis. In this classification task, evaluator one's and two's subjective valence and activation ratings are used to predict the *majority voted* categorical label.

4.2.2 Hidden Markov Models

In the emotion evaluation process, there exists a dependency between assigned evaluation and the temporal acoustic properties of an utterance. We use Hidden Markov Models (HMM) to model the relationship between the temporal fluctuations of the acoustic properties and the resulting reported emotion perception. The HMM classification accuracies provide insight regarding how representative the subjective dimensional tags of the evaluators are of the underlying emotional acoustic properties of the utterances. The accuracies are also used to analyze the effectiveness of averaging subjective, unnormalized evaluations obtained from multiple individuals.

We describe two separate classification tasks: valence and activation. In these tasks, the original five point scale was collapsed into a three point scale to combat a data sparsity issue. Classes one and five were not tagged with sufficient frequency to form models across both evaluators. Class three, representing neutral (for both valence and activation), remained unchanged. Classes one and two (either negative valence or lowly activated) were collapsed into a single class and classes four and five (either positive valence or highly activated) were collapsed into a single class. This resulted in three model

groups for the valence dimension and three models groups for the activation dimension classification tasks.

Original Data		Transformed Data		
Time	Content	Time	Phoneme	Emotional Phoneme Class
0 - 31	silence	0 - 47	sil	sil
48 - 57	what	48 - 51	W	one_liquid
		52 - 54	AH	one_back/mid
		55 - 57	T	one_stop
58 - 70	was	58 - 60	W	one_liquid
		61 - 63	AX	one_back/mid
		64 - 70	Z	one_fricative
71 - 87	that	71 -75	DH	one_fricative
		76 - 83	AE	one_front
		84 - 87	TD	one_stop
88 - 145	silence	88 - 145	sil	sil

Table 4.1: Data format used for HMM categorical emotion training, original sentence: "What was that," expressed with a valence rating of one.

In each model group the data were modeled at the phoneme level. The phonemes were clustered into seven classes to ensure an adequate quantity of training data for each phoneme class. The seven classes included: front vowels, back/mid vowels, dipthong, liquid, nasal, stop consonants, and fricatives (for a detailed mapping see [20]). The utterances had accompanying transcription files at the word and phoneme level, generated using forced alignment [13]. The phoneme-level transcription files were modified for each utterance, replacing the original phonemes with phoneme classes (see Table 4.1 for an example). In each classification task (valence and activation) there were seven phoneme class models for each of the three model groups, plus emotion-independent models for silence and laughter for a total of 23 models.

The HMMs were trained using HTK [123]. Each model had three-states and eight mixture components. The HMMs were trained in two ways: a) using individual-specific

evaluations, and b) using averaged evaluations. The individual-specific HMMs were tested using individual-specific evaluations. The averaged HMMs were tested using both individual-specific evaluations and averaged evaluations. The testing procedure utilized word-level forced alignment using *-I in HVite.* This focused the classification task on the identification of the correct emotional phoneme class, rather than the correct phoneme and the correct emotion class.

The output of the HMM classification consisted of a transcript file containing the estimated emotional phoneme states over specified time windows. The final emotion of the utterance was assigned using majority voting over the estimated emotional phonemes, weighted by the time duration of each assigned emotional phoneme class. The emotion class represented most frequently in the output transcription was assigned as the final class label.

4.3 Results

4.3.1 Naïve Bayes classification of evaluator consistency

The subjective appraisals of valence and activation are linked to the categorical emotion label [101]. This link can be simply modeled using Naïve Bayes (NB). We used an NB classifier, implemented in the Matlab pattern recognition toolkit, PRTools [39], to predict the categorical emotion label of the clip given only the subjective valence and activation evaluations of: a) evaluator one, and b) evaluator two. In all cases, the clips are chosen from the set evaluated by both evaluators one and two and with a categorical

(a) Confusion matrix for evaluator 1, Accuracy = 59.51%

	A	H	S	N
A	61	4	13	22
H	6	74	0	20
S	35	8	23	34
N	10	12	6	73

(b) Confusion matrix for evaluator 2, Accuracy = 66.80%

	A	H	S	N
A	81	2	1	15
H	0	56	0	44
S	23	4	39	35
N	6	3	9	82

Table 4.2: Confusion matrices for the categorical emotion classification task (A = angry, H = happy, S = sad, N = neutral). The results presented in this table are percentages.

label of angry, happy, sad, or neutral. The analysis was performed using five-fold cross-validation. The results show that evaluator one's valence and activation ratings predicted the correct categorical label 59.51% of the time while evaluator two's dimensional evaluations predicted the correct categorical label 66.80% of the time (see Table 4.2 for the evaluator-specific confusion matrices).

4.3.2 HMM classification for correspondence between content and evaluation

In this section, models are referred to as "A - B", where "A" represents the training set and "B" represents the testing set. The accuracies of the models trained with averaged data (models "Ave - (Ave or Ind)" in Tables 4.3 and 4.4) are either better or comparable to the accuracies of the models trained with individual data (models "Ind - Ind" in Tables 4.3 and 4.4). The models trained and tested on averaged data (models "Ave - Ave" in Tables 4.3 and 4.4) had a higher accuracy than either of the individual models (models "Ind - Ind" and "Ave - Ind" in Tables 4.3 and 4.4) for both valence and activation. The classification performance of the "Ave - Ave" model improves significantly only with respect to evaluator one's activation and evaluator two's valence ($\alpha = 0.01$, difference of proportions). In all other conditions, the change in classification performance

Type	Evaluator	1 (%)	2 (%)	3 (%)	Total
Ind - Ind	Evaluator 1	50.00	65.90	23.71	52.18
Ind - Ind	Evaluator 2	37.47	60.28	46.65	44.33
Ave - Ave	Average	47.86	64.70	40.00	52.68
Ave - Ind	Evaluator 1	44.28	61.28	38.44	50.91
Ave - Ind	Evaluator 2	36.01	69.72	41.34	44.39

Table 4.3: Classification: **valence** across the three levels.

between the averaged and individual models is not significant ($\alpha = 0.01, 0.05$, difference of proportions).

The models trained and tested on individual data performed unequally for the valence and activation classification tasks. The evaluator one model outperformed the evaluator two model for the valence task ($\alpha = 0.01$, difference of proportions). The evaluator two model outperformed the evaluator one model for the activation task ($\alpha = 0.01$, difference of proportions). The models trained and tested on individual data did not perform significantly differently than the models trained on averaged data and tested on individual data ($\alpha = 0.01, 0.05$, difference of proportions).

Type	Evaluator	1 (%)	2 (%)	3 (%)	Total
Ind - Ind	Evaluator 1	64.55	23.16	66.76	47.79
Ind - Ind	Evaluator 2	68.81	39.37	62.83	55.79
Ave - Ave	Average	64.50	47.00	65.93	56.86
Ave - Ind	Evaluator 1	41.18	42.20	71.60	47.55
Ave - Ind	Evaluator 2	60.57	49.00	63.70	57.70

Table 4.4: Classification: **activation** across the three levels.

4.4 Discussion

The NB and HMM classification indicated that the evaluation styles and strengths of the two evaluators differed across tasks. However, when the evaluations from both evaluators were combined, the HMM classification accuracies across the valence and activation classification problem either improved or did not change significantly. This suggests that models constructed from averaged evaluator data may capture the emotional acoustic properties of the utterance more closely even given different evaluation styles and internal representations of the relationship between the dimensional and categorical emotion labels.

The NB results suggest reasons for the discrepancies between the classification performance for the HMMs modeled on evaluator-specific data. The NB classification for evaluator one indicates that evaluator one's internal representation of valence is more strictly defined than that of evaluator two (Table 4.2). Evaluator one's confusion matrix demonstrates that based on the subjective valence and activation ratings, there exists a smaller confusion between happiness and other emotions than is observed for evaluator two. Happiness is the only emotion with positive valence and should be differentiable based on the valence rating. It should be noted, that evaluator one's confusion matrix suggests that there is an increased confusion between happiness and sadness. This may be due to a misrepresentation of activation, discussed in the following paragraph. The differences between the inter-evaluator dimensional consistency may explain why the individual-specific HMM valence model for evaluator one outperformed that of evaluator two.

The NB results show an opposite trend for the dimensional ratings of activation. These results suggest that evaluator two's dimensional rating of activation is more internally consistent when compared to that of evaluator one. For example, evaluator one's results indicate that there exists a higher level of confusion between anger and sadness than is observed in evaluator two's results. Anger and sadness are emotion classes that should be differentiable based on their activation (high vs. low, respectively). The difference in evaluator activation consistency is supported by the HMM classification accuracies. The HMM activation classification performance is higher for evaluator two than for evaluator one.

The comparisons between the NB and HMM results suggest that evaluators one and two have different evaluation styles and internal dimensional representation of emotion. However, when the ratings of these two evaluators are combined, the performance of the HMM classification on valence and activation improved (significantly with respect to evaluator one activation and evaluator two valence). This suggests that even given large quantities of data, it may be more beneficial to create averaged models of dimensional evaluation, rather than evaluator-specific models (given evaluation styles that are not divergent).

The user evaluations utilized in this study were not normalized per evaluator. While the results of normalized evaluations may improve overall classification accuracies, such techniques are not necessarily representative of real-world user interactions. It is not good practice to discount the feedback of a user regarding emotion expression. It is important, from a user initiative standpoint, to work with the evaluations as provided. Furthermore, given new users in a human-computer or human-robot interaction scenario, it may not

be possible to develop normalization constants in real time, necessitating the use of raw user input.

It is also important to note that the data utilized in this experiment come from only partially emotionally constrained dyadic acted speech (both scripted and improvised). The utterances were not recorded on a turn-by-turn basis with rigid emotional targets. As a result, the emotional utterances in this database are not archetypal emotion expressions. Consequently, one cannot expect the classification accuracies of these more natural and subtle human emotional expressions to match those of classifications performed on read speech databases.

4.5 Conclusion

This chapter presented evidence suggesting that even given different evaluation styles and different levels of evaluator consistency, averaged models of emotion perception could still outperform individual models. As we move towards a society with ever increasing computing power, we will begin to see emotionally personalized technology. These systems must be able to meet both the interaction needs and expectations of the users with whom they work. This necessitates an understanding and an ability to anticipate these preferences. Initially it may seem wise to model these expectations at a per user level. However, this work suggests that the variability of individuals with respect to their dimensional appraisal may lead to inaccuracies due to user self-misrepresentation. To mitigate this problem, it may be beneficial to adapt averaged models of user perception to accommodate individual users.

Additional work is needed to determine how to integrate and interpret raw user evaluations. Researchers [108] have suggested that new evaluation metrics should be created. Emotion evaluation experimental techniques should also be updated. This may lead to the creation of new emotional ground truthing techniques that are more evaluator-intuitive and evaluator-independent.

A weakness of the work presented in this chapter is the small number of dimensional evaluators considered. Future work includes incorporating the evaluations of additional evaluators. The IEMOCAP database contains both audio and facial motion capture information. Future work also includes utilizing the video information to improve the accuracies and to understand the temporal interaction between the audio information, video information, and user perception.

4.6 Work Published

The work presented in this chapter was published in the following article:

1. **Emily Mower**, Maja J Matarić, Shrikanth Narayanan. "Evaluating Evaluators: A Case Study in Understanding the Benefits and Pitfalls of Multi-Evaluator Modeling." In Proceedings of *International Speech Communication Association (InterSpeech)*. Brighton, England, September 2009.

Chapter 5

Emotion Profiling

The proper design of affective agents requires an a priori understanding of human emotion perception. Models used for the automatic recognition of emotion can provide designers with a method to estimate how an affective interface may be perceived given the feature modulations present in the stimuli. An understanding of the mapping between feature modulation and human perception can foster design improvements in the creation of emotionally relevant and targeted expressions for use in human-computer and human-robot interaction. This understanding will further improve human-centered design, necessary for wide-spread adoption of this affective technology [125].

Human perception of naturalistic expressions of emotion is difficult to compute. This difficultly is in part due to the presence of complex emotions, emotions containing shades of multiple affective classes [75,97,106,107]. In [75], the authors detail a scenario in which evaluators view a clip of a woman learning that her father will remain in jail. Human evaluators tagged these clips with labels including anger, disappointment, sadness, and despair [75]. The lack of emotional purity in natural expressions of emotion must be considered when designing systems to anticipate human perception of non-stereotypical

emotional speech. Classification systems designed to output one emotion label per input utterance may perform poorly if the expressions cannot be well captured by a single emotional label.

Naturalistic emotions can be described by detailing the presence/absence of a set of basic emotion labels (e.g., angry, happy, sad) within the data being evaluated (e.g., a spoken utterance). This multiple labeling representation can be expressed using Emotion Profiles (EP). EPs provide a quantitative measure for expressing the degree of the presence or absence of a set of basic emotions within an expression. They avoid the need for a hard-labeled assignment by instead providing a method for describing the shades of emotion present in the data. These profiles can be used in turn to determine a most likely assignment for an utterance, to map out the evolution of the emotional tenor of an interaction, or to interpret utterances that have multiple affective components.

EPs have been used within the community as a method for expressing the variability inherent in multi-evaluator expressions [108]. These EPs represent the distribution of reported emotion labels from a set of evaluators for a given utterance. The authors compared the entropy of their automatic classification system to that present in human evaluations. We introduced the notion of EPs for classification in a position paper [88]. We described EPs as a method for representing the emotion content of an utterance in terms of the phoneme-level emotion classification output over the utterance. These profiles described the percentage of phonemes classified as one of five emotion classes. In the current work, profiles are extended to represent emotion-specific classifier confidence. Thus, these new profiles can provide a more natural approximation of human

emotion, approximating blends of emotion, rather than time-percentage breakdowns of classification or reported evaluator perception.

EPs are an effective representation for emotion. In this chapter we present an implementation of emotion classification from vocal and motion-capture cues using EPs as an intermediary step. The data are modeled at the utterance-level where an utterance is defined as one sentence within a continuous speaker turn, or, if there is only one sentence in the turn, the entire speaker turn. The EPs are composed of four-binary support vector machine (SVM) classifiers, one for each of the emotions considered (anger, happiness, sadness, neutrality), used to create an estimate of the presence or absence of classes of emotion using classifier confidence. There are two methods that can be used to assign a final label. The first method assigns a class label based on the emotion class with the highest level of confidence, represented by the EP. The second method involves a secondary classification in which the profiles serve as a mid-level representation to a final classification stage. The first method is employed in this chapter. The second method is employed in the following two chapters.

Three data types of varying levels of ambiguity are used in the EP analyses. These data types are based on evaluator reports. They include unambiguous ("prototypical", total evaluator agreement), slightly ambiguous ("non-prototypical majority-vote consensus"), highly ambiguous ("non-prototypical non-majority-vote consensus"), and mixed ("full dataset", both total agreement and majority-vote consensus). We demonstrate that the use of feature selection in conjunction with EP representation results in an overall accuracy of 68.2% and an average of per-class accuracies (unweighted accuracy) of 64.5%, which is comparable to a previous audio-visual study resulting in an unweighted

accuracy of 62.4% [80]. The results are compared to a simplified four-way SVM in which confidences were not taken into account. In all cases, the overall accuracy of the presented method outperforms the simplified system. We also demonstrate that the EP-based system can be extended to interpret utterances lacking a well-defined ground truth. The results suggest that EPs can be used to discriminate between types of highly ambiguous utterances.

This work is novel in that it presents a classification system based on the creation of EPs and uses this technique to interpret emotionally ambiguous utterances. EPs represent emotions as complex blends, rather than discrete assignments. Furthermore, these profiles can be used to disambiguate the emotional content of utterances in expressions that would not otherwise be classified as a single expression of emotion.

5.1 Data Description

The database utilized in this study is the USC IEMOCAP database, collected at the University of Southern California [13]. The USC IEMOCAP database is an audio-visual database, augmented with motion capture recording. The data were evaluated using categorical and dimensional evaluation frameworks. There were at least three evaluators per categorical label and at least two per dimensional label. The database is fully described in Chapter 4, Section 4.1.1.

5.1.1 Emotion Expression Types

Emotional data can be described by the level of evaluator agreement. Thus, the data can be considered either as a cohesive whole or as merged sub-sets of data. The subsets

considered in this work are prototypical, non-prototypical majority-vote consensus ("non-prototypical MV"), and non-prototypical non-majority-vote consensus ("non-prototypical NMV"). These three emotional gradations are derivations of Russell's prototypical and non-prototypical definitions (Chapter 1, Section 1.3.4) and are used to describe the clarity of the emotion presentations.

The three emotion expression types are defined with respect to the categorical emotional evaluators. Prototypical emotion expressions are expressions with clear well-agreed upon emotional content. During the categorical emotion labeling task, these utterances were assigned a categorical label that is assigned by all evaluators' (e.g., for three evaluators, all three evaluators tagged the emotion "angry"). Non-prototypical MV emotions are utterances with identifiable, but ambiguous, emotional content. During the categorical evaluation, there was no single label at the intersection of all of the evaluators' assignments. However, these utterances were tagged by a majority of the evaluators with a single emotional label (e.g., two evaluators tagged an emotion as "angry" and one tagged the emotion as "disgusted"). The final emotional group, the non-prototypical NMV emotions were tagged with an inconsistent set of emotional labels (e.g., one evaluator tagged the emotion as "angry", another as "disgusted", and the final as "sad"). As a result, it is not possible to define a ground-truth label for this group of emotions. It is difficult to make a strong assertion regarding the prototypical or non-prototypical nature of an utterance since there are, on average, only three evaluators per utterance. However, the presented results suggest that the designations are representative of differing amounts of variability within the emotion classes.

