A Framework for Automatic Human Emotion Classification Using Emotion Profiles

Article *in* IEEE Transactions on Audio Speech and Language Processing · August 2011

DOI: 10.1109/TASL.2010.2076804 · Source: IEEE Xplore

Emily Mower Provost
University of Michigan
119 PUBLICATIONS **6,561** CITATIONS

Shrikanth S Narayanan
University of Southern California
1,141 PUBLICATIONS **34,889** CITATIONS

A Framework for Automatic Human Emotion Classification Using Emotion Profiles

Emily Mower, *Student Member, IEEE*, Maja J Matarić, *Senior Member, IEEE*, and Shrikanth Narayanan, *Fellow, IEEE*

***Abstract*—Automatic recognition of emotion is becoming an increasingly important component in the design process for affect-sensitive human–machine interaction (HMI) systems. Well-designed emotion recognition systems have the potential to augment HMI systems by providing additional user state details and by informing the design of emotionally relevant and emotionally targeted synthetic behavior. This paper describes an emotion classification paradigm, based on emotion profiles (EPs). This paradigm is an approach to interpret the emotional content of naturalistic human expression by providing multiple probabilistic class labels, rather than a single hard label. EPs provide an assessment of the emotion content of an utterance in terms of a set of simple categorical emotions: anger; happiness; neutrality; and sadness. This method can accurately capture the general emotional label (attaining an accuracy of 68.2% in our experiment on the IEMOCAP data) in addition to identifying underlying emotional properties of highly emotionally ambiguous utterances. This capability is beneficial when dealing with naturalistic human emotional expressions, which are often not well described by a single semantic label.**

***Index Terms*—Emotion profiles, multimodal emotion classification, nonprototypical emotions.**

I. Introduction

THE proper design of affective agents requires an *a priori* understanding of human emotional perception. Models used for the automatic recognition of emotion can provide designers with a means 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 fosters design improvements for both emotionally relevant and emotionally targeted expressions for use in human–computer and human–robot interaction. This understanding will further improve human-centered design, necessary for widespread adoption of this affective technology [1].

Human perception of naturalistic expressions of emotion is difficult to estimate. This difficultly is in part due to the presence of complex emotions, defined as emotions that contain shades of multiple affective classes [2]–[5]. For example, in [3], 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 [3]. The lack of emotional purity in natural expressions of emotion must be considered when designing systems to anticipate human emotional perception of non-stereotypical 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 (EPs). 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 [6] as a method for expressing the variability inherent in multi-evaluator expressions. These EPs represented 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 [7], 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. 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.

The current study presents 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 utterance-level classification system is composed of four binary support vector machine (SVM) classifiers, one for each of the emotions considered

Manuscript received December 04, 2009; revised April 30, 2010; accepted July 15, 2010. Date of publication September 27, 2010; date of current version May 04, 2011. This work was supported in part by the National Science Foundation, in part by the U.S. Army, in part by the Intel Foundation Fellowship, and in part by the Herbert Kunzel Engineering Fellowship. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Daniel P. W. Ellis.

E. Mower and S. Narayanan are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: mower@usc.edu; shri@sipi.usc.edu).

M. J. Matarić is with the Department of Computer Science, University of Southern California University Park, Los Angeles, CA 90089 USA (e-mail: mataric@usc.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2010.2076804

(anger, happiness, sadness, and neutrality). EPs are created by weighting the output of the four SVM classifiers by an estimate of the confidence of the assignment. The utterance is assigned to the emotion class with the highest level of confidence, represented by the EP.

The results are presented across three data types of varying levels of ambiguity, based on evaluator reports: 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 emotion-specific feature selection in conjunction with emotional profiling-based support vector machines 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% [8]. 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. It extends the EP description of [7] to include a measure of the confidence with which an emotional assessment is made. This extension allows the EPs to represent emotions as complex blends, rather than discrete assignments. Furthermore, 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.

The remainder of the paper will be organized as follows. Section II will describe the work that motivated the creation of EPs. Section III will describe the data utilized in this study. Section IV will describe the features and feature selection method used in this study. Section V will present an overview of the developed classification system. Sections VI and VII will detail the results of the emotional classification task. Finally, Section VIII will provide concluding remarks and future work.

II. Related Work

Ambiguity in emotion expression and perception is a natural part of human communication. This ambiguity can be mitigated through the designation of an utterance as either a prototypical or non-prototypical emotional episode. These labels can be used to provide a coarse description of the ambiguity present in an utterance. These terms are described by Russell in [9]. 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; and there is an alignment of the psychophysiological processes [9]. 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.

