Robust Representations for Out-of-Domain Emotions Using Emotion Profiles

Emily Mower[†‡], *Maja J Matarić*[‡], *Shrikanth Narayanan*[†]

University of Southern California
[†] Signal Analysis and Interpretation Laboratory, [‡]Interaction Laboratory
University Park, Los Angeles, California, USA 90089
mower@usc.edu, mataric@usc.edu, shri@sipi.usc.edu

Abstract

The proper representation of emotion is of vital importance for human-machine interaction. A correct understanding of emotion would allow interactive technology to appropriately respond and adapt to users. In human-machine interaction scenarios it is likely that over the course of an interaction, the human interaction partner will express an emotion not seen during the training of the machine's emotion models. It is therefore crucial to prepare for such eventualities by developing robust representations of emotion that can distinctly represent emotions regardless of whether the data were seen during training of the representation. This novel work demonstrates that an Emotion Profile (EP) representation introduced in [1], a representation composed of the confidences of four binary emotion-specific classifiers, can distinctly represent emotions unseen during training. The classification accuracy increases by only 0.35% over the full dataset when the data excluded from the EP training is included. The results demonstrate that EPs are a robust method for emotion representation.

Index Terms: Emotion Representation, Emotion Classification, Emotion Profiles, Audio-Visual Emotion

1. Introduction

Ambiguous emotional expressions are a natural part of human communication. In 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 distinctly characterize emotions that it has not previously observed. This paper will assess the ability of Emotion Profiles (EP), introduced in [1], 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.

In [1] we introduced the idea of Emotion Profiles (EP) for emotion classification. We extended the EP representation in [2], to analyze how ambiguous emotions could be studied. EPs were also used to fuse different modalities in classification [3]. EP-like representations have also been used to represent the evaluations of a set of evaluators [4, 5] and to represent perception based on actions (as a function of multiple emotions) [6]. In the current work, we further analyze this technique to study its ability to represent out-of-domain data.

EPs describe an utterance in terms of an estimated combination of multiple emotions. The EPs represent the presence or absence of emotions by a confidence score derived from classification. The question remains as to how the EPs should be structured. EPs can contain any number (n) of components, representing the presence or absence of the n emotions linked to those components. In [7] we explored how data-driven clusters can be effectively used to generate profiles suggesting that it may not be necessary for the dimensionality of the profile to match the number of target affective classes. By relying on the semantically meaningful "ideal" clusters (e.g., angry, happy) we demonstrate how five separate emotion categories can be represented using a four-dimensional EP. We utilize data from the classes of anger, happiness, neutrality, sadness, and frustration. We compare the accuracies of this five-class classification problem when using either a four (angry, happy, neutral, sad) or five (plus frustration) dimensional EP to represent the data. The results demonstrate that there is not a significant difference in accuracies (either per-class or overall) between the four and five dimensional representations. This indicates that EPs need not include an exhaustive list of emotion categories. Instead, they should include only the categories necessary to "span" an emotional space. Furthermore, these results suggest that the EPs are a robust representation that can be used to distinctly represent data unseen during the generation of the EPs.

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.

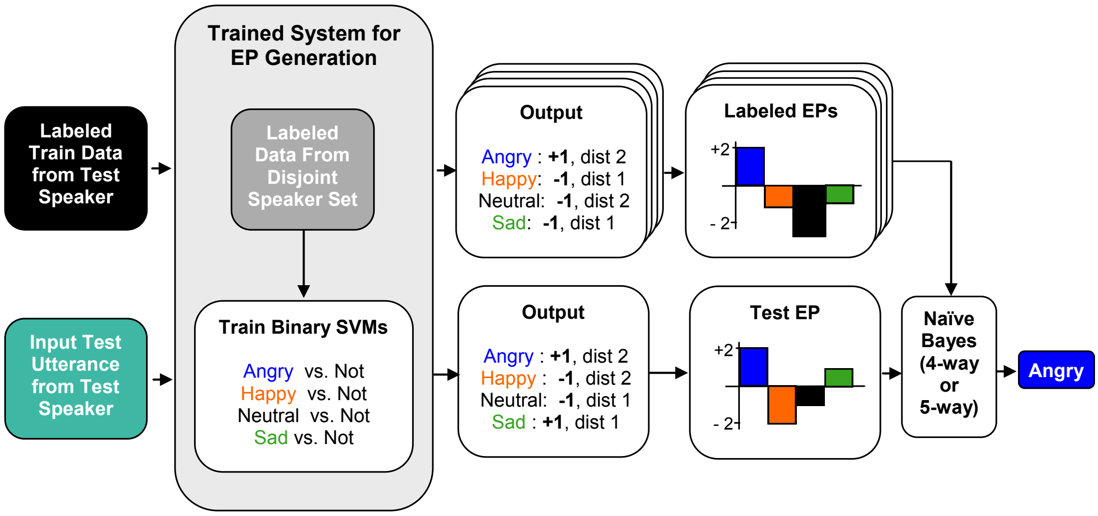


Figure 1: The EP-based classification system diagram. This example demonstrates the correct classification of a nonprototypical angry utterance (a mixture of anger and sadness).