Expression Type	Angry	Happy	Neutral	Sad	Total
Prototypical	284	708	121	309	1422
Non-prototypical MV	316	496	451	315	1578
Non-prototypical NMV 1L	173	17	47	45	282
Non-prototypical NMV 2L	174	142	350	174	420

Table 5.1: The distribution of the classes in the emotion expression types (note: each utterance in the 2L group has two labels, thus the sum of the labels is 840 but the total number of sentences is 420). There are a total of 3,000 utterances in the prototypical and non-prototypical MV group, and 3,702 utterances in total.

In the presented work the utterances considered are tagged with at least one emotion from the emotional set: angry, happy, neutral, sad, excited. In all cases, the classes of happy and excited were merged to combat data sparsity issues into a group referred to as "happy". In the prototypical and non-prototypical MV data, all the utterances had labels from this emotional set. In the non-prototypical NMV group, only utterances tagged by at least one evaluator as angry, happy, neutral, sad, or excited were considered (the classes of happy and excited were again merged). This group is described as either 1L, indicating that one of the labels is in the emotional set, 2L indicating that two of the labels are in this set, or nL indicating that there were more than two labels from the set. The 1L data was extremely biased towards the class of anger (Table 5.1) and there were only 80 utterances in the nL group, therefore, this study will focus only on the 2L emotions. Table 5.1 shows the distribution of the data across the three expression classes.

5.1.2 Data Selection

This work utilized a subset of the USC IEMOCAP database. During the data collection, only one actor at a time was instrumented with motion capture markers. This decision

allowed for an increase in the motion capture marker coverage on the actors' faces. Consequently, only half of the utterances in the database are accompanied by motion capture recordings.

The dataset size was further diminished by eliminating utterances without a single voiced segment. This eliminated utterances of sighs, breaths, and low whispers.

Finally, the dataset size was reduced by the evaluator reported affective label. As previously stated, all utterances analyzed in this chapter are tagged with at least one label from the set: angry, happy/excited, neutral, sad.

5.2 Audio-Visual Feature Extraction

The features utilized in this study were chosen for their perceptual relevance. The initial feature set contained audio and video (motion-capture extracted) features. All features were extracted at the utterance-level and were normalized for each subject using z-normalization. The feature set was reduced to create four emotion-specific feature sets using Information Gain.

5.2.1 Audio Features

The audio features include both prosodic and spectral envelope features. The prosodic features include pitch and energy. These features have been shown to be relevant to emotion perception [104, 106, 107, 119]. The spectral features include Mel Filterbank Coefficients (MFBs). As stated in the previous chapter, MFBs approximate humans' sensitivity to changes in frequencies. As the frequency of a signal increases, humans become less able to differentiate between two distinct frequencies. MFBs capture this

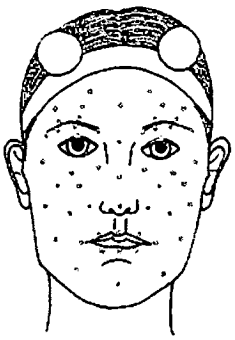
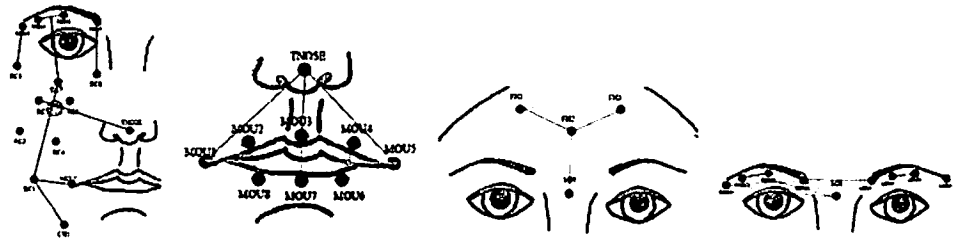


Figure 5.1: The location of the IR markers used in the motion capture data collection.

property by binning the signal with triangular bins of increasing width as the frequency increases. Mel Filterbank Cepstral Coefficients (MFCC) are commonly used in both speech and emotion classification. MFCCs are discrete cosine transformed (DCT) MFBs. The DCT decorrelates the feature space. Previous work has demonstrated that MFBs may contain more emotionally relevant information than Mel Filter Cepstral Coefficients (MFCC) across all phoneme classes, due to the lack of the final de-correlating step of the MFCC calculation [20].

5.2.2 Video Features

The definition of the video features was motivated by Facial Animation Parameters (FAPs). FAPs express distances (x,y,z) between points on the face. The features utilized in this study are based on the features found in [116], adapted to the current facial motion capture configuration. These features were extracted using motion capture markers (Figures 5.1 and 5.2). The cheek features include the distance from the top of the cheek to the eyebrow (approximating the squeeze present in a smile); the distance from the cheek to the mouth, nose, and chin; cheek relative distance features; and an average



(a) Cheek. (b) Mouth. (c) Forehead. (d) Eyebrow.

Figure 5.2: The FAP-inspired facial distance features utilized in classification.

position. The mouth features contain distances correlated with the mouth opening and closing, the lips puckering, and features detailing the distance of the lip corner and top of lip to the nose (correlated with smiles and frowns). The forehead features include features describing the relative distances between points on the forehead and the distance from one of the forehead points to the region between the eyebrows. The eyebrow features include features describing the up-down motion of the eyebrows, features describing eyebrow squish, and features describing the distance to the center of the eyebrows. Each distance is expressed in three features defining the x, y, and z-coordinates in space.

5.2.3 Feature Extraction

The utterance-length feature statistics include mean, variance, range, quantile maximum, quantile minimum, and quantile range. The quantile features were used instead of the maximum, minimum, and range because they tend to be less noisy. The pitch features were extracted only over the voiced regions of the signal. The video motion-capture derived features were occasionally missing values due to camera error or obstructions. To combat this missing data problem, the features were extracted only over the recorded

data for each utterance. These audio-visual features have been used in previous emotion classification problems [55].

The features were normalized over each speaker using z-normalization. The speaker mean and standard deviation were calculated over all of the speaker-specific expressions within the dataset (thus, over all of the emotions). Both the normalized and non-normalized features were included in the feature set.

5.2.4 Feature Selection

There were a total of 685 features extracted. However, there were only 3,000 prototypical and non-prototypical MV utterances utilized for testing and training. The feature set was reduced using Information Gain on a per emotion class basis (e.g., the features for the class of anger differed from those of happiness). Information gain describes the difference between the entropy of the labels in the dataset (e.g., "happy") and entropy of the labels when the behavior of one of the features is known (e.g., "happy" given that the distance between the mouth corner and nose is known) [82]. This feature selection method permits a ranking of the features by the amount of emotion-class-related randomness that they explain. The top features were selected for the final emotion-specific feature sets.

The feature selection was implemented in Weka, a Java-based data mining software package [120]. Information gain has previously been used to select a relevant feature subset in [93] and in the work discussed in Chapters 2 and 3. Information gain does not create an uncorrelated feature set, which is often preferable for many classification

Emotion	Cheek	Eyebrow	Forehead	Mouth	Energy	MFB
Angry	0.03	-	0.04	0.02	0.04	0.87
Happy	0.48	0.11	0.11	0.30	-	-
Neutral	0.48	10.0	0.10	0.28	-	0.05
Sad	-	-	-	-	0.04	0.96

Table 5.2: The *average* percentage of each feature over the 40 speaker-independent emotion-specific feature sets (10 speakers * 4 emotions).

algorithms. However, humans rely on a redundant and correlated feature set for recognizing expressions of emotions. Information gain was chosen to approximate the feature redundancy of human emotion processing.

The features were selected in a speaker-independent fashion. For example, the information gain for the emotion-specific features to be used for speaker 1 were calculated over a database constructed of speakers 2-10 using ten-fold cross-validation.

5.2.5 Final Feature Set

The number of features was determined empirically, optimizing for accuracy. The final feature set included the top 85 features (see Table 5.2 for the feature types selected) for each emotion class. The feature sets for anger and sadness are primarily composed of MFBs. The feature sets of happiness and neutrality are composed primarily of a mixture of cheek and mouth features. The high representation of audio features in the angry and sad feature sets and the low representation in the happy and neutral feature sets reinforce previous findings that anger and sadness are well captured using audio data while happiness is poorly captured using audio data alone [15, 88] .

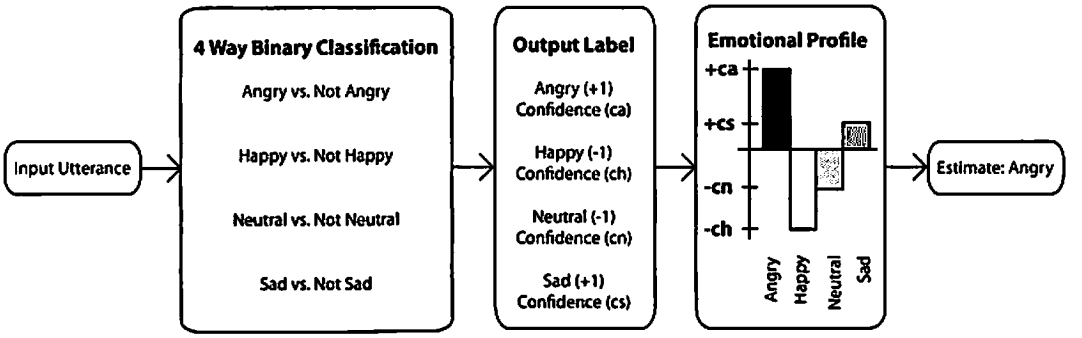


Figure 5.3: The EP system diagram. An input utterance is classified using a four-way binary classification. This classification results in four output labels representing membership in the class (+1) or lack thereof (−1). This membership is weighted by the confidence (distance from the hyperplane). The final emotion label is the most highly confident assessment.

5.3 Classification of Emotion Perception: Emotion Profile Support Vector Machine

The EP representation utilized in this chapter consists of four binary Support Vector Machines (SVM). The EPs were created using the four binary outputs and a measure of classifier confidence. The final label of the utterance is the most confident assignment in the EP (see Figure 5.3 for the system diagram and an example).

5.3.1 Support Vector Machine Classification

SVMs transform input data from the initial dimensionality onto a higher dimension to find an optimal separating hyperplane. SVMs have been used effectively in emotion classification [5, 15, 68, 86, 98]. The SVMs used in this study were implemented using Matlab's Bioinformatics Toolkit. The kernel used is a Radial Basis Function (RBF) with a sigma of eight, determined empirically. The hyperplane is found using Sequential

Minimal Optimization with no data points allowed to violate the Karush-Kuhn-Tucker (KKT) conditions (see [26] for a more detailed explanation of SVM convergence using the KKT conditions).

There were four emotion-specific SVMs trained using the emotion-specific (and speaker-independent) feature sets selected using information gain (Chapter 3, Section 3.2.2). Each of the emotional SVMs was trained discriminatively using a self vs. other training strategy (e.g., angry or not angry). The output of each of the classifications included a ± 1 and the distance from the hyperplane. This training structure is similar to the one utilized in [6], in which the authors estimated the emotion state of a set of speakers from a video signal. The authors transformed the distances from each of the self vs. other SVM classifiers into probability distributions using a softmax function. In the present work, the distances were not transformed because pilot studies demonstrated the efficacy of retaining the distance variations inherent in the outputs of each of the four emotion-specific SVM models. The models were trained and tested using leave-one-speaker-out cross-validation on the emotion-specific feature sets.

5.3.2 Creation of Emotional Profiles

The emotion profiles express the confidence of each of the four emotion-specific binary decisions. Each of the classifiers is trained using an emotion-specific feature set (e.g., the feature set for the angry classifier differs from that for the happy classifier). The outputs of each of these classifiers include a value indicative of how well the models created by each classifier fit the test data. This goodness of fit measure can be used to assess which model fits the data most accurately.

The SVM goodness of fit measure used in this study is the raw distance from the hyperplane. SVM is a maximum margin classifier whose decision hyperplane is chosen to maximize the separability of the two classes. The distance from the margin of each emotion-specific classifier provides a measure of the classifier confidence. The profile components are calculated by weighting each emotion-specific classifier output ± 1 by the absolute value of the distance from the hyperplane (the goodness of fit measure). The EPs are representative of the confidence of each binary yes-no emotion class membership assignment.

The intuition behind this decision comes from the nature of the SVM classifier. SVM identifies a class label using position relative to a separating boundary. Data points that are close to the boundary suggest that, in the feature space (or projected feature space), the class label of the data points are more easily confused than points further away from the boundary. Points that lie far from the separating hyperplane are examples of data that are more differentiable, or are less confusable examples of a given class, than data that lie close to the hyperplane. For example, in the binary angry classification task a point that is far from the decision hyperplane may be a strong example of "angry" suggesting that the data point is in fact not "not angry."

Experimental results demonstrate that the raw distances to the hyperplane are effective measures of the strength of emotion. The accuracy of this statement can be assessed by analyzing the distance to the hyperplane as a function of location in a valence-activation plot. If the raw distance to the hyperplane is an appropriate measure one would expect to see that utterances in regions associated with strong expressions of one of the

basic emotions considered will be further away from the hyperplane than utterances located in other regions of the valence-activation space. In Figure 5.4 the raw distances to the hyperplane are plotted in the valence-activation space for Speaker 1. The four figures represent the distance to the hyperplane for the four emotional components of the EP (anger, happiness, neutrality, and sadness). In the figure, dark red represents distances associated with a strong assertion of class membership while dark blue represents distances associated with a strong rejection of class membership. For example, in Figure 5.4, the "Happy Component" subplot (upper right) shows that utterances that are positively valenced (ranging from calm happiness to excitation) are dark red, the "Angry Component" subplot (upper left) demonstrates that the negatively valenced and highly activated utterances are dark red, and the "Sad Component" (lower right) subplot illustrates that the negatively valenced utterances with low activation are dark red. The "Neutral Component" subplot (lower left) shows that the utterances with neutral valence and lower activation are dark red. The classification between neutral and sad data is notoriously difficult. The comparison between the neutral and sad subplots illustrate that even given a component representation the classes remain confusable.

SVMs were chosen for the EP backbone based on experimental evidence suggesting that this algorithm had the highest performance when compared to other discriminative techniques. Thus, the main results are presented using the SVMs as a backbone. However, EPs can be created using K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), or any classifier that returns a goodness of fit measure (including generative classifiers). Both KNN and LDA have been used in emotion recognition studies [6,33,63].

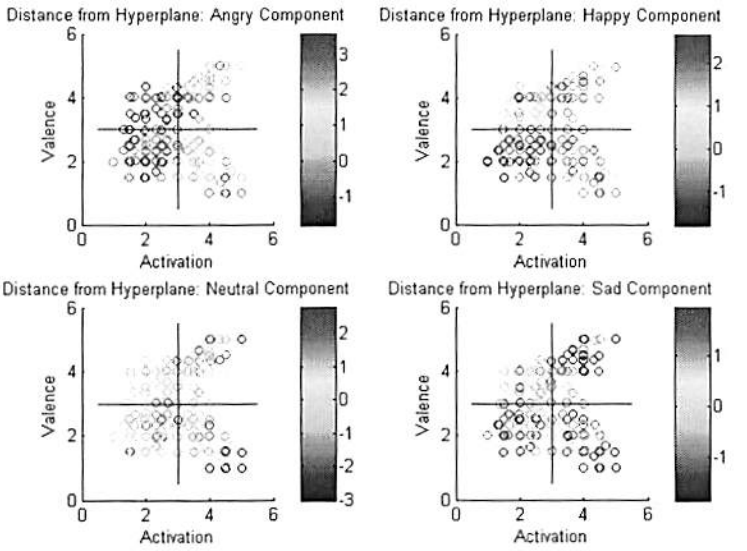


Figure 5.4: The raw distances to the hyperplane for the four emotional components of the EP.

5.3.3 Final Decision

An emotional utterance is assigned to an emotion class depending on the representation of the emotions within the EP. The inherent benefit of such a classification system is that it can handle assessments of emotional utterances with ambiguous emotional content. When the emotional content of an utterance is unclear, the four binary classifiers may return a value suggesting that the emotion is not a member of any of the modeled emotion classes. Absent an EP-based system, it would be difficult to assign an emotional label to such an utterance. However, even in this scenario, it is possible to reach a final emotion assignment by considering the confidences of each of the no-votes.

By definition, ambiguous or non-prototypical emotional utterances fit poorly in the categorical emotion classes. This mismatch may be because the emotional content is from

an emotion class not considered. It may also be because the utterance contains shades of multiple subtly-expressed emotion classes. However, the EP-based classifier is able to recognize these slight presences by a low-confidence rejection. Consequently, even given four no-votes, a final emotion assignment can still be made.

