A Cluster-Profile Representation of Emotion Using Agglomerative Hierarchical Clustering

Emily Mower, Kyu J. Han, Sungbok Lee, Shrikanth Narayanan

University of Southern California
Signal Analysis and Interpretation Laboratory
University Park, Los Angeles, California, USA 90089
mower@usc.edu, kyuhan@usc.edu, sungbokl@usc.edu, shri@sipi.usc.edu

Abstract

The proper representation of emotion is critical to automatic classification systems. In previous research, we demonstrated that emotion profile (EP) based representations are effective for this task. In EP-based representations, emotions are expressed in terms of underlying affective components from the subset of anger, happiness, neutrality, and sadness. The current study explores cluster profiles (CP), an alternate profile representation in which the components are no longer semantic labels, but clusters inherent in the feature space. This unsupervised clustering of the feature space permits the application of a system-level semi-supervised learning paradigm. The results demonstrate that CPs are similarly discriminative to EPs (EP classification accuracy: 68.37% vs. 69.25% for the CP-based classification). This suggests that exhaustive labeling of a representative training corpus may not be necessary for emotion classification tasks.

Index Terms: Emotion Profiles, Agglomerative Hierarchical Clustering, Emotion Representation, Emotion Classification

1. Introduction

Interactive affective technologies require detailed models of human emotion for accurate user state determination. These models are commonly trained using supervised learning algorithms. However, such algorithms typically require labeled training corpora, the collection of which is often expensive and time-intensive. This study presents a system-level semi-supervised approach to user-specific emotion-classification using a novel Cluster-Profile (CP) representation of emotion.

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

In previous work, we demonstrated the efficacy of an Emotion-Profile (EP) based representation for classification [1, 2]. EPs are a quantitative representation of the affective content of an utterance in terms of the presence or absence of a set of component emotions. In these studies the components of the profile were the semantic, or categorical, labels: angry, happy, neutral, and sad. However, it is not clear that the profiles must be constructed using these types of semantic components.

In this study we investigate a system-level semi-supervised approach for emotion classification. The classification system is broken down into four steps: speaker-independent feature selection, speaker-independent clustering, speaker-independent profile generation, and speaker-dependent classification. The feature selection method is the unsupervised Principal Feature Analysis (PFA), an extension of Principal Component Analysis, also used in [3, 4]. The data are clustered using unsupervised agglomerative hierarchical clustering of the emotional space. These clusters are used to train cluster-specific Support Vector Machines (SVM) whose output are the components of the CPs. Finally, the emotion content of the utterance is assessed by classifying over the generated CPs. The system is semi-supervised because the feature selection, clustering, and profile generation are unsupervised while the final classification step is supervised. The unsupervised portion establishes a data-dependent representation for the affective test data using the majority of the training data. The final supervised classification utilizes the generated CPs for Naïve Bayes classification.

The CP classification method outperforms the EP classification by 0.88% absolute (69.25% vs. 68.37%). This result demonstrates the efficacy of the CP-based classification system. The CPs represent emotional utterances in n-components, where n is the number of clusters. This comparable performance of the CP and EP representations suggests that given training sets with expressions from a non-disjoint set of emotion classes, it may be necessary to label only a subset of the data. These results cannot be compared directly to any published work due to the final speaker-dependent classification step. However, this performs comparably to fused GMM-HMM method presented in [4] (62.42%). The novelty of the current work lies in its new definition of a profile and an assessment of the necessity of the semantic profile dimensions utilized in the EPs.

2. Description of Data

2.1. IEMOCAP Database

The discriminative power of the CP-representation is evaluated using the USC IEMOCAP database, collected at the University of Southern California (USC) [5]. This dataset is dyadic and interactive, recorded using a mixed elicitation strategy of emotionally-targeted scripted and improvisational scenarios. There are a total of five dyadic mixed-gender pairs of actors (10 actors total). The recordings include audio, video, and motion-capture measurements (hands, head, face). The motion-capture data was recorded for one actor at a time in each dyadic pair to allow for a greater fidelity in recording. As a result, only the half of the utterances have associated motion-capture data.