2. Description of Data

2.1. IEMOCAP Database

The representative capability of the EP representation was evaluated using the USC IEMOCAP Dataset collected at the University of Southern California [8]. This dataset contains data from five mixed-gender pairs of actors (10 actors total). The data include video, audio, and motion-capture recordings.

The data were collected with two elicitation strategies, scripted dialogues and improvisation. The benefit of using this style of data collection is that it allows for a wider coverage of the emotional range than datasets composed of natural interactions. Furthermore, the collection style permitted the elicitation of natural human interaction patterns [8].

The data were evaluated using categorical and dimensional labeling. Categorical labels were used in this study; details of the dimensional labeling can be found in [8]. The categorical labels describe the contents of an emotional utterance in terms of a semantic label. In this dataset, the utterances were tagged with at least one label (per evaluator) from the set of: anger, happiness, neutrality, sadness, excitation, frustration, surprise, disgust, fear, and other. There were at least three evaluators per utterance. We utilized the utterances with a majority vote from the set of: angry, happy, excited (merged with the class of happiness), neutral, sad, and frustrated.

2.2. Data Definitions

The data were partitioned into groups defined by the level of agreement between evaluators. These groups were labeled prototypical and nonprototypical. These definitions are derived from those of Russell [9]. *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 1.

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

Table 1: The distribution of the emotion classes in the prototypical and nonprototypical categories.

3. Emotion Profiles

One theory of emotion asserts that there exist "basic emotions". An emotion is basic if it is differentiable from all other emotions [10]. 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 as the combination of the basic emotions of anger and sadness [11]. 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 discernable 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.

3.1. Construction of an EP

In this work, the EPs are constructed using Support Vector Machines (SVM). SVMs have been shown to be effective in emotion classification tasks [12, 13, 14, 15]. SVM is a maximum-margin classifier; it projects the input feature space into a

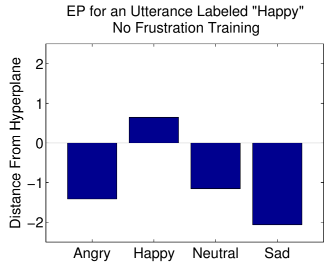


Figure 2: The EP of an utterance tagged as 'happy'. This EP has been trained without frustration data.

space of potentially higher-dimension to find an optimal separating hyperplane that maximizes the distances between the two classes.

Emotion-specific SVMs with Radial Basis Function (RBF) kernels are trained for each class as self vs. other classifiers (e.g., angry vs. not angry) using SMO optimization [16], an analytic approach that avoids the otherwise time consuming calculation of the solution to the large quadratic programming problem required by SVM. The profiles are created by weighting each of the n-memberships ($\pm$ 1) by the raw distances to the hyperplane (see Figure 1 for the system diagram). The raw distance to the hyperplane is defined by the distance, in the higher dimensional space, from the individual point to the hyperplane boundary.

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 the class label of the data points are more easily confused than points further away from the boundary in the feature space (or projected feature space). 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."

The EPs 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 are 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 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 2 for an example of a four-dimensional EP.

3.2. 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 2 the label would be happy). This approach was employed in [2]. However, this voting-based labeling does not take advantage of the information in the minority components of the EP. Instead of relying on choosing the maximal confidence, the final emotion can be selected after classifying the generated profile. In this work, we use Naïve Bayes classification.

3.3. Speaker-Dependent and Speaker-Independent Components

The 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 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.

4. 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, maximum, minimum, range, 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 [4, 17, 18, 19]. 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 [20].

The video features are based on Facial Animation Parameters (FAP) [21]. 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 [2].

4.1. Feature Selection

The initial feature set consists of 685 features. The feature selection method utilized is Principle Feature Analysis (PFA) [22]. 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, determined empirically. This feature selection algorithm has been used in emotion classification tasks on the USC IEMOCAP dataset [3, 23].

5. 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. 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 previously published work [3] 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 [3] 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.

	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 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.