The neutral emotion class is difficult to classify because there exists a large variability in the emotion expression expressed within this class. Neutral expressions may be colored by shades of anger, happiness, or sadness. Evaluators also may assign a class of neutrality when no other emotion is distinctly expressed. EPs can also be used to capture this phenomenon. If all of the rejection confidences, described by the EP, are above a threshold, then there is a strong indication that there is no clear emotion expressed. These utterances are assigned to the class of neutrality. This threshold is defined by calculating the mean and subtracting one standard deviation of the confidences over each of the emotions. If the EP indicated that the emotion with the highest confidence is chosen with a no-vote confidence outside this threshold (i.e. the profile expressed high-confidence that the utterance did not contain that emotion), it is assumed that there is no clear emotion present, and the utterance is assigned to the class of neutrality. This neutral assignment method is similar to the one implemented in [81].

5.4 Results and Discussion: the Prototypical, Non-Prototypical MV, and Mixed Datasets

The results presented describe the system performance over utterances labeled as angry, happy (the merged happy–excited class), neutral, or sad. The results are divided into

three categories: general results, prototypical emotion results, and non-prototypical MV results.

The general, prototypical, and non-prototypical results are compared to a baseline classification system and chance. The baseline is a simplified version of the EP classifier. In this baseline, instead of utilizing the EP representation (weighting the output by the distance from the boundary), the decisions are made using three steps. If only one classifier returns a value of +1, then the emotion label is assigned to this class. If multiple classifiers return +1, the utterance is assigned to the selected class with the higher prior probability. If no classifiers return +1, the emotion is assigned to the class with the highest prior probability (of the four emotion classes).

The baseline represents SVM classification without considering relative confidences. Emotion is often expressed subtly. This subtle expression of emotion is often not well recognized by classifiers trained to produce a binary decision (acceptance vs. rejection). The comparison between the EP and the baseline will demonstrate the importance of considering the confidence of classification results (e.g., a weak rejection by one of the classifiers may indicate a subtle expression of emotion, not the absence of the emotion) rather than just the binary result. The chance classification result assigns all utterances to the emotion most highly represented within the four (i.e., general, prototypical, and non-prototypical MV) data sub-sets.

5.4.1 General Results

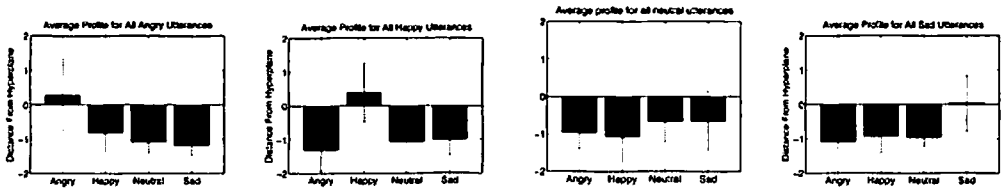
The first set of classification results is obtained by training and testing on the full dataset (prototypical and non-prototypical MV utterances). The overall classification accuracy

Data Type	Emotion	Precision	Recall	F
Full EP	Angry	0.67	0.75	0.71
	Happy	0.77	0.81	0.79
	Neutral	0.54	0.28	0.37
	Sad	0.60	0.75	0.67
	Weighted: 0.68		Unweighted: 0.65	
Baseline	0.59			
Prot EP	Angry	0.75	0.80	0.77
	Happy	0.89	0.88	0.88
	Neutral	0.65	0.34	0.45
	Sad	0.76	0.89	0.82
	Weighted: 0.82		Unweighted: 0.72	
Baseline	0.76			
NonProt MV EP	Angry	0.58	0.71	0.64
	Happy	0.60	0.70	0.65
	Neutral	0.46	0.39	0.42
	Sad	0.55	0.41	0.47
	Weighted: 0.55		Unweighted: 0.55	
Baseline	0.42			

Table 5.3: The EP and baseline classification results for three data divisions: full (a combination of prototypical and non-prototypical MV), prototypical, and non-prototypical MV. The baseline result (simplified SVM) is presented as a weighted accuracy.

using the EP representation is 68.2% (Table 5.3). This outperforms both chance (40.1%) and the simplified SVM (55.9%). The difference between the EP method and baseline method is significant at $\alpha \leq 0.001$ (difference of proportions test). The unweighted accuracy (an average of the per-class accuracies) is 64.5%. This result is comparable to the work of Metallinou et al. [80] (described in Chapter 1, Section 1.3.4) with an unweighted accuracy of 62.4%, demonstrating an efficacy of the approach for a dataset with varying levels of emotional ambiguity.

The average profiles for all utterances demonstrate that in the classes of angry, happy, and sad there is a clear difference between the representation of the reported and non-reported emotions within the average profiles (Figure 5.5). All four profiles demonstrate



(a) Angry. (b) Happy. (c) Neutral. (d) Sad.

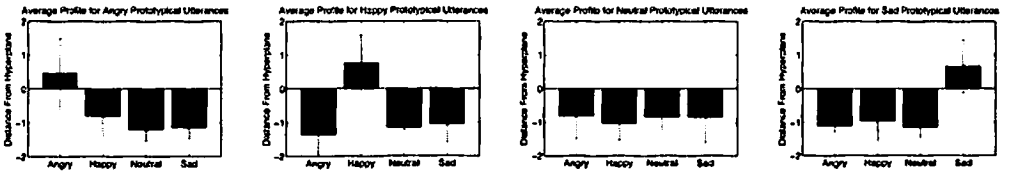
Figure 5.5: The average emotional profiles for all (both prototypical and non-prototypical) utterances. The error bars represent the standard deviation.

the necessity of considering confidence in addition to the binary yes-no label in the classification of naturalistic human data. For example, the angry EP indicates that even within one standard deviation of the average confidence the angry classifier returned a label of "not angry" for angry utterances. The use of and comparison between the four emotional confidences allowed the system to determine that, despite the lack of a perfect match between the angry training and testing data, the evidence indicated that the expressed emotion was angry (F-measure = 0.71).

As mentioned earlier, the EP technique can be performed using a variety of classification algorithms. The results are presented using an SVM backbone. Results can also be presented for an EP-KNN (k = 35, 66.4%) and an EP-LDA (diagonal covariance matrix, 60.3%).

5.4.2 Prototypical Classification

The prototypical classification scenario demonstrates the ability of the classifier to correctly recognize utterances rated consistently by evaluators. The overall accuracy for the prototypical EP classifier was 81.7% (Table 5.3). This outperformed chance (49.8%) and the simplified SVM (75.5%). The difference between the EP and baseline is significant at



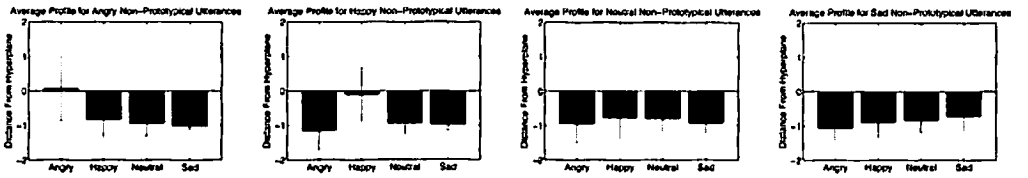
(a) Angry. (b) Happy. (c) Neutral. (d) Sad.

Figure 5.6: The average emotional profiles for prototypical utterances. The error bars represent the standard deviation.

$\alpha \leq 0.001$ (difference of proportions test). The high-performance of the simplified SVM is due in part to the prevalence of happiness in the prototypical data (49.8%). This bias affected the final results because both ties were broken and null-results were converted to a class assignment using class priors.

The simplified SVM left 391 utterances unclassified (all classifiers returned -1), representing 27.5% of the data.

The average profiles for prototypical utterances (Figure 5.6) demonstrate that there is a difference between the representation of the reported emotion and non-reported emotions in the EPs for the classes of angry, happy, and sad. The barely-differentiated neutral EP clearly demonstrates the causes behind the poor classification performance of the neutral data. The performance increase in the angry, happy, and sad classifications can be visually explained by comparing Figures 5.5(a) to 5.6(a), 5.5(b) to 5.6(b), and 5.5(d) to 5.6(d). The mean confidence value for the angry, happy, and sad data were higher when training and testing on prototypical data.



(a) Angry. (b) Happy. (c) Neutral. (d) Sad.

Figure 5.7: The average emotional profiles for non-prototypical utterances. The error bars represent the standard deviation.

5.4.3 Non-prototypical Majority-Vote (MV) Classification

The classification of non-prototypical MV utterances using EPs resulted in an overall accuracy of 55.4%. This accuracy is particularly of note when compared to the simplified SVM baseline classification whose overall accuracy is 42.2%. This difference is significant ($\alpha \leq 0.001$, difference of proportions test). The EP also outperforms chance (31.4%). The class-by-class comparison can be seen in Table 5.3. In 62.3% of the data (983 utterances), none of the binary classifications in the simplified SVM classifier returned any values of +1. This indicates that the four-way binary classification alone is not sufficient to detect the emotion content of ambiguous emotional utterances. In the EP method there is a higher level of confusion between all classes and the class of neutrality. This suggests that the emotional content of utterances defined as "neutral" may not belong to a well-defined emotion class, but may instead be representative of the lack of any clear and emotionally meaningful information.

The average profiles for non-prototypical MV utterances (Figure 5.7) demonstrate that the EP representation strongly differentiates between reported and non-reported emotions given non-prototypical MV data in the classes of anger and happiness. The non-prototypical EPs also provide additional evidence for the importance of comparing

the confidences in emotional assessments between multiple self vs. other classification schemes. The simplified baseline demonstrated that in 62.3% of the data all four binary classifiers returned non-membership labels, indicating that, in this subset, the feature properties of the training and testing data differ more markedly here than in the prototypical training–testing scenario. However, the similarity between the properties of a specific emotion class in the training data were closer to those of the same emotion class in the testing data, rather than a different emotion class. This suggests that the EP-based method is more robust to the differences in within-class emotional modulation than conventional SVM techniques.

The neutral classification of the non-prototypical MV data was more accurate than that of either the prototypical or full datasets. The neutral EP modeled using the non-prototypical MV data (Figure 5.7(c)) was better able to capture the feature properties of the neutral data than those modeled using either the prototypical or full data (compare to Figures 5.6(c) and 5.7(c)). This suggests that it may be beneficial to create models based on emotionally variable data (e.g., the non-prototypical MV data) when considering inherently ambiguous emotion classes, such as neutrality.

5.4.4 Emotion Profiles as a Minor Emotion Detector

The previous sections demonstrated that EP-based representations can be used in a classification framework. This section will assess the ability of the EPs to capture both the majority and minority reported emotions (e.g., for the rating "angry-angry-sad", the major emotion is anger and the minor emotion is sadness). In this assessment, the EP is trained and tested on the non-prototypical MV data.

The ability of the profile to correctly represent the major and minor emotions is studied in two ways. First, using utterances whose major emotion was correctly identified by the EP and whose minor emotion is from the set of angry, happy, neutral, and sad and second, using utterances whose major and minor emotions were the two most confident assessments (in either order). There are a total of 748 utterances with minor emotions in the targeted set. Utterances with minor labels outside of this set were not considered as the EPs only include representations of anger, happiness, neutrality, and sadness confidences and cannot directly represent emotions outside of this set.

The major-minor emotion trends can be seen in Table 5.3(a). The proportion of the non-prototypical MV emotions with secondary emotions from the considered set differs with respect to the majority label. For example, 81.04% of the original happy data is included in the new set while only 6.01% of the angry data has secondary labels in the set. The most common secondary label for the angry data is frustration (74.05%), an emotion not considered in this study due to a large degree of overlap between the classes. The distribution of the secondary emotions suggests that the most common combination in the considered affective set is a majority label of happy and a minority label of neutral (Table 5.3(a)). This combination represents 51.06% of the major-minor emotion combinations in the considered set. It should also be noted that across all major emotions, the most common co-occurrence emotion was neutrality.

In an ideal case, the EP would be able to represent both the majority and the minority emotions correctly, with the majority emotion as the most confident assessment and the minority emotion as the second most confident assessment. There are a total of 211 profiles (28.2% of the data) that correctly identify the major and the minor emotions.

(a) Total number of emotions with secondary labels in the angry, happy, neutral, sad set.

Major ↓	Angry	Happy	Neutral	Sad	Total
Angry	-	4	13	2	19
Happy	4	-	382	16	402
Neutral	12	129	-	56	197
Sad	11	25	94	-	130
Total	27	158	489	74	748

(b) Results where the major and minor emotions are correctly identified

Major ↓	Angry	Happy	Neutral	Sad	Total
Angry	-	2	6	1	9
Happy	1	-	136	5	142
Neutral	1	9	-	19	29
Sad	1	2	28	-	31
Total	3	13	170	25	211

(c) Results where the major and minor emotions were both in the top two reported labels

Major ↓	Angry	Happy	Neutral	Sad	Total
Angry	-	2	6	1	9
Happy	2	-	161	6	169
Neutral	2	28	-	33	63
Sad	1	11	47	-	59
Total	5	41	214	40	300

Table 5.4: The major–minor emotion analysis.

Over the non-prototypical MV data, there were 406 utterances with a correctly identified major label. Thus, the 211 profiles represent 52.0% of the correctly labeled data. This indicates that the majority of profiles that correctly identified the major emotion also correctly identified the minor emotion. This suggests that EPs can accurately assess emotionally clear and emotionally subtle aspects of affective communication. The major-minor pairing results can be found in Table 5.3(b).

In emotion classification a commonly observed error is the swapping of the major and minor emotions (i.e., the major emotion is reported as the minor and vise versa). This phenomenon was also studied. Table 5.3(c) presents the emotions whose major and

minor emotions were recognized in the two most confidently returned emotion labels (in either order). The results demonstrate that of the utterances with minor emotions in the target affective set, the EPs identified both the major and minor emotions in the top two components 40.1% of the time. This percentage varied across the major labels. The angry non-prototypical MV data had both components recognized in 47.4% generated EPs, while they were both represented in 42.0% of the happy EPs, 32% of the neutral EPs, and 45% of the sad EPs.

These results suggest that the EP technique is capable of representing subtle emotional information. It is likely that this method does not return a higher level of accuracy because the expression of the major emotion was already subtle. Therefore, the expression of the minor emotion was not only subtle, but not observed by all evaluators. Therefore, this minority assessment may have been due to a characteristic of the data or the attention level of the evaluator. In this light, the ability of the EP method to capture these extremely subtle, and at times potentially tenuous, bits of emotional information should further recommend the method for the quantification of emotional information.

5.5 Results and Discussion: the Non-Prototypical NMV Dataset

One of the hallmarks of the EP method is its ability to interpret ambiguous utterances. EPs can be used to detect whatever information is possible given inherently ambiguous data. In this chapter, the goal is to utilize utterances that have at least one label from

the target emotional set and to identify at least one of the emotions reported by the evaluators.

The non-prototypical NMV utterances have no majority-voted label. These utterances were labeled with one (or more) of the labels from the set: angry, happy, neutral, sad. No ground-truth can be defined for these utterances because there was no evaluator agreement. EPs are ideally suited to work with this type of data because they provide information describing the emotional makeup of the utterances rather than a single hard label.

Two experiments were conducted on the non-prototypical NMV data. The first experiment was a classification study in which the non-prototypical NMV data were classified using models trained on the full dataset, the prototypical data only, and the non-prototypical MV data only. This study determines how well suited the EP representation method, trained on labeled data, is for recognizing ambiguous emotional content. This problem is difficult because the classifiers must be able to identify the categorical emotion labels when the evaluators themselves could not. The evaluator confusion implies that the feature properties of the utterances are not well described by a single label. The second experiment was a statistical study designed to understand the differences in the representations of the emotions within the EPs. This study provides evidence validating the returned EPs. It demonstrates the differences that exist between EPs of specific ambiguous emotion classes. These results suggest that the EP method returns meaningful information in the presence of emotional noise.

There were a total of 420 non-prototypical NMV 2L utterances considered in the two experiments.

5.5.1 Experiment One: Classification

In the classification study, three train-test scenarios were analyzed. In each study, the modeling goal was to recognize at least one of the labels tagged by the evaluators in the 2L dataset using the EPs. This modeling demonstrates the ability of the EPs to capture emotional information in the presence of highly ambiguous emotional content. Classifier success is defined as the condition in which the classified emotion (the top estimate from the EP) is in the set of labels reported by the categorical evaluators. In the 2L dataset, there are two possible correct emotions, as explained in Section 5.1.

There were three training-testing scenarios. In the first scenario, the models were trained on all the full data (prototypical and non-prototypical MV). In the second scenario, the models were trained on only prototypical utterances. In the final experiment, the models were trained only on non-prototypical MV utterances. The three experiments analyze the generalizability of the models trained on utterances with varying levels of ambiguity in expression.