There are many sources of 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 or more 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 the difference between two classes of emotion (e.g., irritation and anger) are not strongly differentiated. Finally, ambiguity may also occur when a sequence of emotions is expressed consecutively within the boundary of one utterance [10]. 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 received attention. At the Interspeech Conference in 2009, there was an Emotion Challenge [11] special session to focus on the classification of emotional ambiguous utterances. 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 [2], [12] and synthesis [13], [14].

EPs can 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 [2], [15], classification output [7], and perception (as a combination of multiple emotions) resulting from one group's actions towards another group [16]. EPs can be thought of as a quantified description of the properties that exist in the emotion classes considered. In [17], 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, expressions labeled as angry must be sufficiently similar to each other and sufficiently different from the expressions labeled as other emotions. 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. As shown in 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 [15].

Engineering models provide an important avenue for developing 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 [5], [18][19]. For a survey of the field, see [1].

The classification technique employed in this paper, SVMs, has been used previously in emotion classification tasks [2], [18], [20]. 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.

The feature selection method utilized in this study is information gain, which has also been used widely in the literature [2], [20]. Information gain is used to estimate the importance of the features using a classifier independent method. Information gain does not decorrelate the feature space. However, humans rely on a set of correlated features to disambiguate emotional utterances. Thus, our decision to utilize this type of feature selection was motivated by our desire to better understand the features that are important to humans in emotion identification. The purpose of this work is not to demonstrate the efficacy of either the SVM 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.

The data utilized in this study are from the USC IEMOCAP dataset, collected at the University of Southern California [21] (discussed in more detail in Section III). The USC IEMOCAP dataset[1] is an emotional audio-visual database with facial motion-capture information. This database has been used for studies ranging from interaction modeling [14], [22] to evaluator modeling [23]. This dataset has been previously used for two classification studies. The first study [7] classified audio utterances using hidden Markov models (HMMs) into one of five states: anger, happiness, neutrality, sadness, and frustration. The accuracies ranged from 47.3% for the classification of emotionally well-defined, or prototypical, utterances to 35.1% for the classification of emotionally ambiguous, or non-prototypical, utterances.

In the presented work, we do not consider the class of frustration. The class of frustration within our dataset overlaps with both anger and sadness [21]. The goal of the presented work is to create a multi-emotion representation of affective utterances. Such a representation seeks to describe the affective content of an utterance in terms of its similarity to other affective classes. Consequently, the components of the representation should include emotion classes without high degrees of mutual overlap to decrease redundancy. Future work includes classifying frustration using EPs constructed from the classes of anger, happiness, neutrality, and sadness to further verify the efficacy of the technique.

In [8], the authors performed a profiling-based multimodal classification experiment on the USC IEMOCAP database. The authors utilized Mel filterbank coefficients (MFBs), head motion features, and facial features selected using emotion-independent principal feature analysis (PFA) [24]. 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 outputs 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 62.4%.

III. Data Description

The dataset utilized in this study is the USC IEMOCAP database [21]. 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 challenges of and benefits to utilizing actor-solicited data are discussed more fully in [25]–[27].

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 temporal order, thus 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. The kappa for the dataset over all utterances and emotional labels was 0.27. The kappa over only utterances where the evaluators reached a majority consensus was 0.40. When the classes of happiness and excitation were merged (as they were in the work presented in this paper) and the classes of disgust, fear, and surprised were merged with other, the overall kappa increased to 0.35 and the kappa over sentences in which there was majority consensus increased to 0.48. More detail on the evaluation of this dataset can be found in [21].

In the dimensional evaluations, the evaluators were asked to rate the valence (positive versus negative), activation (calm versus excited), and dominance (passive versus aggressive) properties of the emotion expression. The dimensional evaluation task was completed by a separate set of six evaluators, again evaluating overlapping subsets of the data. Each emo-

[1] We are currently releasing Session 1 data: http://sail.usc.edu/iemocap/.

tional utterance within the database was labelled by at least two dimensional evaluators [21].

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 [21].

A. Emotion Expression Types

Emotional data can be considered either as a cohesive whole or as merged subsets 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 (Section II) 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 the intersection of all of the evaluators' categorical ratings (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 label at the intersection of all of the evaluators' evaluated sets. 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.

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 into a group referred to as "happy" to combat data sparsity issues. 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 to indicate that one of the labels is in the emotional set, 2L to indicate that two of the labels are in this set, or nL to indicate that there were more than two labels from the set. The 1L data were extremely biased toward the class of anger Table I and there were only 80 utterances in the nL group; therefore, this study will focus only on the 2L emotions. Table I shows the distribution of the data across the three expression classes.