6.1. Classification with EP Frustration Training

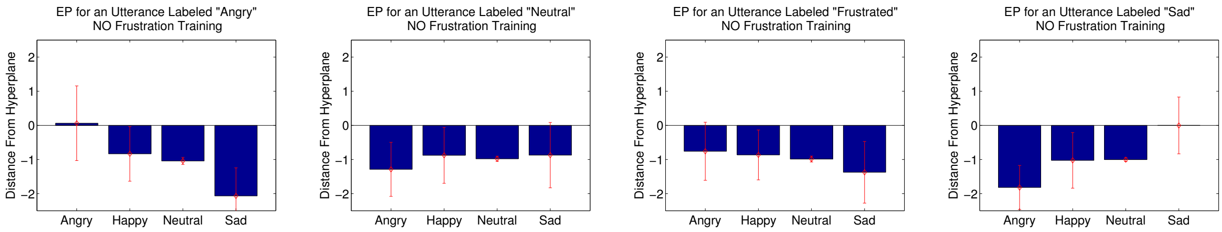
This set of results demonstrates the classification performance when a five-dimensional EP representation is employed. The hypothesis is that training EPs with frustration will not provide significant benefit to the overall five-class classification accuracy when compared with the five-class classification 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 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 [8] 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% 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.

6.2. Classification without EP Frustration Training

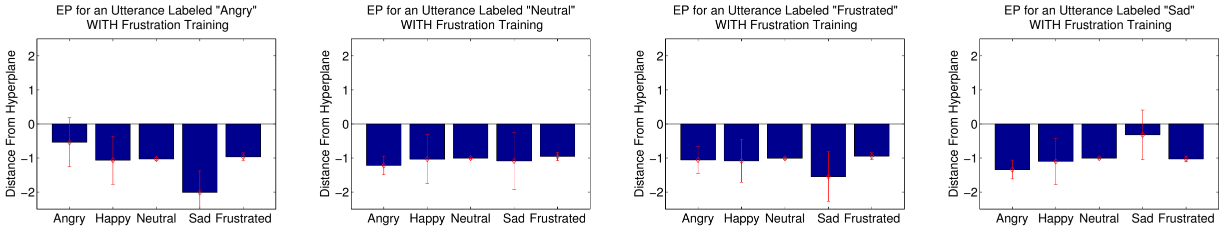
In the final scenario the EPs are trained only with angry, happy, neutral, and sad data, while the Naïve Bayes classifier must classify emotions from all five categories. In this training scenario, the EPs must distinctly represent an emotion not seen during training. The results will be compared to 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 perfor-



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

Figure 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 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.

mance 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 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.

6.3. EP Representation of Frustration

Previous work has demonstrated that the utterances labeled as frustrated in this database are confused both by human evaluators [8] and by machine learning algorithms [1] (audio-only analysis). The graphs of Figures 3 and 4 further support the inherent difficulty in characterizing this ambiguous emotion. Figure 4 demonstrates that on average the emotion of "frustration" is represented as not present for emotions labeled as 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 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 the EP, can be more accurately characterized than neutral utterances, which have been historically difficult to characterize in this database [2, 1, 3].

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

Table 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 average EPs of Figures 3 and 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 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 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).

7. Conclusions

This paper demonstrates the efficacy of a novel 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.

8. References

[1] 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, pp. 1 – 8.

[2] E. Mower, M. Matarić, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, Accepted for Publication.

[3] 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.

[4] 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.

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

[6] C.A. Cottrell and S.L. 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.

[7] E. Mower, K. J. Han, S. Lee, and S. Narayanan, "A Cluster-Profile Representation of Emotion Using Agglomerative Hierarchical Clustering: New Techniques in Profile-Based Representations for Emotion Classification," in *InterSpeech 2010*, Makuhari, Japan, September 2010.

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

[9] J.A. Russell and L.F. 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.

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

[11] J.M. Zelenski and R.J. 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.

[12] 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 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, CA, 2005, vol. 2, pp. 568–573.

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

[14] 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.

[15] 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. 5, pp. 843–855, 2009.

[16] J.C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*. MIT press, 1999, pp. 185–208.

[17] 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 *InterSpeech*, Antwerp, Belgium, 2007, pp. 2253–2256.

[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 *Int. Conf. on Computer Vision Theory and Applications (VISAPP)*, Madeira, Portugal, 2008, pp. 145–151.

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

[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] 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., chapter 9, pp. 141–167. John Wiley & Sons, Ltd., 2002.

[22] Yijuan Lu, Ira Cohen, Xiang Sean Zhou, and Qi Tian, "Feature selection using principal feature analysis," in *Int. Conf. on Multimedia*, New York, NY, USA, 2007, pp. 301–304, ACM.

[23] 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, pp. 2474 – 2477.