The results demonstrate (Table 5.5) that the emotional profiling technique is able to effectively capture the emotion content inherent even in ambiguous utterances. In all results presented, the per-class evaluation measure is precision, and the overall measure is accuracy. This decision was motivated by the evaluation structure. The goal of the system is to correctly identify either one or both of the two possible answers. Consequently, a per-class measure that necessitates a calculation of all of the utterances tagged as a certain emotion is not relevant, because two components of the EP would then be in direct opposition for the per-class accuracy measure. However, accuracy over the entire

Dataset	Train	Angry	Happy	Neutral	Sad	Accuracy
2L	All	0.76	0.61	0.90	0.65	0.71
	Prot	0.77	0.60	0.88	0.59	0.65
	Non-prot	0.70	0.55	0.89	0.68	0.73
	Baseline			0.42		

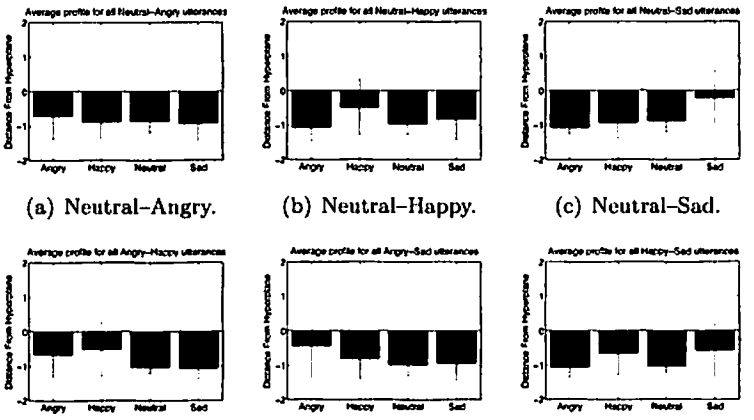
Table 5.5: The results of the EP classification on the 2L non-prototypical NMV data. The results are the precision, or the percentage of correctly returned class designations divided by the total returned class designations.

classified set is relevant because a classifier that returns either of the two listed classes can be defined as performing correctly. The chance accuracy (assigning a specific utterance to the class with the highest prior probability) the 2L dataset was calculated by finding the emotion that co-occurred with the other emotion labels most commonly. The class of neutrality occurred in 41.6% of the labels. Thus, chance was 41.6%.

The maximal accuracy of the 2L dataset was achieved in the non-prototypical MV training scenario with 72.6% (Table 5.5). The class-by-class precision results demonstrate that specific data types are more suited to identifying the affective components of emotionally ambiguous utterances.

The results indicate that anger was precisely identified 70-77% of the time. This is of particular note because in this data, humans could not agree on the label; yet, when training with the non-prototypical MV data, the EP could accurately identify the presence of anger.

The results further indicate that the EP is able to reliably detect one of the emotional labels of the utterances from the 2L dataset. The overall accuracy of 72.6% is far above the chance accuracy of 41.6%. Furthermore, since the chance classifier is only capable of



(a) Neutral–Angry. (b) Neutral–Happy. (c) Neutral–Sad.

(d) Angry–Happy. (e) Angry–Sad. (f) Happy–Sad.

Figure 5.8: The average emotional profiles for the non-prototypical NMV utterances. The error bars represent the standard deviation.

detecting neutrality, this supports the more precise detection of the EP over a range of emotions.

The EP method is able to capture information that cannot be captured by the simplified baseline SVM discussed earlier. In Figure 5.8 all of the histograms demonstrate that, on average, all four binary classifiers return non-membership results (-1). The confidence component allows the EP to disambiguate the subtle emotional content of the non-prototypical NMV utterances. The average profiles of Figure 5.8 demonstrate that the EPs are able to capture the emotion content of these utterances.

The EP method is able to capture information that cannot be captured by the simplified baseline SVM discussed earlier. In Figure 5.8 the majority of the histograms demonstrate that, on average, all four binary classifiers return non-membership results (-1). The confidence component allows the EP to disambiguate the subtle emotional content

of the non-prototypical NMV utterances. The average profiles of Figure 5.8 demonstrate that the EPs are able to capture the emotion content of these utterances.

5.5.2 Experiment Two: ANOVA of EP based representations

In this statistical study, two ANOVA analyses are performed on the profiles to determine the statistical significance of the representations of the emotions within the profiles. These studies investigate the ability of the EPs to differentiate between the reportedly present and absent emotional content.

The results presented in this section are two-tailed ANOVAs. These analyses were performed on the *2L dataset* with EP models trained using the full dataset (prototypical and non-prototypical MV). This study will demonstrate that EPs are able to capture multiple reported emotions. In the results described below, the two reported labels for an utterance in the 2L dataset will be referred to as the ***co-occurring*** labels or group (e.g., neutral and angry). Labels that are not reported are referred to as the ***non-occurring*** labels or group (e.g., happy and sad). Each ANOVA analysis studies sets of EPs grouped by the co-occurring emotions (e.g., the neutral–angry group). These groups will be referred to as ***EP-sets***.

The first analysis studies the representation of pairs of emotions in individual EP-sets by comparing the co-occurrence (e.g., neutral and angry) group mean to the non-occurrence (e.g., happy and sad) group mean. This study asks whether the EP representation is able to capture the difference between reportedly present and absent emotions. This analysis will be referred to as the *Individual EP-Set* experiment.

Co-occurring emotions		P-value
Emotion 1	Emotion 2	
Neutral	Angry	-
Neutral	Happy	**
Neutral	Sad	**
Angry	Happy	
Angry	Sad	***
Happy	Sad	

Table 5.6: ANOVA analysis of the difference in group means between co-occurring and non-occurring emotions within an EP-set (Individual EP-set experiment). (- = $\alpha \leq 0.1$, * = $\alpha \leq 0.05$, ** = $\alpha \leq 0.01$, *** = $\alpha \leq 0.001$)

The *Individual EP-Set* experiment demonstrates that in general the representation of the co-occurrence group in an EP-set differs from that of the non-occurrence group. In the angry–sad EP-set this difference was significant at $\alpha \leq 0.001$, in the neutral–happy and neutral–sad, this difference was significant at $\alpha \leq 0.01$. In the neutral–angry, this difference was significant at $\alpha \leq 0.1$. In the angry–happy and happy–sad EP-sets, this difference was not significant. This suggests that in the majority of the cases, the individual EP-sets were able to differentiate between the presence and absence of the co-occurrence labels in the emotional utterances (Table 5.6).

The next study builds on the results of the *Individual EP-Set* results to determine if the representation of these co-occurring emotion groups differs between their ***native*** EP-set and a different (***non-native***) EP-set (e.g., compare the representation of neutral and angry in the neutral–angry EP-set to the neutral–angry representation in the happy–sad EP-set). This will be referred to as the *Group* experiment.

The *Group* experiment found that in most cases, the co-occurrence group mean differed between the native EP-set and the non-native EP-sets when the co-occurrence emotions of the non-native set was disjoint from the co-occurrence emotions of the native

set. This was observed most starkly with the Angry–Sad EP-set. The representation of the co-occurrence emotions differed from their native EP-set only when compared with their representation in the Neutral–Happy EP-set, sets where the co-occurrence emotions were entirely disjoint. This demonstrates that EP-sets must be differentiated based on more than their co-occurrence emotions (Table 5.7).

The following two analyses determine if the representation of the individual co-occurring emotions differs between the native and non-native sets. These will be referred to as the Emo_1 and Emo_2 experiments (e.g., compare the representation of neutral in the neutral–angry EP-set to the neutral representation in the happy–sad EP-set).

The Emo_1 and Emo_2 experiments demonstrate that the difference in the representation of the individual co-occurrence emotions of anger, happiness, and sadness between their native and non-native EP-sets occurs most frequently and most significantly when the co-occurrence emotion pair is neutrality (Table 5.7).

The *Individual EP-Set*, *Group*, Emo_1 and Emo_2 analyses demonstrate that aspects of the EP-sets are differentiable. The final analysis compares the differences between the EP-set representations as a whole. The result is an interaction term between the analysis of the difference between the representation of each emotion in the EP-sets and the difference between the two EP-sets' values when grouped together (e.g., compare the neutral–angry EP-set to the happy–sad EP-set). This will be referred to as the EP_1 *vs.* EP_2 experiment.

The EP_1 *vs.* EP_2 experiment demonstrates that in 26 of the 30 cases, the representation of the EP-sets differs significantly between the sets. Furthermore, the cases in which the EP-sets are not significantly different occur when the emotions represented by the

two co-occurrence pairs share a similar co-occurrence emotion. The co-occurrence pairs of angry–happy and angry–sad can both represent tempered anger. Consequently, the EP-sets' inability to strongly differentiate between the two emotion types should not be seen as a failure, but instead as the EP-sets' ability to recognize the inherent similarity in the emotion expression types (Table 5.7).

These results suggest that certain EP-sets distinctly represent the underlying emotions reported by evaluators. This further suggests that these EP-sets (rather than a single confident label) can be used during classification to detect the differences between ambiguous emotional utterances. This application of EP-sets will be explored in future work.

5.6 Conclusion

Natural human expressions are combinations of underlying emotions. Models aimed at automated processing of these emotions should reflect this aspect. Conventional classification techniques provide single emotion class assignments. However, this assignment can be very noisy when there is not a single label that accurately describes the presented emotional content. Instead these utterances should be described using a method that identifies multiple emotion hypotheses. If a single label is necessary, it can be divined from the information contained within the EP. However, when a hard label is not required, the entirety of the emotional content can be retained for higher-level emotional interpretation.

The EP technique performs reliably both for prototypical and non-prototypical emotional utterances. The results also demonstrate that the presented EP-based technique can capture the emotional content of utterances with ambiguous affective content. EPs provide a method for describing the emotional components of an utterance in terms of a pre-defined affective set. This technique provides a method for either identifying the most probable emotional label for a given utterance or the relative confidence of the available emotional tags.

The neutral emotion class is difficult to classify because there exists a wide range in the variability of emotion expressed within this class. Neutral expressions may be colored by shades of anger, happiness, or sadness. Evaluators also may assign a class of neutrality when no other emotion is distinctly expressed.

One of the strengths of the EP-based method is its relative insensitivity to the selection of the base classifier. This study presented results utilizing an SVM four-way binary classifier. The SVM classifier can be replaced by any classifier that returns a measure of confidence. The results demonstrate that other classification methods (KNN, LDA) can also serve as the backbone of an EP-based method.

Future work will include several investigations of the utility of the EP-representation. In the presented work, the final emotion assessments are made by selecting the most confident emotion assessment from the generated profile. However, this does not take into account the relationship between the individual emotional components of the EP. Chapters 6 and 7 will investigate classification of the generated profiles. Frustration is not included in either the EP testing or training. Chapter 6 will also investigate whether frustration should be included as a component of the profile, or if the EP representation is

sufficiently powerful to represent frustration without including it as a component. Finally, Chapter 7 will also investigate the utility of emotion-based components for representation, rather than data-driven clusters as the relevant components in the profile construction. These analyses will provide further evidence regarding the efficacy of profile-based representations of emotion.

This chapter presented a novel Emotional Profiling method for automatic emotion classification. The results demonstrate that these profiles can be used to accurately interpret naturalistic and emotionally ambiguous human expressions and to generate both hard- and soft-labels for emotion classification tasks. Furthermore, EPs-based methods are relatively robust to classifier selection. Future work will include utilizing EPs to interpret dialog-level emotion expression and utilizing EPs for user-specific modeling.

5.7 Work Published

The work presented in this chapter was published in the following articles:

1. **Emily Mower**, Maja J Matarić and Shrikanth S. Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotional Profiles." *IEEE Transactions on Audio, Speech and Language Processing.* Accepted for publication, August 2010.
2. **Emily Mower**, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, Shrikanth Narayanan. "Interpreting Ambiguous Emotional Expressions." In Proceedings of *ACII Special Session: Recognition of Non-Prototypical Emotion from Speech- The Final Frontier?.* Amsterdam, The Netherlands, September 2009.

EP-Set 1		EP-Set 2		Co-occurring			EP_1 vs. EP_2
Emo_1	Emo_2	$Other_1$	$Other_2$	Group	Emo_1	Emo_2	(Interaction Term)
Neu	Ang	Neu	Hap	***	*	***	***
		Neu	Sad	***	-	-	***
		Ang	Hap		*		**
		Ang	Sad		-	-	*
		Hap	Sad	**	-	-	***
Neu	Hap	Neu	Ang	***	*	***	***
		Neu	Sad	***	*	***	***
		Ang	Hap				*
		Ang	Sad	*		*	***
		Hap	Sad				
Neu	Sad	Neu	Ang	***		***	***
		Neu	Hap	***	*	***	***
		Ang	Hap	***	*	***	***
		Ang	Sad	***	-	***	***
		Hap	Sad	*	-	-	**
Ang	Hap	Neu	Ang	*		**	**
		Neu	Hap		***		*
		Neu	Sad	***	***	**	***
		Ang	Sad				
		Hap	Sad	-	*		**
Ang	Sad	Neu	Ang		-		*
		Neu	Hap	**	***		***
		Neu	Sad		***	***	***
		Ang	Hap				
		Hap	Sad		*	*	**
Hap	Sad	Neu	Ang	**	-	*	***
		Neu	Hap				
		Neu	Sad		*	-	**
		Ang	Hap			*	**
		Ang	Sad	*		*	**

Table 5.7: ANOVA analyses of the differences between reported emotions in profiles in which they were reported vs. profiles in which they weren't. Note that the EP_1 vs. EP_2 is an interaction of an ANOVA analysis of the set EP_1 vs. EP_2 and an ANOVA analysis of the representation of the individual emotions in each EP-set. (- = $\alpha \leq 0.1$, * = $\alpha \leq 0.05$, ** = $\alpha \leq 0.01$, *** = $\alpha \leq 0.001$)

Chapter 6

The Robustness of Emotion Profiling

In the previous chapter EPs were shown to be effective representations for emotional utterances. The components of the profile allowed for a descriptive characterization of the emotion components present in the utterance. However, it is necessary to demonstrate that the EPs can also represent the component properties of out-of-domain emotions. Ambiguous emotional expressions are a natural part of human communication. Consequently, in a human-machine interaction (HMI), a system's affective awareness capabilities are limited both by its ability to recognize emotions on which it has been trained and to reconcile emotions that it has not previously observed. This chapter will assess the ability of EPs to discriminatively represent emotions unseen during training.

Emotion classification requires the quantification of affective utterances via mathematical representation. These representations attempt to disambiguate affective data by maintaining the flexibility needed to capture the essence of the expression while allowing for the variance inherent in human emotions. However, during an interaction with a human, a system will invariably be faced with representing an emotion unseen during its training. The representation employed by the machine must be able to capture the

emotional content of the data in a way that will allow for future classification, even if the emotional category has not previously been observed. This ability to characterize utterances may allow future HMI systems to adapt to the emotion speaking style of their users.

EPs have been used to fuse different modalities in classification [80]. EP-like representations have also been used to represent the evaluations of a set of evaluators [57, 107] and to represent perception based on actions (as a function of multiple emotions) [23]. In this chapter, we will further analyze this technique to study the ability of this technique to represent out-of-domain data.

In the previous chapter the EPs were four-dimensional and the classification was four-way. In this chapter the utility of adding additional representations will be explored. The driving hypothesis is that the four emotional components of the EP should be able to represent emotions that are combinations of the components. This chapter will demonstrate that four semantically meaningful "ideal" clusters (e.g., angry, happy) can be used to represent five separate emotion categories, using frustration as a case study. The EPs will again be composed of angry, happy, neutral, and sad emotional components with an optional frustration component. The comparison of the four and five dimensional EP classifications will demonstrate the robustness of the EPs for the representation of out-of-domain emotional data. The classification accuracies did not significantly differ between the four and five dimensional EP classifications suggesting that EPs need only contain the emotions necessary to "span" the emotional space. The classification results and statistical analyses presented in this chapter suggest that EPs are a robust representation for emotion, both in- and out-of-domain.

The results demonstrate that the EP-representation can be effectively used to characterize the data in an n-way (where $n = 4, 5$) speaker-dependent emotion classification task using Naïve Bayes. This speaker-dependent classification is representative of the user personalization component inherent in long-term human-machine interaction. The presented classification framework obtains an accuracy of 68.43% over the four-class emotion classification problem (angry, happy, neutral, and sad) over the full dataset. However, its true power lies in its ability to characterize emotions unseen during the generation of the representation. EPs trained only on angry, happy, neutral, and sad data can classify a test set composed of angry, happy, neutral, sad, and frustrated utterances with a classification accuracy of 58.20%. This represents a decrease of performance of only 0.35% when compared to the results obtained by including frustration in the EP-training. This study's novelty is in its demonstration that EPs, a new representation for emotional utterances, can be used to discriminatively characterize emotions unseen during the training of the EPs.

6.1 Description of Data

6.1.1 IEMOCAP Database

The representative capability of the EP representation was evaluated using the USC IEMOCAP Dataset collected at the University of Southern California [14]. This dataset contains data from five mixed-gender pairs of actors (10 actors total). The data include video, audio, and motion-capture recordings. A full description of the data can be found in Chapter 4, Section 4.1.1.

Data Type	Angry	Happy	Neutral	Sad	Frustrated
Prototypical	284	709	121	309	353
	15.99%	39.92%	6.81%	17.40%	19.88%
Nonprototypical	316	496	451	315	598
	14.52%	22.79%	20.73%	14.48%	27.48%
Combined	600	1205	572	624	951
	15.18%	30.49%	14.47%	15.79%	24.06%

Table 6.1: The distribution of the emotion classes in the prototypical and nonprototypical categories.