TABLE I
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

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

B. 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 paper are tagged with at least one label from the set: angry; happy/excited; neutral; sad.

IV. 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.

A. 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 [2], [5], [18], [20]. The spectral features include Mel frequency coefficients (MFBs). 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 property by binning the signal with triangular bins of increasing width as the frequency increases. Mel filterbank cepstral coefficients (MFCCs) are commonly used in both speech and emotion classification. MFCCs are discrete cosine transformed (DCT) MFBs. The DCT decorrelates the feature space. Previous research has demonstrated that MFBs are better discriminative features than MFCCs across all phoneme classes for emotion classification [28].

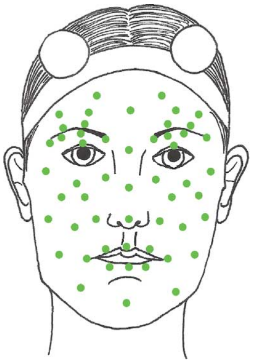
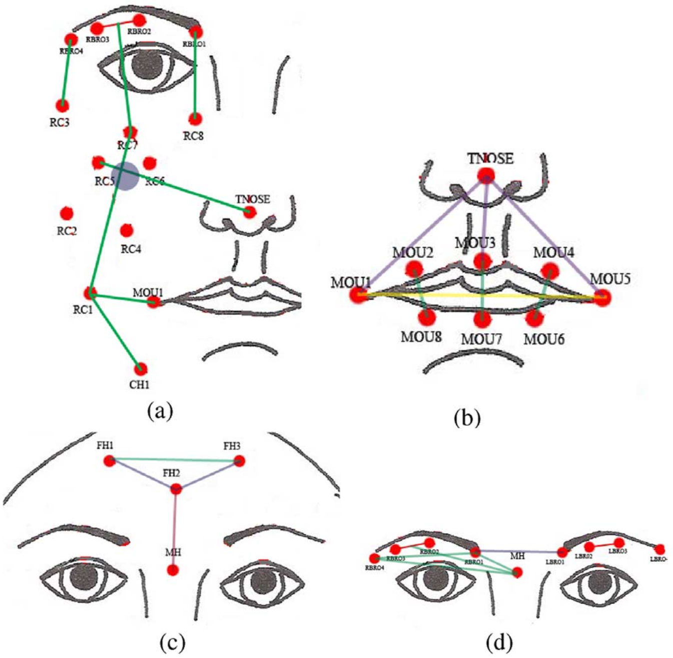


Fig. 1. Location of the IR markers used in the motion-capture data collection.



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

Fig. 2. Facial features. (a) Cheek. (b) Mouth. (c) Forehead. (d) Eyebrow.

B. 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 [29], adapted to our facial motion-capture configuration. These features were extracted using motion-capture markers (Figs. 1 and 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 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 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 distances describing the up-down motion of the eyebrows, eyebrow squish, and the distance to the center of the eyebrows. Each distance is expressed in three features defining the x, y, and z-coordinates in space.

C. 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 [12].

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.

D. Feature Selection

There were a total of 685 features extracted. However, there were only 3000 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) [30]. 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 [31]. Information gain has previously been used to select a relevant feature subset in [32] and [33]. Information gain does not create an uncorrelated feature set, which is often preferable for many classification 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 tenfold cross-validation.

E. Final Feature Set

The number of features was determined empirically, optimizing for accuracy. The final feature set included the top 85 features (see Table II 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 [7], [19].