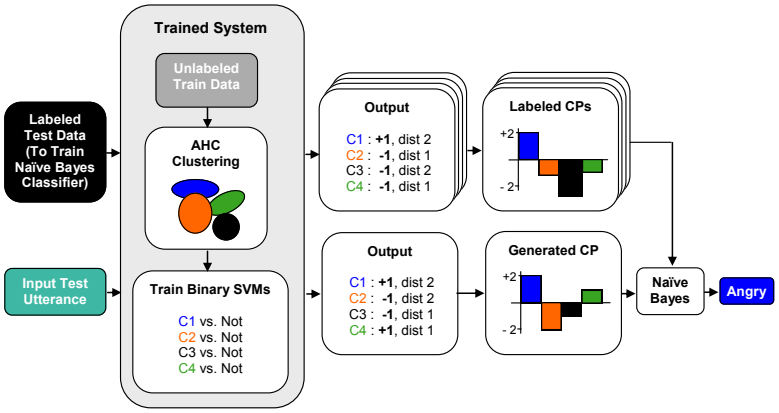

Figure 1: The CP-based classification system diagram. This example demonstrates the correct classification of a nonprototypical angry utterance (a mixture of anger and sadness).

The data were evaluated using two labeling strategies: categorical and dimensional labeling. Dimensional labels capture the properties of an affective utterance such as valence (positive vs. negative) and activation (calm vs. excited). Categorical labels are assignments of semantic labels (e.g., angry, happy, neutral, etc.). The ten categorical labels in this dataset are: angry, happy, neutral, sad, frustrated, excited, surprised, fearful, disgust, and other. Each utterance was labeled by at least two dimensional evaluators and at least three categorical labelers. The categorical labels were exclusively used in this study.

3. Emotion and Cluster Profiles

Profile-based representations describe affective utterances over a set of affective components rather than in terms of a single mathematical (e.g., a valence of '3') or semantic (e.g., 'angry') label. This added flexibility is beneficial when the emotional character of the speech is subtle. In previous work [1, 2], EPs were implemented as four-dimensional representations of emotion. The dimensions expressed the degree of presence or absence of each of the emotions: angry, happy, neutral, and sad. This subset was chosen to minimize affective overlap in our experimental dataset. In the current work, we explore profile generation using an unsupervised component-generation approach (Figure 1).

3.1. Description of the Train and Test Sets

The dataset considered consists of 4,806 utterances across the ten emotional labels and ten-speakers. The profile generation ("training") is speaker-independent while the final classification ("testing") is speaker-dependent (Figure 1). For each speaker, the training data (for unsupervised clustering and profile generation) consist of all of the utterances not spoken by the speaker. These data contain unlabeled emotions from all 10 emotion categories. The testing data consist only of utterances spoken by the speaker from the set: angry, happy, neutral, and sad.

3.2. Unsupervised Clustering for CPs

The feature space is clustered using the unsupervised agglomerative hierarchical clustering (AHC) over the unlabeled training data. This hierarchical clustering strategy circumvents the initialization issues common to other clustering approaches (e.g., k-means or GMM-EM [6, 7]). AHC is a bottom-up process, which is more computationally efficient than top-down (divisive) clustering. Research has demonstrated that AHC can be applied to many clustering tasks and is effective. This clustering approach is of particular popularity in the field of speaker clustering and diarization [8].

Initially, AHC considers each data point a cluster. Then, at every iteration, it selects the closest pair of clusters to merge. This merging procedure continues until a pre-set stopping criterion is satisfied. Generalized likelihood ratio (GLR) [9] is used to measure inter-cluster distance at every stage of AHC. The stopping criterion is a manually pre-set number of clusters, n. This work will explore the utility of considering different numbers of clusters in the CP construction.

3.3. Construction of a Profile

EPs and CPs are both constructed using the output from Support Vector Machines (SVM). The efficacy of SVMs has been demonstrated in emotion classification [10–14]. SVM is a maximum margin classifier that projects input data into a higher dimensional space to find an optimal separating hyperplane between two classes. The distance from one point in the projected space to the hyperplane can be interpreted as the confidence of the classifier's assessment. Points closer to the hyperplane are representative of data that are more easily confused in the projected-space. These points represent utterances that cannot be as confidently labeled as utterances further from the decision hyperplane.

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

The performance of the cluster-based profile representation will be compared to that of the pre-specified emotion-based profile representation. In the EPs, the original clusters are defined by the emotion labels: angry, happy, neutral, and sad. Binary self vs. other SVMs are trained over each of the emotion clusters. As in the CP formulation, the EPs are composed of the membership function weighted by the distance from the hyperplane for each of the outputs for a given utterance. These profiles are four-dimensional.

The final step is performing classification over the gener-

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

Table 1: The CP-based classification results. The entries in bold font indicate the best accuracy or f-measure recorded.

ated profiles (both CP and EP). This n-dimensional classification is performed using Naïve Bayes. Gaussian Mixture Models, KNN, and Discriminant Analysis were also explored, but were not as effective. Only Naïve Bayes results will be reported.