6.1.2 Data Definitions

As discussed in the previous chapter, the data were partitioned into groups defined by the level of agreement between evaluators. These groups were labeled prototypical and nonprototypical majority-vote (hereafter referred to as nonprototypical). These definitions are derived from those of Russell [101]. *Prototypical* utterances have clear emotional content with total evaluator agreement; the utterance's majority emotional tag was selected by all of the evaluators. The *nonprototypical* utterances have emotional content that is less clear than that of the prototypical utterances, the majority emotion tag was the tag selected by only a majority of the evaluators. The distribution of the data can be seen in Table 6.1.

6.2 Emotion Profiles

One theory of emotion asserts that there exist "basic emotions". An emotion is basic if it is differentiable from all other emotions [40]. The set of basic emotions can be thought of as a subset of the space of human emotion, forming an approximate basis for the emotional space. More complex, or secondary, emotions can be created by blending combinations of the basic emotions. For example, the secondary emotion of jealousy can be thought of

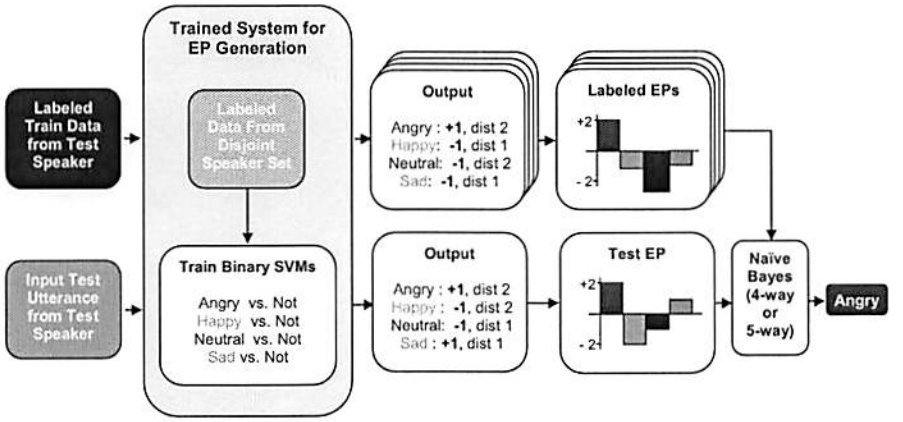


Figure 6.1: The EP-based classification system diagram. This example demonstrates the correct classification of a nonprototypical angry utterance (a mixture of anger and sadness).

as the combination of the basic emotions of anger and sadness [124]. There are often four emotions postulated as basic. This emotion list includes anger, happiness, sadness, and fear. The basic emotions utilized in this work are a subset of this basic emotion list and include: anger, happiness, sadness, and an additional emotion, neutrality, usually defined as the absence of discernible emotional content.

Thus, EPs represent emotional utterances using a set of emotional bases. The EPs quantify the presence or absence of a set of emotions in a given utterance. This subset of emotional labels is chosen to minimize class overlap and correlation. This work assesses the utility of extending the EP representation to include additional emotions that are correlated with the emotional bases previously described.

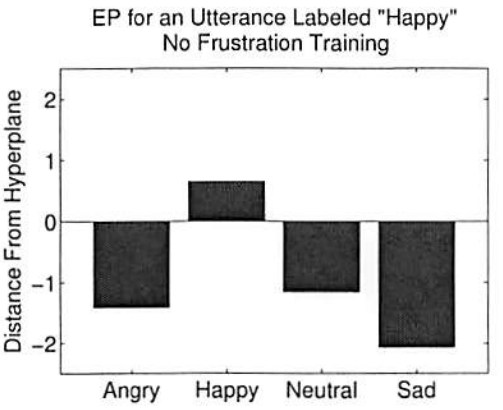


Figure 6.2: The EP of an utterance tagged as 'happy'. This EP has been trained without frustration data.

6.2.1 Construction of an EP

As discussed in the previous chapter, the EPs are constructed using Support Vector Machines (SVM) and are speaker-independent. The models are trained using a disjoint speaker set (e.g., the EP for Speaker 1 is generated using data from Speakers 2-10). This training data is clustered into the semantic classes using the labels angry, happy, neutral, sad, and when applicable, frustrated. The EPs are constructed by testing the held out speaker data (e.g., Speaker 1) on the trained SVM models (Figure 6.1). Each EP contains n-components, one for the output of each emotion-specific SVM. The number of components is either four (angry, happy, neutral, and sad) or five (angry, happy, neutral, sad, and frustrated). See Figure 6.2 for an example of a four-dimensional EP.

6.2.2 Classification with EP-Based Representations

There are two ways to transform an n-dimensional EP into a final classification label. The simpler of the two approaches is to assign a label to an input utterance based on

the maximal component of the profile (e.g., in Figure 6.2 the label would be happy). This approach was employed in Chapter 5. However, as observed in the previous chapter (Section 5.4.4), the minority component also contains emotional relevant information. Voting-based labeling does not take advantage of the information in the minority components. Instead of relying on choosing the maximal confidence, the final emotion can be selected after classifying the generated profile in a speaker dependent method. In this work, we use Naïve Bayes classification.

6.2.3 Speaker-Dependent and Speaker-Independent Components

he classification framework employed in this study is motivated by speaker personalization. Speaker personalization involves two stages, a speaker-dependent and a speaker-independent stage. In speaker personalization, a system is initialized with a baseline set of models. The personalization stage is then the process of adapting the system's models to the current speaker. Speaker personalization is important in emotion-aware technology as emotion production varies across individuals.

In this framework the classification system is composed of the described speaker-independent and speaker-dependent components. In the speaker-independent stage, emotion-specific SVMs are trained using the labeled emotional (angry, happy, neutral, sad, and frustrated, if applicable) data from nine speakers. These four or five emotion-specific SVMs are used to generate the four or five-dimensional EPs for the held out speaker. These EPs are used as the features in the speaker-dependent classification stage. In the speaker-dependent classification stage, the held out speaker's EPs are classified in a

speaker-dependent fashion using Naïve Bayes (Figure 6.1). The results are assessed using leave-one-out cross-validation over the generated EPs for each speaker. For example, Speaker 1 has m EPs after the speaker-independent EP construction. The final emotion class assignment of an utterance (represented by an EP) is determined by training a Naïve Bayes classifier on the remaining $m - 1$ EPs. This process is repeated over all of the generated EPs. Preliminary results suggest that Naïve Bayes classification is more effective in this task than K-Nearest Neighbors, Discriminant Analysis, and Gaussian Mixture Models.

6.3 Feature Extraction and Selection

The features utilized in this study are extracted from the audio and motion-capture information. In both cases utterance-level features are used. The statistics used in this study include: mean, variance, upper quantile, lower quantile, and quantile range.

The audio features include the first thirteen Mel Filterbank Coefficients (MFB), pitch, and intensity. Pitch and intensity are commonly used in emotion classification tasks and have been found to be effective [104, 106, 107, 119]. As discussed in the previous chapter, Mel Filterbank Cepstral Coefficients (MFCC) are also commonly used in both speech and emotion recognition. MFCCs are not used because previous work has demonstrated that MFBs are more effective for emotion classification than MFCCs [16].

As stated in the last chapter, the video features are based on Facial Animation Parameters (FAP) [116]. These features are adapted for the motion capture configuration

present in the USC IEMOCAP dataset. FAPs specify the (x,y,z) distances between specific points on the face. The video features were broken down into regions defined by the cheeks, eyebrow, forehead, and mouth. A more detailed description of the video features can be found in Chapter 5, Section 5.2.2.

6.3.1 Feature Selection

The initial feature set consists of 685 features. The feature selection method utilized is Principle Feature Analysis (PFA) [71]. PFA is an extension of Principle Component Analysis (PCA) that returns interpretable features (from the original feature space) rather than linear combinations of features. In PFA, as in PCA, the eigenvalues and eigenvectors are calculated. The features are clustered in the PCA space. The features closest to the mean of each of the clusters are returned as the final feature set. The PFA feature selection was speaker-independent (e.g., features were selected for Speaker 1 using Speakers 2-10) over the prototypical and nonprototypical utterances labeled as angry, happy, neutral, or sad. The final feature set contained 30-features for each speaker. This feature selection algorithm has been used in emotion classification tasks on the USC IEMOCAP dataset [79, 80].

6.4 Methods

There are two train-test scenarios presented to analyze the ability of the EP tool to generalize to unseen data. In both scenarios, the EP performance when the training and test contain the same emotions is used as a benchmark. In the first scenario, the

EPs are augmented to include a frustration component, in the second scenario the EPs contain only the angry, happy, neutral, and sad data. In both conditions, the EPs are tested on the angry, happy, neutral, sad, *and* frustrated data. The goal is to assess the ability of the EP to uniquely represent unseen test data. The hypothesis is that frustration test utterances will be represented in the EPs sufficiently differently from that of the other affective classes. This result is anticipated because frustration has a high degree of overlap with the classes of anger, happiness, and sadness. Consequently, EPs trained on the set of angry, happy, neutral, and sad emotions should be able to represent frustration. This result would suggest that EPs used for n-way classification need not contain n components.

6.5 Results

This section will demonstrate the efficacy of EP-based representation for the emotional classes of angry, happy, neutral, sad, and frustrated. The classification performance will be analyzed across three data conditions: prototypical only, combined prototypical and nonprototypical, and nonprototypical. The classification results on a baseline set of angry, happy, neutral, and sad data are provided as a reference. In the previous chapter and previously published work [80] the classification was entirely speaker-independent. Consequently, the results presented in this study cannot be compared directly to any of the published work due to the final user-dependent classification step. However, in [80] the authors obtained a speaker-independent unweighted accuracy of 62.42% (accuracy across the four emotion categories) on combined prototypical and nonprototypical data

across the classes of angry, happy, neutral, and sad using a fused GMM-HMM approach. The authors used a profile-based technique to fuse the facial (motion-capture) and vocal modalities. While, the current unweighted accuracy of 66.52% is not directly comparable, however, it demonstrates that the EP-based classification technique is effective for this database.

6.5.1 Classification with EP Frustration Training

This set of results demonstrates the classification performance when a five-dimensional EP representation is employed. The hypothesis is that training EPs with frustration will not provide significant benefit to the overall five-class classification accuracy when compared with the five-class classification of the data without first training the EPs on the frustration data.

In this scenario, both the EPs and Naïve Bayes classifier are trained with data from the set of angry, happy, neutral, sad, and frustrated utterances. The results demonstrate that over both the prototypical and combined datasets the classification performance for each of the emotions decreases when the train and test sets are augmented with frustration (Table 6.2, compare the left-most and middle result columns). These results are anticipated due to the high degree of overlap with the angry, sad, and neutral emotional classes. In [14] the authors demonstrate that within the human evaluations frustration overlaps with the classes of anger, sadness, and neutrality. In the human evaluations, utterances labeled as frustration were also labeled as anger, happiness, neutrality, and sadness 11%, 0%, 7% and 4% of the time, respectively. Utterances labeled as anger, happiness, neutrality, and sadness were also labeled as frustration 17%, 1%, 13%, and 8%

	Prototypical	Four class EP	Frustration Augmentation	
			EP Train	No EP Train
F- measure	Angry	0.82	0.69	0.71
	Happy	0.90	0.86	0.85
	Neutral	0.59	0.51	0.53
	Sad	0.82	0.80	0.78
	Frustrated	-	0.58	0.56
Weighted Accuracy (%)		83.69	74.32	73.54
Unweighted Accuracy (%)		79.29	69.09	69.01
	Combined	Four class EP	Frustration Augmentation	
			EP Train	No EP Train
F- measure	Angry	0.73	0.54	0.56
	Happy	0.78	0.75	0.75
	Neutral	0.45	0.27	0.30
	Sad	0.67	0.61	0.61
	Frustrated	-	0.50	0.46
Weighted Accuracy (%)		68.43	58.55	58.20
Unweighted Accuracy (%)		66.52	54.19	54.30
	Nonprototypical	Four class EP	Frustration Augmentation	
			EP Train	No EP Train
F- measure	Angry	0.66	0.37	0.40
	Happy	0.61	0.58	0.57
	Neutral	0.47	0.29	0.33
	Sad	0.54	0.49	0.48
	Frustrated	-	0.45	0.42
Weighted Accuracy (%)		56.53	44.72	44.44
Unweighted Accuracy (%)		57.89	44.42	43.83

Table 6.2: Classification results (F-measure) across the three datasets: prototypical, combined, and nonprototypical. "EP Train" indicates five-dimensional EPs, "No EP Train" indicates four-dimensional EPs.

of the time, respectively. Consequently, one would expect the classification performance of those three classes to decrease when frustration is added to the train and test sets. In the nonprototypical dataset there was also a decrease in performance in the happy classification. This may be due to the increasingly vague definition of the emotion of happiness.

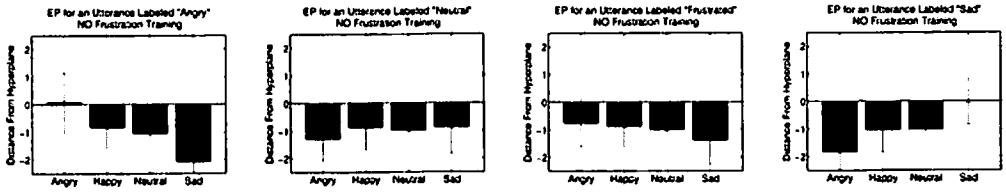
6.5.2 Classification without EP Frustration Training

In the final scenario the EPs are trained only with angry, happy, neutral, and sad data, while the Naïve Bayes classifier must classify emotions from all five categories. In this training scenario, the EPs must distinctly represent an emotion not seen during training. The results will be compared to the previous training scenario in which frustration was used to train the EPs. The hypothesis is that since frustration overlaps with the other classes already represented in the profile, the profile does not need a frustration component, as that information is redundant.

The results demonstrate that there is no significant difference between including frustration in the training of the profiles and merely training on the profiles resulting from only the angry, happy, neutral, and sad training. The greatest performance disparity occurred in the prototypical dataset where the weighted accuracy decreased by only 0.78%, the unweighted by 0.08%. In the combined and nonprototypical datasets, the weighted accuracy decreased by 0.35% and 0.28%, respectively (Table 6.2). These performance differences are not significant at $\alpha = 0.05$. The small discrepancies in performance suggest that the EPs are a robust representation for emotion.

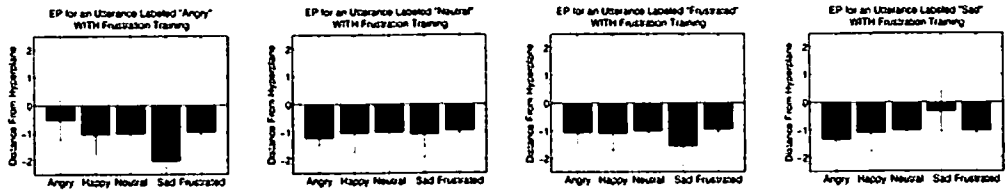
6.5.3 EP Representation of Frustration

Previous work has demonstrated that the utterances labeled as frustrated in this database are confused both by human evaluators [14] and by machine learning algorithms [88] (audio-only analysis). The graphs of Figures 6.3 and 6.4 further support the inherent difficulty in characterizing this ambiguous emotion. Figure 6.4 demonstrates that on average the emotion of "frustration" is represented as not present for emotions labeled as



(a) Angry. (b) Neutral. (c) Frustrated. (d) Sad.

Figure 6.3: The average EPs for the prototypical and nonprototypical utterances when the EPs were trained *without* frustration data. The error bars represent the standard deviation. The happy EP is not included in this plot; the trends follow those of the angry and sad EPs.



(a) Angry. (b) Neutral. (c) Frustrated. (d) Sad.

Figure 6.4: The average EPs for the prototypical and nonprototypical utterances when the EPs were trained *with* frustration data. The error bars represent the standard deviation. The sad EP is not included in this plot; the trends follow those of the angry and happy EPs.

frustration. However, in both training conditions, frustration is recognized well above the chance level, which is 19.88% for prototypical data and 27.48% for nonprototypical data (Table 6.2). This indicates that the feature variations characteristic of frustration are captured by both methods. This supports the assignment of frustration to a secondary, rather than a basic emotion since it can be similarly described using a combination of basic emotions. This further supports the idea that an emotional utterance should be characterized by what is present, but also by what is confidently identified as absent. It should be noted that frustration, even when not modeled during the construction of

EP Type	Angry	Happy	Neutral	Sad
4-Dim	ANS	AHNS	AS	AHNS
5-Dim	ANSF	AHNSF	ANS	ANSF

Table 6.3: ANOVA analysis of the component-by-component comparison between the frustrated and other emotional EPs. The emotion components are labeled by the first letter of their class (e.g., angry EP component = 'A'). All dimensions listed in this table are statistically different with $p < 0.001$.

the EP, can be more accurately characterized than neutral utterances, which have been historically difficult to characterize in this database [80,87,88].