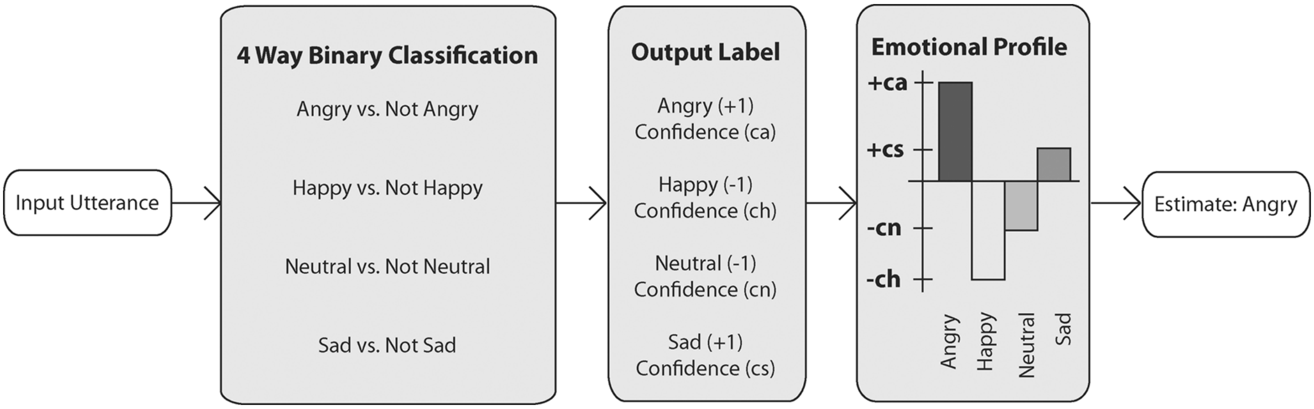


Fig. 3. EP-SVM system diagram. An input utterance is classified using a four-way binary classification. This classification results in four output labels representing membership in the class $(+1)$ or lack thereof (-1). This membership is weighted by the confidence (distance from the hyperplane). The final emotion label is the most highly confident assessment.

TABLE II
Average Percentage of Each Feature Over the 40 Speaker-Independent Emotion-Specific Feature Sets (10 Speakers * 4 Emotions)

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	0.10	0.10	0.28	–	0.05
Sad	–	–	–	–	0.04	0.96

V. Classification of Emotion Perception: Emotion Profile Support Vector Machine (EP-SVM)

The classification system utilized in this experiment consists of four binary SVMs. The EPs were created using the four binary outputs and an approximate measure of classifier confidence. The final label of the utterance is the most confident assignment in the EP (see Fig. 3 for the system diagram and an example).

A. 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 [19], [33]–[36]. 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 [37] 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 (Section IV.D). Each of the emotional SVMs was trained discriminatively using a self versus 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 [38], 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 versus 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.

B. Creation of Emotional Profiles

The emotional 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. In feature space, data points that are closer to the margin are more easily confused with the opposing class than data points further from the margin. Thus, 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.

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 EP-SVM technique. However, EPs can be created using K-nearest neighbors (KNNs), 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 [38]–[40].

C. Final Decision

An emotional utterance is assigned to an emotion class based 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 fully integrated system is referred to as an Emotion Profile Support Vector Machine, or EP-SVM.

VI. 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-SVM 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 versus rejection). The comparison between the EP-SVM 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 three (i.e., general, prototypical, and non-prototypical MV) data subsets.

A. 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 accuracy for the EP-SVM classification system is 68.2% (Table III). This outperforms both chance (40.1%) and the simplified SVM (55.9%). The difference between the EP-SVM 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.* [8] (described in Section II) 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 (Fig. 4). All four profiles demonstrate 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 $(\text{F-measure} = 0.71)$.

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%).

B. 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-SVM classifier was 81.7% (Table III). This outperformed chance (49.8%) and the simplified SVM (75.5%). The difference between the EP-SVM and baseline is significant at $\alpha \leq 0.001$ (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 (Fig. 5) 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 Figs. 4(a) to 5(a), 4(b) to 5(b), and 4(d) to 5(d). The mean confidence value for the angry, happy, and sad data were higher when training and testing on prototypical data.

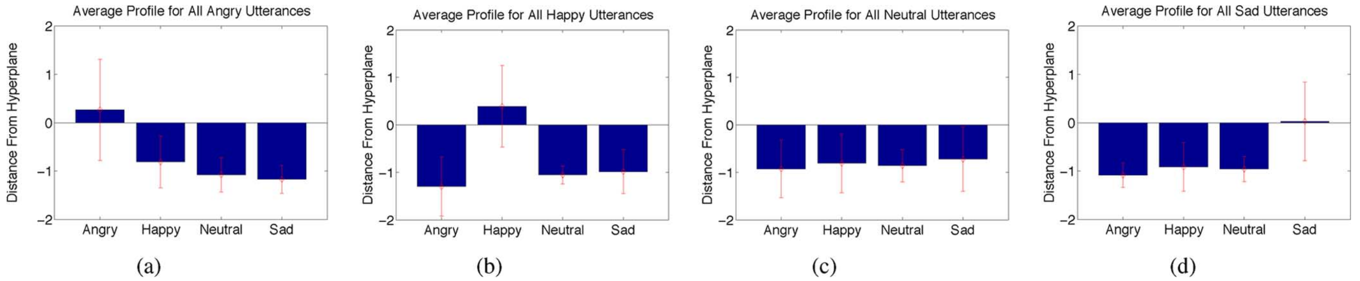


Fig. 4. Average emotional profiles for all (both prototypical and non-prototypical) utterances. The error bars represent the standard deviation. (a) Angry. (b) Happy. (c) Neutral. (d) Sad.

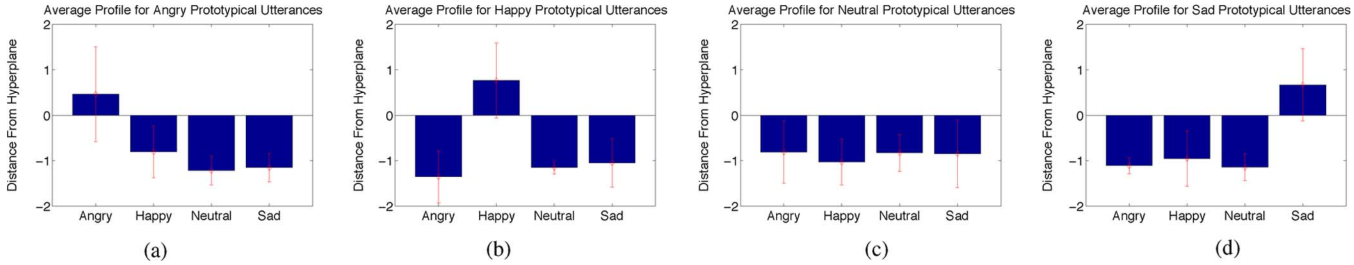


Fig. 5. Average emotional profiles for prototypical utterances. The error bars represent the standard deviation. (a) Angry. (b) Happy. (c) Neutral. (d) Sad.

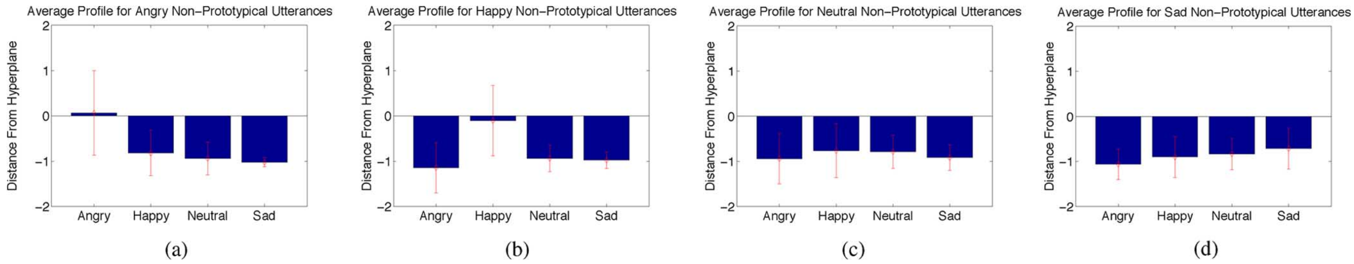


Fig. 6. Average emotional profiles for non-prototypical utterances. The error bars represent the standard deviation. (a) Angry. (b) Happy. (c) Neutral. (d) Sad.

TABLE III
EP-SVM 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

Data Type	Emotion	Precision	Recall	F
Full EP-SVM	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-SVM	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-SVM	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			

C. 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 not significant ($\alpha \leq 0.001$, difference of proportions test). The EP-SVM also outperforms chance (31.4%). The class-by-class comparison can be seen in Table III. 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-SVM classification 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 (Fig. 6) 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 versus 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 [Fig. 6(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 Figs. 5(c) and 6(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.

D. 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-SVM 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-SVM 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 IV(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 IV(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-SVM 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. 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 IV(b).

TABLE IV
MAJOR–MINOR EMOTION ANALYSIS (a) TOTAL NUMBER OF EMOTIONS WITH SECONDARY LABELS IN THE ANGRY, HAPPY, NEUTRAL, SAD SET, (b) RESULTS WHERE THE MAJOR AND MINOR EMOTIONS ARE CORRECTLY IDENTIFIED, (c) RESULTS WHERE THE MAJOR AND MINOR EMOTIONS WERE BOTH IN THE TOP TWO REPORTED LABELS

(a)

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)

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)

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

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 vice versa). This phenomenon was also studied. Table IV(b) 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-SVM 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-SVM 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.

VII. Results and Discussion: The Non-Prototypical NMV Dataset

One of the hallmarks of the EP-SVM method is its ability to interpret ambiguous utterances. EPs can be used to detect subtle emotional information given inherently ambiguous data. In this paper, 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-SVM 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-SVM 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.

A. 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 EP-SVMs. 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 III-A.

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.

TABLE V
Results of the EP-SVM 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

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				

The results demonstrate (Table V) 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-SVM would then be in direct opposition for the per-class accuracy measure. However, accuracy over the entire 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) of the 2L dataset was calculated by finding the emotion that co-occurred with the other emotion labels most commonly. The class of neutrality occurred with 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%. 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-SVM could accurately identify the presence of anger.