The average EPs of Figures 6.3 and 6.4 suggest that there is not a large difference between the characterization of neutral and frustrated data. Such a finding would imply that frustration, like neutrality, is not so much captured as defaulted to a generic "none of the above" representation. However, statistical analyses support the differentiation of these two emotion classes in line with the semantic understanding of these emotion classes. In the four-dimensional EPs the frustration EPs are differentiated from the neutrality EPs along the anger and sadness dimensions with $p < 0.001$ (ANOVA, Table 6.3), where anger is more strongly represented and sadness is less strongly represented in the frustration EP than in neutrality EP. This suggests that frustrated utterances can be differentiated from neutral utterances based on the presence of angry components ($p < 0.001$, ANOVA, Table 6.3), although these components are less strongly defined when compared to the angry utterances ($p < 0.001$, one-way t-test, difference of means). It is also interesting to note that the comparison of the sad components in the frustration and anger EPs suggests that sadness is represented more strongly in frustrated utterances than in angry utterances ($p < 0.001$, one-way t-test, difference of means).

6.6 Conclusions

This chapter demonstrates the efficacy of EP-based classification for out-of-domain audio-visual emotional data. In all three data types there was no significant difference between the classification accuracies (weighted or unweighted) of the EPs trained on frustrated data and trained only on angry, happy, neutral, and sad data. The decrease in the emotion-specific F-measures between the EPs trained and not trained on frustration was less than or equal to 0.04 in all cases and in some cases increased (prototypical anger and neutrality, combined neutrality, nonprototypical anger and neutrality). It should be noted that all emotions are recognized above the chance level. This indicates that EPs whose components span the target emotional space are sufficiently flexible to represent unseen emotions and offer robust representations for emotional communication.

The representative power of an EP is dependent on the employed emotional basis. The EPs in this study were able to represent frustration because frustration can be described as combination of the emotion classes included in the EPs. The ability of the EPs to distinctly represent emotions that do not overlap with the EP components has not yet been assessed. Future work will include the investigation of techniques to derive additional component representations for EPs. Future work will also include analyses of the ability of the EPs to represent additional more highly ambiguous emotion classes.

The F-measures for the classes of neutral and frustration were comparatively low. This may be a result of ambiguous class definitions, the ambiguous expression of neutral and frustrated speech prevalent in human interactions, or perhaps a suboptimal feature set. Future work includes the investigation of techniques to improve these accuracies.

However, the success of such future work is not guaranteed. The lower performance of frustration classification can be explained in part by the high-degree of overlap in human evaluations between the classes of frustration and anger, neutrality, and sadness. Such a large degree of overlap suggests that there is a lower upper-bound for frustration classification.

This work demonstrates a method for quantifying out-of-domain emotional data. Such representations are necessary as human-machine interactive technology continues to develop and speaker personalization becomes increasingly important. As human-interactive technologies become more prevalent, interfaces must be able to interpret truly ambiguous information, utterances without human-labeled ground truths. Future work includes extending this representation to the domain of these truly ambiguous emotional utterances.

6.7 Work Published

The work presented in this chapter was published in the following articles:

1. **Emily Mower**, Maja J Matarić and Shrikanth S. Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotional Profiles." *IEEE Transactions on Audio, Speech and Language Processing.* Accepted for publication, August 2010.
2. **Emily Mower**, Maja J Matarić and Shrikanth S. Narayanan, "Robust Representations for Out-of-Domain Emotions Using Emotion Profiles." In Proceedings of *IEEE Workshop on Spoken Language Technology (SLT)*, Berkeley, CA, December 2010.
3. **Emily Mower**, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, Shrikanth Narayanan. "Interpreting Ambiguous Emotional Expressions." In Proceedings of *ACII Special Session: Recognition of Non-Prototypical Emotion from Speech- The Final Frontier?*. Amsterdam, The Netherlands, September 2009.

Chapter 7

Cluster Profiling

The previous two chapters introduced and described Emotion Profiles (EP), a method for representing the affective content of human utterances. This representation is important for interactive affective technologies, which detailed models of human emotion for accurate user state determination. These models are commonly trained using supervised learning algorithms. However, such algorithms typically require labeled training corpora, the collection of which is often expensive and time-intensive. This chapter presents a system-level heuristic semi-supervised approach to user-specific emotion-classification using a novel Cluster-Profile (CP) representation of emotion.

In user-adapted emotion classification systems, two types of data are necessary: a large amount of emotional data from multiple speakers and a smaller amount of data from the target speaker. The labels from the target speaker are directly relevant to the classification task while those from the disjoint speakers are needed only for training. An approach requiring only the labels of the target speaker's utterances would drastically reduce the time needed for database preparation.

In the previous chapters the efficacy of an Emotion-Profile (EP) based representation for classification was demonstrated. EPs are a quantitative representation of the affective content of an utterance in terms of the presence or absence of a set of component emotions. The profile were the semantic, or categorical, labels: angry, happy, neutral, and sad. However, it is not clear that the profiles must be constructed using these types of semantic components.

In this chapter we investigate a system-level heuristic semi-supervised approach for emotion classification. The classification system is broken down into four steps: speaker-independent feature selection, speaker-independent clustering, speaker-independent profile generation, and speaker-dependent classification. The feature selection method is the method utilized in the previous chapter, unsupervised Principal Feature Analysis (PFA), an extension of Principal Component Analysis, also used in [79, 80]. The data are clustered using unsupervised agglomerative hierarchical clustering of the emotional space. These clusters are used to train cluster-specific Support Vector Machines (SVM) whose output are the components of the CPs. Finally, the emotion content of the utterance is assessed by classifying over the generated CPs. The system is a heuristic semi-supervised approach because the feature selection, clustering, and profile generation are unsupervised while the final classification step is supervised. The unsupervised portion establishes a data-dependent representation for the affective test data using the majority of the training data. The final supervised classification utilizes the generated CPs for Naïve Bayes classification.

The CP classification method outperforms the EP classification by 0.88% absolute (69.25% vs. 68.37%). This result demonstrates the efficacy of the CP-based classification

system. The CPs represent emotional utterances in n-components, where n is the number of clusters. This comparable performance of the CP and EP representations suggests that given training sets with expressions from a non-disjoint set of emotion classes, it may be necessary to label only a subset of the large training data. These results cannot be compared directly to any published work due to the final speaker-dependent classification step. However, this performs comparably to fused GMM-HMM method presented in [80] (62.42%). The novelty of the current work lies in its new definition of a profile and an assessment of the necessity of the semantic profile dimensions utilized in the EPs.

7.1 Description of Data

7.1.1 IEMOCAP Database

The discriminative power of the CP-representation is evaluated using the USC IEMOCAP database, collected at the University of Southern California (USC) [14] and described in Chapter 4, Section 4.1.1.

7.2 Emotion and Cluster Profiles

Profile-based representations describe affective utterances over a set of affective components rather than in terms of a single mathematical (e.g., a valence of '3') or semantic label (e.g., 'angry'). This added flexibility is beneficial when the emotional character of the speech is subtle. In the previous two chapters the EPs were implemented as either four or five dimensional representations of emotion. The dimensions expressed the degree of presence or absence of each of the emotions: angry, happy, neutral, sad, and

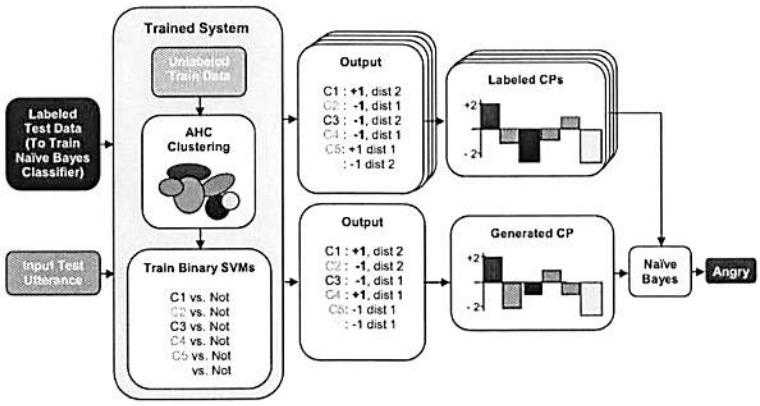


Figure 7.1: The CP-based classification system diagram. This example demonstrates the correct classification of a nonprototypical angry utterance (a mixture of anger and sadness).

optionally frustrated. This subset was chosen to minimize affective overlap in our experimental dataset. In this chapter, we explore profile generation using an unsupervised component-generation approach (Figure 7.1).

7.2.1 Description of the Train and Test Sets

The dataset considered consists of 4,806 utterances across the ten emotional labels and ten-speakers. The profile generation ("training") is speaker-independent while the final classification ("testing") is speaker-dependent (Figure 7.1). For each speaker, the training data (for unsupervised clustering and profile generation) consist of all of the utterances not spoken by the speaker. These data contain unlabeled emotions from all 10 emotion categories. The testing data consist only of utterances spoken by the speaker from the set: angry, happy, neutral, and sad.

7.2.2 Unsupervised Clustering for CPs

The feature space is clustered using the unsupervised agglomerative hierarchical clustering (AHC) over the unlabeled training data. This hierarchical clustering strategy circumvents the initialization issues common to other clustering approaches (e.g., k-means or GMM-EM [38,121]). AHC is a bottom-up process, which is more computationally efficient than top-down (divisive) clustering. Research has demonstrated that AHC can be applied to many clustering tasks and is effective. This clustering approach is of particular popularity in the field of speaker clustering and diarization [114].

Initially, AHC considers each data point a cluster. Then, at every iteration, it selects the closest pair of clusters to merge. This merging procedure continues until a pre-set stopping criterion is satisfied. Generalized likelihood ratio (GLR) [48] is used to measure inter-cluster distance at every stage of AHC. The stopping criterion is a manually pre-set number of clusters, n. This work will explore the utility of considering different numbers of clusters in the CP construction.

7.2.3 Construction of a Profile

EPs and CPs are both constructed using the output from Support Vector Machines (SVM). As described in the Chapter 5, Section 5.3, SVM is a maximum margin classifier that projects input data into a higher dimensional space to find an optimal separating hyperplane between two classes. The distance from one point in the projected space to the hyperplane can be interpreted as the confidence of the classifier's assessment. Points closer to the hyperplane are representative of data that are more easily confused in the

projected-space. These points represent utterances that cannot be as confidently labeled as utterances further from the decision hyperplane.

In the CP approach, n speaker-independent binary self vs. other SVMs are trained for each of the clusters generated using AHC. Each cluster-specific SVM returns a membership value ($\pm$ 1) and a distance from the hyperplane. The profiles are created by weighting the membership by the raw distance from the hyperplane. A sigmoid function is often used to convert the range of SVM hyperplane distances to the range 0–1. However, the raw distances were retained because pilot studies demonstrated the efficacy of utilizing the raw, rather than the sigmoid-transformed, distances in the profile-based representations (see Chapter 5, Section 5.3 and Figure 5.4). The final profile is an n-dimensional representation of the n-classifier confidences.

The performance of the cluster profile representation will be compared to that of the emotion-profile representation (Chapter 5). The final step is performing classification over the generated profiles (both CP and EP). This n-dimensional classification is performed using Naïve Bayes. Gaussian Mixture Models, KNN, and Discriminant Analysis were also explored, but were not as effective. Only Naïve Bayes results will be reported.

7.3 Hypotheses

Hypothesis 1: The data-driven CP representation will be more accurately classified than the EP representation. The CP can represent more emotion-specific fluctuations than the EP because it can contain a larger number of components. This will allow the CP to capture more of the inherent variation in the affective data. A negative result would

suggest that the semantic emotional labels are more effective at clustering the affective space than the employed data-driven technique.

Hypothesis 2: The EP representation is a more compact representation of the affective components of human speech. Semantic emotion labels describe clusters of the data that are objectively recognized by large numbers of people. Consequently, it is expected that the clusters generated using these affective labels will be highly representative of the feature-level properties of the emotional utterances.

7.4 Features Extraction and Selection

The EPs and CPs are constructed using utterance-level features extracted from the audio and motion-capture modalities. The statistics used in this study include: mean, variance, upper quantile, lower quantile, and quantile range. All features were normalized using speaker-dependent z-normalization. Utterances were rejected if any of the audio or motion-capture features were undefined. The features utilized in the CP analysis are the same as those utilized in the previous chapter.

The set of audio features included: intensity, pitch, and the first 13 Mel Filterbank Coefficients (MFB). Intensity and pitch have been used successfully in emotion classification studies [104, 106, 107, 119]. In this work, MFBs are also used. MFBs are less common than Mel-Frequency Cepstral Coefficients (MFCC) in emotion research. However, previous work has demonstrated that MFB features are more effective for emotion classification than MFCCs across all broad phoneme classes [16].

The motion-capture features utilized in this work are derived from Facial Animation Parameters (FAP) [116]. These features are part of the MPEG-4 standard and represent distances between points on the face. The FAPs were adapted to the motion-capture configuration used in the USC IEMOCAP data recording. The facial features were broken into groups by facial region. These regions included: mouth, cheeks, forehead, and eyebrows. These features were also used in Chapters 5 and 6.

7.4.1 Feature Selection

The initial feature set has 685 features. The feature set size is reduced using the unsupervised method of Principal Feature Analysis (PFA) [71] (Chapter 6, Section 6.3.1). The feature sets were identified in a speaker-independent fashion. For example, the selected features for Speaker 1 were analyzed using the data from Speakers 2-10. The final feature set size was 20-features.

7.5 Experimental Methods

The goal of this analysis is to determine if an unsupervised data clustering algorithm can find relevant clusters within the data for use in the profile-based classification. A successful result would indicate that exhaustive labeling of the training space is not necessary. Instead, the data-dependent clusters inherent in the space can be used as components of the profile for a final supervised training on a much smaller proportion of the data.

The EP-based classification is presented as a baseline performance metric. The EPs are trained in a speaker-independent fashion (e.g., EPs for Speaker 1 are trained using

the data from Speakers 2-10) over the semantic labels of angry, happy, neutral, and sad. In CP-based classification the speaker-independent training data are first clustered into n-clusters using the aforementioned clustering approach. The CPs are then constructed using the output from the n-SVMs trained on each cluster's data (one SVM for each of the n clusters). In both profile-based methods, the final emotion assessment is made using Naïve Bayes over the generated profiles. The performance of the Naïve Bayes classifier is assessed using leave-one-out cross-validation (see system diagram, Figure 7.1).

7.6 Results

7.6.1 EP Classification

The EP-based classification will serve as a comparative baseline for the CP-based classification results. In the EP-based classification, the accuracy was 68.37%. The emotion-specific results can be seen in Table 7.1. The classes of anger, happiness, and sadness were well recognized (F-measure > 0.69). The class of neutrality was relatively poorly recognized. This trend is common in this database, where neutrality remains an emotion class that is not well understood [80, 87].

7.6.2 CP Classification

In this task, the maximal accuracy occurred with 15 clusters. The maximal accuracy was 69.25% (Table 7.1). The emotions of anger, happiness, and sadness were again well recognized (F-measure > 0.69). It should be noted that in the CP-representation, the F-measure for the class of neutrality increased to 0.54. This represents a 9% absolute

	Emotion	EP	Number of Clusters								
			3	5	7	9	11	13	15	17	19
F-Meas.	Angry	**0.73**	0.51	0.60	0.64	0.64	0.68	0.68	0.69	0.68	0.69
	Happy	**0.77**	0.74	0.76	0.76	0.76	**0.77**	**0.77**	**0.77**	**0.77**	**0.77**
	Neutral	0.45	0.34	0.40	0.49	0.49	0.53	**0.54**	**0.54**	0.53	0.53
	Sad	0.69	0.52	0.60	0.67	0.67	**0.70**	**0.70**	**0.70**	**0.70**	**0.70**
Acc.	Weighted	0.68	0.57	0.62	0.66	0.66	**0.69**	**0.69**	**0.69**	**0.69**	**0.69**

Table 7.1: The CP-based classification results. The entries in bold font indicate the best accuracy or F-measure recorded.

and 20.00% relative improvement. This result suggests that CP-based representations are more effective for capturing inherently ambiguous classes of emotion than EP-based representations.

It should be further noted that the CP-based classification outperformed the EP-based classification by 0.88% absolute (1.29% relative). This result is not statistically significant at $\alpha = 0.05$, indicating that the CP and EP representations are both effective for emotion classification. This equivalence suggests that it is not necessary to exhaustively label a large dataset for user-adapted emotion classification tasks.

7.7 Discussion

The studies in this paper were motivated by two hypotheses: 1) the CP-based technique would outperform the EP-based technique and 2) the EP representation would offer a more compact representation of affective content. The results demonstrate that the CP-based classification outperforms EP-based classification by 0.88% absolute (1.29% relative) with 15-clusters. This suggests that the CP-based representation can adequately

represent emotion in an unsupervised manner. However, the assertion that the CP representation can more accurately represent emotion content of utterances cannot be supported at this time.

The second hypothesis is also supported. CP-based classification required at least 11-clusters to match the accuracy obtained by EP-based classification. The F-measures obtained in the EP-based classification for angry, happy, neutral, and sad was never obtained in the CP-based classification for anger and required 11, 13, and 11 clusters respectively for the classes of happiness, neutrality, and sadness. This suggests that the EP-based representation is more compact than this implementation of the CP-based representation. This further suggests that the clusters generated from the semantic labels of angry, happy, neutral, and sad are extremely effective for capturing the affective feature properties of the utterances, supporting the use of the components of anger, happiness, neutrality, and sadness in the EP-based representation.