The results further indicate that the EP-SVM 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 detecting neutrality, this supports the more precise detection of the EP-SVM over a range of emotions.

The EP-SVM method is able to capture information that cannot be captured by the simplified baseline SVM discussed earlier. In Fig. 7, all of the histograms demonstrate that, on average, all four binary classifiers return non-membership results (-1). The confidence component allows the EP-SVM to disambiguate the subtle emotional content of the non-prototypical NMV utterances. The average profiles of Fig. 7 demonstrate that the EPs are able to capture the emotion content of these utterances.

B. 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 rep-

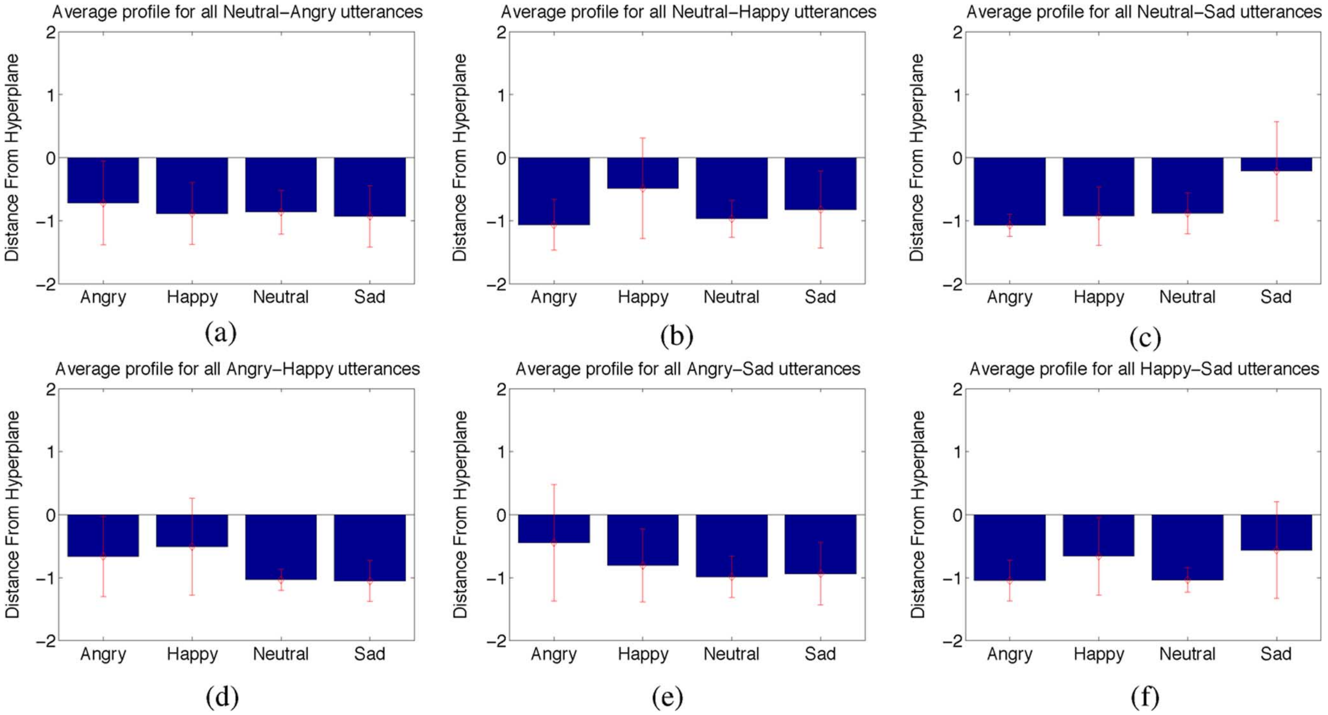


Fig. 7. Average emotional profiles for the non-prototypical NMV utterances. The error bars represent the standard deviation. (a) Neutral–Angry. (b) Neutral–Happy. (c) Neutral–Sad. (d) Angry–Happy. (e) Angry–Sad. (f) Happy–Sad.

resentations 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.

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 VI).

TABLE VI
ANOVA ANALYSIS OF THE DIFFERENCE IN GROUP MEANS BETWEEN CO-OCCURRING AND NON-OCCURRING EMOTIONS WITHIN AN EP-SET (INDIVIDUAL EP-SET EXPERIMENT). ($- = \alpha \leq 0.1, * = \alpha \leq 0.05, ** = \alpha \leq 0.01, *** = \alpha \leq 0.001$)

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

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 were 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, an EP-set whose co-occurrence emotions were entirely disjoint. This demonstrates that EP-sets must be differentiated based on more than their co-occurrence emotions (Table VII).

The following two analyses determine if the representation of the individual co-occurring emotions differs between the native

TABLE VII

ANOVA Analyses of the Differences Between Reported Emotions in Profiles in Which They Were Reported Versus Profiles in Which They Were Not. Note That the EP$_1$ Versus EP$_2$ Is an Interaction of an ANOVA Analysis of the Set EP$_1$ Versus EP$_2$ and an ANOVA Analysis of the Representation of the Individual Emotions in Each EP-Set. ($- = \alpha \leq 0.1, * = \alpha \leq 0.05, ** = \alpha \leq 0.01, *** = \alpha \leq 0.001$)

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	*		*	**

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 VII).

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 versus EP$_2$ experiment.

The EP$_1$ versus 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 VII).

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.

VIII. Conclusion

Natural human expressions are combinations of underlying emotions. Models aimed at automated processing of these emotions should reflect this fact. 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-SVM 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 predefined 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. Future work will investigate classification of the generated profiles. Furthermore, as discussed in Section I, frustration is not included in either the EP testing or training. Future work 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, future work 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 paper 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.

REFERENCES

[1] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, no. 1, pp. 39–58, Jan. 2009.

[2] 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," in *Proc. InterSpeech*, 2008, pp. 601–604.

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

[4] R. Plutchik, *Emotion: A Psychoevolutionary Synthesis*. New York: Harper & Row, 1980.

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

[6] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, ""Of all things the measure is man": Automatic classification of emotions and interlabeler consistency," in *Proc. ICASSP'05*, 2005, vol. 1, pp. 317–320.

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

[8] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Dallas, TX, Mar. 2010, pp. 2462–2465.

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

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

[11] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proc. Interspeech*, Brighton, U.K., 2009, pp. 312–315.

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

[13] 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.

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

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

[16] C. Cottrell and S. Neuberg, "Different emotional reactions to different groups: A sociofunctional threat-based approach to prejudice," *J. Personal. Soc. Psychol.*, vol. 88, no. 5, pp. 770–789, 2005.

[17] L. Barrett, "Are emotions natural kinds?," *Perspectives on Psychol. Sci.*, vol. 1, no. 1, pp. 28–58, 2006.

[18] 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 *Proc. Int. Conf. Comput. Vis. Theory Applicat. (VISAPP)*, Madeira, Portugal, 2008, pp. 145–151.

[19] 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 *Proc. Int. Conf. Multimodal Interfaces*, State Park, PA, Oct. 2004, pp. 205–211.

[20] 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," in *Proc. Interspeech*, 2007, pp. 2253–2256.

[21] 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," *J. Lang. Res. Eval.*, pp. 335–359, Nov. 5, 2008.

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

[23] E. Mower, M. Mataric, and S. Narayanan, "Evaluating evaluators: A case study in understanding the benefits and pitfalls of multi-evaluator modeling," in *Proc. 10th Annu. Conf. Interspeech*, Brighton, U.K., Sep. 2009.

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

[25] 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.

[26] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: A closer look," in *Proc. 2nd Int. Workshop Emotion: Corpora for Research on Emotion and Affect, Int. Conf. Lang. Res. Eval. (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[27] F. Enos and J. Hirschberg, "A framework for eliciting emotional speech: Capitalizing on the actor's process," in *Proc. 1st Int. Workshop Emotion: Corpora for Research on Emotion and Affect (Int. Conf. Lang. Res. Eval. (LREC 2006))*, Genoa, Italy, May 2006, pp. 6–10.

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

[29] 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. New York: Wiley, 2002, ch. 9, pp. 141–167.

[30] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.

[31] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with Java implementations," in *Proc. ANNES Int. Workshop Emerging Eng. Connectionnist-Based Inf. Syst.*, Dunedin, New Zealand, 1999, vol. 99, pp. 192–196.

[32] P. Oudeyer, "The production and recognition of emotions in speech: Features and algorithms," *Proc. Int. Human-Comput. Studies*, vol. 59, no. 1–2, pp. 157–183, 2003.

[33] 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 Trans. Multimedia*, vol. 11, no. 4, pp. 843–855, 2009.

[34] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Recognizing facial expression: Machine learning and application to spontaneous behavior," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recogn.*, 2005, vol. 2, pp. 568–573.

[35] Y. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Aug. 2005, vol. 8, pp. 4898–4901.

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

[37] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines: And Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[38] 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 *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, 2004, vol. 1, pp. 592–597.

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

[40] O. Kwon, K. Chan, J. Hao, and T. Lee, "Emotion recognition by speech signals," in *Proc. 8th Eur. Conf. Speech Commun. Technol.*, 2003, pp. 32–35.