7.8 Conclusions

This chapter presents a novel system-level heuristic semi-supervised technique to classify the emotion content of utterances using a profile-based technique. The CP-based classification non-significantly outperformed the EP-based classification by 0.88% with 15 clusters. This suggests that both data-driven and knowledge-driven clusters are effective for profile generation. The CP-based representation alleviates the need for exhaustive labeling of the training corpus, requiring instead a labeling of a small subset of the data.

Although, as stated earlier, the results presented in this paper cannot be directly compared to previously published methods, both the EP- and CP-based classification systems produce similar accuracies to those seen in the literature (62.42%) [80]. This demonstrates that both profile-based representations are effective for emotion classification tasks.

The results are presented on the USC IEMOCAP database. Future research will investigate the relative robustness of the EP or CP methods across multiple databases. The lower complexity of the EP representation suggests that the emotional clusters (angry, happy, neutral, and sad) may be a more orthogonal "basis" representation in the IEMOCAP database. This may indicate that the EP components are a more perceptually salient representation than the CP components. However, the CP representation in this database provides better functional definitions for the components. Future work will investigate the relevance of the EP and CP representations with respect to human perception. Future work will also include the analysis of additional clustering methods to determine the effect of these techniques on the classification accuracy of the system. Finally, we plan to investigate user-personalization methods for the final emotion assessment such as using Collaborative Filtering.

This chapter presents a foray into heuristic semi-supervised learning for emotion classification. Semi-supervised emotion classification has the potential to make user-personalization more tractable by incorporating unlabeled emotional data for deriving an affective representation. As affective interactive technologies continue to grow in popularity, these techniques will only become more important.

7.9 Work Published

The work presented in this chapter was published in the following articles:

1. **Emily Mower**, Maja J Matarić and Shrikanth S. Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotional Profiles." *IEEE Transactions on Audio, Speech and Language Processing.* Accepted for publication, August 2010.
2. **Emily Mower**, Kyu Jeong Han, Sungbok Lee and Shrikanth S. Narayanan. "A Cluster-Profile Representation of Emotion Using Agglomerative Hierarchical Clustering." In Proceedings of *International Speech Communication Association (InterSpeech)*, Makuhari, Japan. September 2010.
3. **Emily Mower**, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, Shrikanth Narayanan. "Interpreting Ambiguous Emotional Expressions." In Proceedings of *ACII Special Session: Recognition of Non-Prototypical Emotion from Speech- The Final Frontier?.* Amsterdam, The Netherlands, September 2009.

Chapter 8

Conclusions and future work

This thesis presented a new method for quantifying and identifying the emotion present in naturalistic human utterances. This thesis also presented classification and statistical studies demonstrating the ways in which individuals process and interpret emotional cues and the trends in user evaluation styles. Finally, this thesis presented methods for interpreting highly ambiguous emotional utterances. These utterances are often not considered in emotion classification tasks as they either are too noisy or cannot be assigned a ground truth. This thesis presented results that indicate that even given highly ambiguous utterances, due either to natural human speech or intentional emotional mismatch, there exists degrees of consistency in the reported user perception; clear and interpretable expression within the feature streams; and meaningful EP-based quantification results.

This thesis explored the link between reported perception and feature modulation using a database composed of congruent and conflicting emotional cues. A statistical analysis of the reported perception indicated that in emotionally congruent utterances, utterances in which the emotion expressed in the audio and video channels match, evaluators integrate the information expressed within the channels to arrive at an emotional

description dependent on both the audio and video expressions. This highlights the importance of the proper design of both the audio and video components of emotional expression even given a simplified (synthetic) video channel. The results also suggested that given emotionally conflicting utterances, utterances in which the emotions expressed in the audio and video channels do not match, evaluators tended to rely more heavily on the more expressive channel than the less expressive channel. However, this result varies across the emotion dimensions. Our results suggest that evaluators tended to rely on audio (the more expressive modality and the modality correlated with activation) for activation information and both the audio and video channels for valence information. This finding further stresses the importance of proper video design, even when the video is much less expressive than the audio, for recognizable emotion expression.

The thesis also presented methods to evaluate evaluators based on classification metrics. Our results indicate that it is often more accurate to model an averaged evaluator than an individual evaluator. These studies also demonstrated that there is no significant decrease in accuracy when an individual evaluator is classified utilizing models trained on an averaged evaluator's data. Furthermore, the pervasive difficulty in classification suggests that both categorical and dimensional descriptions of emotion do not fully describe the emotional landscape. These findings highlighted the importance of developing an understanding of the relationship between feature modulation, evaluator state, and the resulting reported emotion perception.

The results from the perceptual and evaluator-modeling experiments suggested that traditional characterizations of emotion (i.e., dimensional or categorical representations) are not sufficient to describe the emotion content of affective utterances. This insight led

to our development of Emotion Profiles (EP), a new method for quantifying emotions. EPs are a multidimensional representation of emotion that incorporate aspects of the categorical and dimensional representations to arrive at a rich and interpretable method for expressing the affective content of utterances.

In human emotion expression, naturally occurring utterances are often complex expressions of emotion. These complex emotional utterances can appear as combinations or blendings of several emotions, combinations which cannot be well captured by a single label. Furthermore, these complex utterances may not be well described by dimensional labels because separate subtle emotion classes may overlap in the dimensional space, leading to interpretability problems.

EP-based quantification techniques are an integration of the categorical and dimensional descriptions of emotion. These techniques are ideally suited to describe the affective content of natural utterances through their characterization of an emotion in terms of the degree of presence or absence of a set of emotions, leading to a richer emotional description. The results presented in this thesis demonstrate that EPs are an effective measure for quantifying reported human emotion perception. EPs can be used as an intermediary step during classification or as a method to characterize highly ambiguous emotional utterances. The EP method is able to not only accurately classify emotions with affective ground truths, but is also able to interpret the affective content of emotionally ambiguous utterances, allowing for their inclusion in natural human-machine interactions.

This thesis also presented detailed analyses of the EP representations through an investigation of the robustness of the representation and through a study analyzing the efficacy of using data-driven, rather than semantic emotional, components of the EP. The

results suggest that EPs can robustly represent unseen secondary emotions that can be described as a combination of the profile components. We demonstrate that frustration, described as a combination of sadness and anger, can be represented sufficiently distinctly using four dimensional EPs (angry, happy, neutral, sad) or five dimensional EPs (angry, happy, neutral, sad, frustrated). This result suggests that EPs can be sued to represent emotions that are combinations of the profile components.

However, EP components are not required to be semantic emotional labels. Although semantic components allow the profile to represent and quantify emotion in an interpretable way, the generation of these components require a large amount of labeled training data in order to accurately model the confidence of the component assertions (i.e., degree of presence or absence of a given emotion class). We demonstrate that profiles can be created using data-driven components. These components represent emotions in terms of the presence vs. absence, not of specific emotion classes, but of clusters within the feature space. By modeling emotion as a collection of these cluster-level confidences it is possible to characterize emotion without requiring the input training data to have emotional labels. Like EPs, these profiles, called Cluster Profiles (CP), can be used as a mid-level representations in a classification system. The results suggest that EPs and CPs function as similarly effective mid-level representations. This suggests that both semantic emotional clusters and data-driven clusters can be used to characterize the affective makeup of the utterance.

8.1 Research Goals for Future Work

This thesis demonstrated that conventional methods for quantifying emotional content do not fully describe the emotional space. This finding motivated the development of the EP-based frameworks. Future work will explore the efficacy of this quantification with respect to the findings illustrated in the evaluator-specific and congruent-conflicting chapters in this document and additional information not yet considered.

Accurate evaluator modeling is important in human-machine interactions. A machine that cannot adapt to user preferences or perception styles may have difficulty maintaining user interest in the long run. The results presented in the evaluator-specific modeling chapter indicated that the modeling of individual evaluators was less accurate than the modeling of averaged evaluators. The work also suggested that when predicting the perception of a single evaluator it was no less accurate to train models based on an averaged evaluator than on the individual evaluator. This finding suggests that different emotion quantification methods are needed to model the user evaluations. We plan to extend these studies to create user-specific models of reported emotion perception using EP-based techniques. It is our belief that the EP-based representations will more accurately capture individual variations than the previously utilized HMMs.

Accurate user modeling also hinges on a proper understanding of how humans integrate audio and video information. The statistical studies presented in this document indicated that evaluators are biased primarily by audio information when evaluating audio-visual emotional expressions. However, the results of our studies may have been affected by the presence of a synthetic face, making it difficult to determine if the observed

effects were due to the limited expressivity of the utilized face or an innate quality of dynamic audio-visual emotion perception. We plan to extend this congruent-conflicting work by utilizing clips composed of both human facial and human vocal information. These studies will provide an opportunity for the study of the relationship between facial and vocal cues of equal levels of expression. We will also utilize an EP-based framework to assess the ability of the EP to detect the two emotions present in the conflicting utterances.

The context in which an emotion is viewed also has a strong impact on the resulting perception of the user. Currently, none of our studies have employed contextual estimates. We plan to incorporate contextual information via emotion profiles. We will use this information during classification to obtain a richer understanding of the emotional landscape.

Finally, we will demonstrate the effectiveness of our studies in an application context. Previous work has demonstrated that children with an Autism Spectrum Disorders (ASD) diagnosis have a difficult time identifying the emotional content of expressions [22,49,59, 72]. We are developing an interactive computer avatar designed to assist children with ASD in recognizing socially relevant emotion states. We believe that the inclusion of an avatar will allow the children to learn to recognize and, in the future, utilize these emotion states.

References

[1] R. Abelson and V. Sermat, "Multidimensional scaling of facial expressions," *Journal of Experimental Psychology*, vol. 63, no. 6, pp. 546–554, 1962.

[2] A. Arya, L. N. Jefferies, J. T. Enns, and S. DiPaola, "Facial actions as visual cues for personality," *Computer Animation and Virtual Worlds*, vol. 17, no. 3–4, pp. 371–382, 2006.

[3] T. Banziger and K. Scherer, "Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus," *Lecture Notes in Computer Science*, vol. 4738, p. 476, 2007.

[4] L. Barrett, "Are emotions natural kinds?" *Perspectives on Psychological Science*, vol. 1, no. 1, pp. 28–58, 2006.

[5] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: machine learning and application to spontaneous behavior," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 568–573, 2005.

[6] M. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," in *IEEE international conference on systems, man and cybernetics*, vol. 1, 2004, pp. 592–597.

[7] J. Bates, "The role of emotion in believable agents," *Communications of the ACM*, vol. 37, no. 7, pp. 122–125, 1994.

[8] S. Biersack and V. Kempe, "Tracing vocal emotion expression through the speech chain: do listeners perceive what speakers feel," in *ISCA Workshop on Plasticity in Speech Perception*, London, UK, June 2005, pp. 211–214.

[9] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49 - 59, 1994.

[10] S. Buisine, S. Abrilian, R. Niewiadomski, J. Martin, L. Devillers, and C. Pelachaud, "Perception of blended emotions: From video corpus to expressive agent," *Lecture Notes in Computer Science*, vol. 4133, p. 93, 2006.

[11] M. Bulut, C. Busso, S. Yildirim, A. Kazemzadeh, C. Lee, S. Lee, and S. Narayanan, "Investigating the role of phoneme-level modifications in emotional speech resynthesis," in *Interspeech*, Lisbon, Portugal, Sept. 4–8 2005, pp. 801–804.

[12] M. Bulut, S. Lee, and S. Narayanan, "Recognition for synthesis: automatic parameter selection for resynthesis of emotional speech from neutral speech." in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 4629 - 4632.

[13] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, pp. 335–359, Nov. 5 2008.

[14] ——, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, pp. 335–359, Nov. 5 2008.

[15] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the International Conference on Multimodal Interfaces*, State Park, PA, Oct. 2004, pp. 205–211.

[16] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *InterSpeech*, Antwerp, Belgium, Aug. 2007, pp. 2225–2228.

[17] C. Busso and S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, Nov. 2007.

[18] ——, "Joint analysis of the emotional fingerprint in the face and speech: A single subject study," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Chania, Greece, Oct. 2007, pp. 43–47.

[19] ——, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[20] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *InterSpeech*, Antwerp, Belgium, Aug. 2007, pp. 2225–2228.

[21] C. Busso and S. S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Proceedings of InterSpeech*, Brisbane, Australia, Sep. 2008, pp. 257–260.

[22] G. Celani, M. Battacchi, and L. Arcidiacono, "The understanding of the emotional meaning of facial expressions in people with autism," *Journal of Autism and Developmental Disorders*, vol. 29, no. 1, pp. 57–66, 1999.

[23] C. Cottrell and S. Neuberg, "Different emotional reactions to different groups: A sociofunctional threat-based approach to prejudice.," *Journal of Personality and Social Psychology*, vol. 88, no. 5, pp. 770–789, 2005.

[24] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[25] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, pp. 32–80, Jan. 2001.

[26] N. Cristianini and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods.* Cambridge Univ Pr, 2000.

[27] K. Dautenhahn, C. Numaoka, and AAAI, *Socially intelligent agents.* Springer, 2002.

[28] J. Davitz, *The language of emotion.* Academic Pr, 1969.

[29] B. de Gelder, "The perception of emotions by ear and by eye," *Cognition & Emotion*, vol. 14, no. 3, pp. 289–311, 2000.

[30] B. de Gelder, K. Böcker, J. Tuomainen, M. Hensen, and J. Vroomen, "The combined perception of emotion from voice and face: early interaction revealed by human electric brain responses," *Neuroscience Letters*, vol. 260, no. 2, pp. 133–136, 1999.

[31] L. De Silva, T. Miyasato, and R. Nakatsu, "Facial emotion recognition using multi-modal information," in *International Conference on Information, Communications and Signal Processing (ICICS)*, vol. I, Singapore, 1997, pp. 397–401.

[32] B. DeGelder and P. Bertelson, "Multisensory integration, perception, and ecological validity," *Trends in Cognitive Sciences*, vol. 7, no. 10, pp. 460–467, Oct. 2003.

[33] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," *ICSLP*, 1996.

[34] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[35] S. D'Mello, R. Picard, and A. Graesser, "Toward an affect-sensitive AutoTutor," *IEEE Intelligent Systems*, pp. 53–61, 2007.

[36] S. D'Mello, S. Craig, B. Gholson, S. Franklin, R. Picard, and A. Graesser, "Integrating affect sensors in an intelligent tutoring system," in *Affective interactions: The computer in the affective loop workshop at 2005 International Conference on intelligent user interfaces*, 2005, pp. 7–13.

[37] E. Douglas-Cowie, L. Devillers, J. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox, "Multimodal databases of everyday emotion: Facing up to complexity," in *9th European Conference on Speech Communication and Technology (Interspeech'2005)*, Lisbon, Portugal, Sept. 2005, pp. 813–816.

[38] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* Wiley-Interscience Publication, 2000.

[39] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. De Ridder, and D. Tax, "Prtools, a matlab toolbox for pattern recognition," *Delft University of Technology*, 2004.

[40] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, pp. 45–60, 1999.

[41] P. Ekman and W. Friesen, *Facial Action Coding System.* Palo Alto, CA: Consulting Psychologists Press, 1978.

[42] R. el Kaliouby and M. Goodwin, "iSET: interactive social-emotional toolkit for autism spectrum disorder," in *Proceedings of the 7th international conference on Interaction design and children.* ACM, 2008, pp. 77–80.

[43] T. Engen, N. Levy, and H. Schlosberg, "The dimensional analysis of a new series of facial expressions," *Journal of Experimental Psychology*, vol. 55, no. 5, pp. 454–458, 1958.

[44] F. Enos and J. Hirschberg, "A framework for eliciting emotional speech: Capitalizing on the actors process," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, Genoa,Italy, May 2006, pp. 6–10.

[45] S. Fagel, "Emotional McGurk Effect," in *Proceedings of the International Conference on Speech Prosody*, vol. 1, Dresden, 2006.

[46] N. Fragopanagos and J. Taylor, "Emotion recognition in human-computer interaction," *Neural Networks*, vol. 18, no. 4, pp. 389–405, 2005.

[47] N. Frijda, *The laws of emotion.* Lawrence Erlbaum Associates, 2007.

[48] H. Gish, M. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1991, pp. 873–876.

[49] O. Golan, S. Baron-Cohen, J. Hill, and M. Rutherford, "The Reading the Mind in the Voicetest-revised: A study of complex emotion recognition in adults with and without autism spectrum conditions," *Journal of autism and developmental disorders*, vol. 37, no. 6, pp. 1096–1106, 2007.

[50] F. Gosselin and P. G. Schyns, "Bubbles: a technique to reveal the use of information in recognition tasks," *Vision Research*, vol. 41, no. 17, pp. 2261 - 2271, 2001.

[51] J. Gratch, W. Mao, and S. Marsella, "Modeling social emotions and social attributions," *Cognition and multi-agent interaction: from cognitive modeling to social simulation*, p. 219, 2006.

[52] J. Gratch and S. Marsella, "A domain-independent framework for modeling emotion," *Cognitive Systems Research*, vol. 5, no. 4, pp. 269–306, 2004.

[53] M. Grimm and K. Kroschel, "Evaluation of Natural Emotions Using Self Assessment Manikins," *Proc. IEEE WSh. ASRU*, 2005.

[54] ——, "Rule-based emotion classification using acoustic features," in *Conf. on Telemedicine and Multimedia Communication*, Kajetany, Poland, Oct. 2005, p. 56.

[55] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[56] A. Hanjalic, "Extracting moods from pictures and sounds: Towards truly personalized tv," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 90–100, March 2006.

[57] U. Hess, S. Senécal, G. Kirouac, P. Herrera, P. Philippot, and R. Kleck, "Emotional expressivity in men and women: Stereotypes and self-perceptions," *Cognition and Emotion*, vol. 14, no. 5, pp. 609–642, 2000.

[58] J. Hietanen, J. Leppänen, M. Illi, and V. Surakka, "Evidence for the integration of audiovisual emotional information at the perceptual level of processing," *European Journal of Cognitive Psychology*, vol. 16, no. 6, pp. 769–790, 2004.

[59] R. Hobson, J. Ouston, and A. Lee, "Emotion recognition in autism: Coordinating faces and voices," *Psychological Medicine*, vol. 18, no. 4, pp. 911–923, 1988.

[60] C. Izard, *Human emotions*. Springer, 1977.

[61] A. Jaimes and N. Sebe, "Multimodal human–computer interaction: A survey," *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 116–134, 2007.

[62] E. Konstantinidis, M. Hitoglou-Antoniadou, A. Luneski, P. Bamidis, and M. Nikolaidou, "Using affective avatars and rich multimedia content for education of children with autism," in *Proceedings of the 2nd International Conference on PErvsive Technologies Related to Assistive Environments*. ACM, 2009, p. 58.

[63] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003, pp. 32–35.

[64] R. Lazarus, J. Averill, and E. Opton Jr, "Towards a cognitive theory of emotion," in *Feeling and emotion: The Loyola Symposium*, 1970, pp. 207–232.

[65] C.-C. Lee, C. Busso, S. Lee, and S. S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *InterSpeech*, Brighton, UK, Sep. 2009.

[66] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 320–323.

[67] C.-C. Lee, S. Lee, and S. S. Narayanan, "An analysis of multimodal cues of interruption in dyadic spoken interactions," in *InterSpeech*, Brisbane, Australia, Sep. 2008, pp. 1678–1681.

[68] Y. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," *Proc. of Int. Conf. on Machine Learning and Cybernetics*, vol. 8, pp. 4898–4901, Aug. 2005.

[69] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Applied Signal Processing*, pp. 1672–1687, Sept., 1 2004.

[70] C. Liu, K. Conn, N. Sarkar, and W. Stone, "Affect recognition in robot assisted rehabilitation of children with autism spectrum disorder," in *Proc. of the 15th IEEE Intl. Conf. on Robotics and Automation.* Citeseer, 2006.

[71] Y. Lu, I. Cohen, X. S. Zhou, and Q. Tian, "Feature selection using principal feature analysis," in *Int. Conf. on Multimedia.* New York, NY, USA: ACM, 2007, pp. 301–304.

[72] H. Macdonald, M. Rutter, P. Howlin, P. Rios, A. Le Conteur, C. Evered, and S. Folstein, "Recognition and expression of emotional cues by autistic and normal adults," *Journal of Child Psychology and Psychiatry*, vol. 30, no. 6, pp. 865–77, 1989.

[73] M. Madsen, R. El Kaliouby, M. Goodwin, and R. Picard, "Technology for just-in-time in-situ learning of facial affect for persons diagnosed with an autism spectrum disorder," in *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility.* ACM, 2008, pp. 19–26.

[74] G. Mandler, *Mind and emotion.* Wiley, 1975.

[75] J. Martin, R. Niewiadomski, L. Devillers, S. Buisine, and C. Pelachaud, "Multimodal complex emotions: Gesture expressivity and blended facial expressions," *International Journal of Humanoid Robotics*, vol. 3, no. 3, pp. 269–292, 2006.

[76] D. Massaro, "Fuzzy logical model of bimodal emotion perception: Comment on" The perception of emotions by ear and by eye" by de Gelder and Vroomen," *Cognition & Emotion*, vol. 14, no. 3, pp. 313–320, 2000.

[77] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[78] H. K. M. Meeren, C. C. R. J. van Heijnsbergen, and B. de Gelder, "Rapid perceptual integration of facial expression and emotional body language," *Proceedings of the National Academy of Sciences*, vol. 102, no. 45, pp. 16518–16523, 2005.

[79] A. Metallinou, C. Busso, S. Lee, and S. S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, March 2010.

[80] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Dallas, Texas, March 2010.

[81] ——, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *In submission*, 2010.

[82] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[83] D. Mobbs, N. Weiskopf, H. C. Lau, E. Featherstone, R. J. Dolan, and C. D. Frith, "The Kuleshov Effect: the influence of contextual framing on emotional attributionseffect: the influence of contextual framing on emotional attributions," *Social Cognitive and Affective Neuroscience*, vol. 1, no. 2, pp. 95–106, 2006.

[84] E. Mower, S. Lee, M. J. Matarić, and S. Narayanan, "Human perception of synthetic character emotions in the presence of conflicting and congruent vocal and facial expressions," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, NV, Apr. 2008, pp. 2201–2204.

[85] ——, "Joint-processing of audio-visual signals in human perception of conflicting synthetic character emotions," in *IEEE International Conference on Multimedia & Expo (ICME)*, Hannover, Germany, 2008, pp. 961–964.

[86] E. Mower, M. Matarić, and S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Transactions on Multimedia*, vol. 11, no. 4, 2009.

[87] ——, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, Accepted for Publication.

[88] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *ACII Special Session: Recognition of Non-Prototypical Emotion from Speech- The Final Frontier?*, Amsterdam, The Netherlands, September 2009.

[89] O. Mowrer, *Learning theory and behavior*. Wiley New York, 1960.

[90] M. Nicolao, C. Drioli, and P. Cosi, "Voice GMM modelling for FESTIVAL/MBROLA emotive TTS synthesis," in *International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, Sept. 17-21 2006, pp. 1794–1797.

[91] A. Ortony and T. Turner, "What's basic about basic emotions," *Psychological review*, vol. 97, no. 3, pp. 315–331, 1990.

[92] A. ORTONY and A. COLLINS, *The cognitive structure of emotions.* Cambridge university press, 1988.

[93] P. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies*, vol. 59, no. 1–2, pp. 157–183, 2003.

[94] M. Pantic, A. Pentland, A. Nijholt, and T. Huang, "Human computing and machine understanding of human behavior: a survey," *Artifical Intelligence for Human Computing*, pp. 47–71, 2007.

[95] M. Pantic, N. Sebe, J. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proceedings of the 13th annual ACM international conference on Multimedia.* ACM New York, NY, USA, 2005, pp. 669–676.

[96] R. W. Picard, *Affective Computing.* MIT Press, 1997.

[97] R. Plutchik, *Emotion: A psychoevolutionary synthesis.* Harper & Row, New York, 1980.

[98] P. Rani, C. Liu, and N. Sarkar, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Analysis & Applications*, vol. 9, no. 1, pp. 58–69, May 2006.

[99] P. Robbel, M. Hoque, and C. Breazeal, "An integrated approach to emotional speech and gesture synthesis in humanoid robots," in *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots.* ACM, 2009, pp. 1–4.

[100] J. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[101] J. Russell and L. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant," *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 805–819, 1999.

[102] S. Schacter and J. Singer, "Cognitive and emotional determinants of emotional states," *Psychological Review*, vol. 69, pp. 379–399, 1962.

[103] H. Schlosberg, "Three dimensions of emotion," *Psychological review*, vol. 61, no. 2, pp. 81–88, 1954.

[104] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," *Interspeech*, pp. 2253–2256, 2007.

[105] B. Schuller, S. Steidl, , and A. Batliner, "The interspeech 2009 emotion challenge," in *Interspeech*, Brighton, UK, 2009, pp. 312–315.

[106] N. Sebe, I. Cohen, T. Gevers, and T. Huang, "Emotion recognition based on joint visual and audio cues," in *Int. Conf. on Pattern Recognition*, 2006, pp. 1136–1139.

[107] D. Seppi, A. Batliner, B. Schuller, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, and V. Aharonson, "Patterns, Prototypes, Performance: Classifying Emotional User States," *InterSpeech*, pp. 601–604, 2008.

[108] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, ""Of all things the measure is man": Automatic classification of emotions and inter-labeler consistency," in *ICASSP, 2005.*, vol. 1, 2005, pp. 317–320.

[109] S. Sutton, R. Cole, J. de Villiers, J. Schalkwyk, P. Vermeulen, M. Macon, Y. Yan, E. Kaiser, B. Rundle, K. Shobaki *et al.*, "Universal speech tools: the cslu toolkit," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, Nov. – Dec. 1998, pp. 3221–3224.

[110] W. Swartout, J. Gratch, R. Hill, E. Hovy, S. Marsella, J. Rickel, and D. Traum, "Toward virtual humans," *AI Magazine*, vol. 27, no. 2, pp. 96–108, 2006.

[111] W. Swartout, J. Gratch, R. Hill Jr, E. Hovy, S. Marsella, J. Rickel, and D. Traum, "Toward virtual humans," *AI Magazine*, vol. 27, no. 2, p. 96, 2006.

[112] M. Swerts and E. Krahmer, "The importance of different facial areas for signalling visual prominence," in *International Conference on Spoken Language (ICSLP)*, Pittsburgh, PA, USA, Sept. 2006, pp. 1280–1283.

[113] F. Thomas and O. Johnston, *Disney animation: The illusion of life.* Abbeville Press New York, 1981.

[114] S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, 2006.

[115] K. Truong, M. Neerincx, and D. van Leeuwen, "Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, Sept. 2008, pp. 318–321.

[116] N. Tsapatsoulis, A. Raouzaiou, S. Kollias, R. Cowie, and E. Douglas-Cowie, "Emotion recognition and synthesis based on mpeg-4 fap's," in *MPEG-4 Facial Animation: The Standard, Implementation, and Applications*, I. S. Pandzic and R. Forchheimer, Eds. John Wiley & Sons, Ltd., 2002, ch. 9, pp. 141–167.

[117] V. Vapnik, *Statistical Learning Theory.* Wiley, New York, 1998.

[118] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *IEEE International Conference on Multimedia & Expo (ICME)*, Los Alamitos, CA, USA, 2005, pp. 474–477.

[119] M. Wimmer, B. Schuller, D. Arsic, G. Rigoll, and B. Radig, "Low-level fusion of audio and video feature for multi-modal emotion recognition," in *Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, Madeira, Portugal, 2008, pp. 145–151.

[120] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with Java implementations," in *Proceedings of ANNES International Workshop on emerging Engineering and Connectionnist-based Information Systems*, vol. 99, Dunedin, New Zealand, 1999, pp. 192–196.

[121] R. Xu and D. Wunsch, *Clustering*. Wiley-IEEE Press, 2008.

[122] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech." in *International Conference on Spoken Language Processing International Conference on Spoken Language Processing (ICSLP)*, Jeju Island, South Korea, 2004, pp. 2193–2196.

[123] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*. Entropic Cambridge Research Laboratory Cambridge, England, 1997.

[124] J. Zelenski and R. Larsen, "The distribution of basic emotions in everyday life: A state and trait perspective from experience sampling data," *Journal of Research in Personality*, vol. 34, no. 2, pp. 178–197, 2000.

[125] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, pp. 39–58, 2009.

[126] Z. Zeng, J. Tu, B. Pianfetti, and T. Huang, "Audio–Visual Affective Expression Recognition Through Multistream Fused HMM," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.

Appendix

Journal Papers

1. textbfEmily Mower, Maja J Matarić and Shrikanth S. Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotional Profiles." *IEEE Transactions on Audio, Speech and Language Processing.* Accepted for publication, August 2010.

2. **Emily Mower**, Maja J Matarić and Shrikanth S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information." *IEEE Transactions on Multimedia*, 11:5(843-855). August 2009.

3. Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, **Emily Mower**, Samuel Kim, Jeannette Chang, Sungbok Lee and Shrikanth S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database." *Journal of Language Resources and Evaluation*, 42:4(335-359). December 2008.

4. Michael Grimm, Kristian Kroschel, **Emily Mower** and Shrikanth S. Narayanan, "Primitives-based evaluation and estimations of emotions in speech." *Speech Communication*, 49:10-11(787-800). November 2007.

Conference Papers

1. **Emily Mower**, Maja J Matarić and Shrikanth S. Narayanan, "Robust Representations for Out-of-Domain Emotions Using Emotion Profiles." In Proceedings of *IEEE Workshop on Spoken Language Technology (SLT)*, Berkeley, CA, December 2010.

2. **Emily Mower**, Kyu Jeong Han, Sungbok Lee and Shrikanth S. Narayanan. "A Cluster-Profile Representation of Emotion Using Agglomerative Hierarchical Clustering." In Proceedings of *International Speech Communication Association (InterSpeech)*, Makuhari, Japan. September 2010.

3. Dongrui Wu, Thomas Parsons, **Emily Mower** and Shrikanth S. Narayanan. "Speech Emotion Estimation in 3D Space." In Proceedings of *IEEE International Conference on Multimedia & Expo (ICME)*, Singapore, July 2010.

4. **Emily Mower**, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, Shrikanth Narayanan. "Interpreting Ambiguous Emotional Expressions." In Proceedings of *ACII Special Session: Recognition of Non-Prototypical Emotion from Speech- The Final Frontier?*. Amsterdam, The Netherlands, September 2009.

5. **Emily Mower**, Maja J Matarić, Shrikanth Narayanan. "Evaluating Evaluators: A Case Study in Understanding the Benefits and Pitfalls of Multi-Evaluator Modeling." In Proceedings of *International Speech Communication Association (InterSpeech)*. Brighton, England, September 2009.

6. Chi-Chun Lee, **Emily Mower**, Carlos Busso, Sungbok Lee and Shrikanth S. Narayanan, Emotion recognition using a hierarchical binary decision tree approach, In Proceedings of *International Speech Communication Association (InterSpeech)*, Brighton, England, September. 2009. [Emotion Challenge Winner]

7. **Emily Mower**, Maja J Matarić, Shrikanth Narayanan. "Selection of Emotionally Salient Audio-Visual Features for Modeling Human Evaluations of Synthetic Character Emotion Displays." In Proceedings of *IEEE International Symposium on Multimedia (ISM)*. Berkeley, California, December 2008.

8. **Emily Mower**, Sungbok Lee, Maja J Matarić, Shrikanth Narayanan. "Joint-processing of audio-visual signals in human perception of conflicting synthetic character emotions." In Proceedings of *IEEE International Conference on Multimedia & Expo (ICME)*, Hannover, Germany, June 2008.

9. **Emily Mower**, Sungbok Lee, Maja J Matarić, Shrikanth Narayanan. "Human perception of synthetic character emotions in the presence of conflicting and congruent vocal and facial expressions." In Proceedings of *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Las Vegas, Nevada, March-April 2008.

10. **Emily K. Mower**, David J. Feil-Seifer, Maja J Matarić, and Shrikanth Narayanan. "Investigating Implicit Cues for User State Estimation in Human-Robot Interaction Using Physiological Measurements." In Proceedings of *IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*, Jeju Island, South Korea, Aug 2007.

11. Michael Grimm, **Emily Mower**, Kristian Kroschel, and Shrikanth Narayanan. "Combining categorical and primitives-based emotion recognition." In Proceedings of *European Signal Processing Conference (EUSIPCO)*, Florence, Italy, September 2006.

12. Wenting Zhou, Weichen Wu, Nathan Palmer, **Emily Mower**, Noah Daniels, Lenore Cowen, Anselm Blumer. "Microarray Data Analysis of Survival Times of Patients with Lung Adenocarcinomas Using ADC and K-Medians Clustering." In Proceedings of *Critical Assessment of Massive Data Analysis (CAMDA)*. Durham, North Carolina, November 2003.

Organization Committees

1. Co-organizer (with Frank Broz and Marek Michalowski) *It's All In the Timing: Representing and Reasoning About Time in Interactive Behavior* at the AAAI Spring Symposium, Stanford University, Palo Alto, CA (March 2010).

2. Co-organizer (with Alan Wagner, Rosemarijn Looije, Maxim Makatchev, and Marek Michalowski) *HRI Young Pioneers Workshop* at the IEEE International Conference on Human-Robot Interaction (HRI), La Jolla, CA (March 2009).

Awards and Honors

- Herbert Kunzel Engineering Fellowship (2010-2011)
- Achievement Rewards For College Scientists (ARCS 2009)
- Emotion Challenge, Classification Challenge Winner, International Speech Communication Association Conference (InterSpeech 2009)
- Intel Foundation Fellowship (2008-2010)
- Herbert Kunzel Engineering Fellowship (2007-2008)
- NSF Graduate Student Fellowship (2004-2007)
- Graduated summa cum laude, Dean's list all semesters
- Received Undergraduate Thesis Honors (2004)
- Tau Beta Pi (2004)
- Undergraduate Computing Research Association Distributed Mentor Project, CRA DMP (Summer 2003)
- Eta Kappa Nu- Electrical and Computer Engineering Honor Society (2003)