Emily Mower (S'07) received the B.S. degree in electrical engineering from Tufts University, Boston, MA, in 2004 and the M.S. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2007. She is currently pursuing the Ph.D. degree in electrical engineering as a member of the Signal Analysis and Interpretation Laboratory (SAIL), USC.

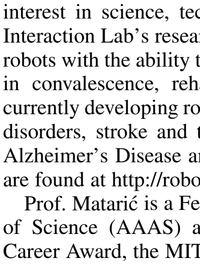
Her research interests are in human-centered speech and video processing and multimodal interfaces design. The goals of her research are motivated by the complexities of human emotion generation and perception. She seeks to provide a computational account of how humans perceive emotional utterances ("emotion perception") and combine this with knowledge gleaned from perception estimation studies ("emotion recognition") to develop systems capable of interpreting naturalistic expressions of emotion utilizing a new quantification measure ("emotion profiles"). She has published several articles in these areas and is a contributor to the winning paper in the classifier category of the 2009 Interspeech Emotion Challenge.

Ms. Mower is a member of Tau-Beta-Pi and Eta-Kappa-Nu. She has been awarded the National Science Foundation Graduate Research Fellowship (2004–2007), the Herbert Kunzel Engineering Fellowship from USC (2007–2008, 2010–2011), and the Intel Research Fellowship (2008–2010).

Maja J Matarić (SM'06) received the B.S. degree in computer science from the University of Kansas, Lawrence, in 1987 and the M.S. degree in computer science and the Ph.D. degree in computer science and artificial intelligence from the Massachusetts Institute of Technology (MIT), Cambridge, in 1990 and 1994, respectively.

She is a Professor of computer science, neuroscience, and pediatrics at the University of Southern California (USC), Los Angeles, founding Director of the USC Center for Robotics and Embedded Systems, Co-Director of the USC Robotics Research Lab, and Senior Associate Dean for Research in the USC Viterbi School of Engineering. She is actively involved in K-12 educational outreach, having obtained federal and corporate grants to develop free open-source curricular materials for elementary and middle-school robotics courses in order to engage student interest in science, technology, engineering, and math (STEM) topics. Her Interaction Lab's research into socially assistive robotics is aimed at endowing robots with the ability to help people through individual non-contact assistance in convalescence, rehabilitation, training, and education. Her research is currently developing robot-assisted therapies for children with autism spectrum disorders, stroke and traumatic brain injury survivors, and individuals with Alzheimer's Disease and other forms of dementia. Details about her research are found at http://robotics.usc.edu/interaction/.

Prof. Matarić is a Fellow of the American Association for the Advancement of Science (AAAS) and recipient of the Okawa Foundation Award, NSF Career Award, the MIT TR100 Innovation Award, and the IEEE Robotics and Automation Society Early Career Award. She served as the elected president of the USC faculty and the Academic Senate. At USC, she was awarded the Viterbi School of Engineering Service Award and Junior Research Award, the Provost's Center for Interdisciplinary Research Fellowship, the Mellon Mentoring Award, the Academic Senate Distinguished Faculty Service Award, and a Remarkable Woman Award. She is featured in the science documentary movie "Me & Isaac Newton," in *The New Yorker* ("Robots that Care" by Jerome Groopman, 2009), *Popular Science* ("The New Face of Autism Therapy," 2010), the IEEE Spectrum ("Caregiver Robots," 2010), and is one of the LA Times Magazine 2010 Visionaries. She is an associate editor of three major journals and has published extensively. She serves on a number of advisory boards, including the National Science Foundation Computing and Information Sciences and Engineering (CISE) Division Advisory Committee, and the Willow Garage and Evolution Robotics Scientific Advisory Boards.



Shrikanth S. Narayanan (M'95–SM'02–F'09) is the Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, and holds appointments as Professor of Electrical Engineering, Computer Science, Linguistics, and Psychology and as the Founding Director of the Ming Hsieh Institute. Prior to USC, he was with AT&T Bell Labs and AT&T Research from 1995 to 2000. At USC , he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered information processing and communication technologies with a special emphasis on behavioral signal processing and informatics.

Prof. Narayanan is a Fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an Editor for the *Computer Speech and Language Journal* and an Associate Editor for the IEEE Transactions on Multimedia, IEEE Transactions on Affective Computing, and the *Journal of the Acoustical Society of America*. He was also previously an Associate Editor of the IEEE Transactions of Speech and Audio Processing (2000–2004) and the IEEE Signal Processing Magazine (2005–2008). He is a recipient of a number of honors including Best Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C. M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010–2011. He has published over 400 papers and has eight granted U.S. patents.