4. Features Extraction and Selection

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

The set of audio features included: intensity, pitch, and the first 13 Mel Filterbank Coefficients (MFB). Intensity and pitch have been used successfully in emotion classification studies [15–18]. In this work, MFBs are also used. MFBs are less common than Mel-Frequency Cepstral Coefficients (MFCC) in emotion research. However, previous work has demonstrated that MFB features are more effective for emotion classification than MFCCs across all broad phoneme classes [19].

The motion-capture features utilized in this work are derived from Facial Animation Parameters (FAP) [20]. These features are part of the MPEG-4 standard and represent distances between points on the face. The FAPs were adapted to the motion-capture configuration used in the USC IEMOCAP data recording. The facial features were broken into groups by facial region. These regions included: mouth, cheeks, forehead, and eyebrows. These features were also used in [2].

4.1. Feature Selection

The initial feature set has 685 features. The feature set size is reduced using the unsupervised method of Principal Feature Analysis (PFA) [21]. This feature selection technique has been used successfully on the USC IEMOCAP dataset [3, 4]. PFA is an extension of Principal Component Analysis (PCA) in which the generated PCA components are used to identify representative features in the original feature space. As in PCA, the top p-eigenvectors are identified. The features are then clustered using their projections over the top eigenvectors. The final feature set consists of the the features closest to each cluster mean. The number of clusters was chosen empirically.

The feature sets were identified in a speaker-independent fashion. For example, the selected features for Speaker 1 were analyzed using the data from Speakers 2-10. The final feature set size was 20-features.

5. Experimental Methods

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

The EP-based classification is presented as a baseline performance metric. The EPs are trained in a speaker-independent fashion (e.g., EPs for Speaker 1 are trained using the data from Speakers 2-10) over the semantic labels of angry, happy, neutral, and sad. In CP-based classification the speaker-independent training data are first clustered into n-clusters using the aforementioned clustering approach. The CPs are then constructed using the output from the n-SVMs trained on each cluster's data (one SVM for each of the n clusters). In both profile-based methods, the final emotion assessment is made using Naïve Bayes over the generated profiles. The performance of the Naïve Bayes classifier is assessed using leave-one-out cross-validation (see system diagram, Figure 1).

6. Results

6.1. EP Classification

The EP-based classification will serve as a comparative baseline for the CP-based classification results. In the EP-based classification, the accuracy was 68.37%. The emotion-specific results can be seen in Table 1. The classes of anger, happiness, and sadness were well recognized (f-measure > 0.69). The class of neutrality was relatively poorly recognized. This trend is common in this database, where neutrality remains an emotion class that is not well understood [2, 4].

6.2. CP Classification

In this task, the maximal accuracy occurred with 15 clusters. The maximal accuracy was 69.25% (Table 1). The emotions of anger, happiness, and sadness were again well recognized (f-measure > 0.69). It should be noted that in the CP-representation, the f-measure for the class of neutrality increased to 0.54. This represents a 9% absolute and 20.00% relative improvement. This result suggests that CP-based representations are more effective for capturing inherently ambiguous classes of emotion than EP-based representations.

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

7. Conclusions

This paper presents a novel system-level semi-supervised technique to classify the emotion content of utterances using a profile-based technique. The CP-based classification nonsignificantly outperformed the EP-based classification by 0.88% with 15 clusters. This suggests that both data-driven and knowledge-

driven clusters are effective for profile generation. The CP-based representation alleviates the need for exhaustive labeling of the training corpus, requiring instead a labelling of a small subset of the data.

CP-based classification required at least 11-clusters to match the accuracy obtained by EP-based classification. The f-measures obtained in the EP-based classification for angry, happy, neutral, and sad was never obtained in the CP-based classification for anger and required 11, 7, and 11 clusters respectively for the classes of happiness, neutrality, and sadness. This suggests that the EP-based representation is more compact than this implementation of the CP-based representation. This further suggests that the clusters generated from the semantic labels of angry, happy, neutral, and sad are very effective for capturing the affective feature properties of the utterances, supporting the use of the components of anger, happiness, neutrality, and sadness in the EP-based representation.

Although, as stated in the Introduction, the results presented in this paper cannot be directly compared to previously published methods, both the EP- and CP-based classification systems produce similar accuracies to those seen in the literature (62.42%) [4]. This demonstrates that both profile-based representations are effective for emotion classification tasks.

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

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

8. Acknowledgements

This work was supported by the National Science Foundation, the US Army, and the Intel Foundation PhD Fellowship.

9. 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.

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

[3] 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.

[4] 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.

[5] 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.

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

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

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

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

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

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

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

[13] 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.

[14] 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. on Multimedia*, vol. 11, no. 4, 2009.

[15] 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.

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

[17] M. Woellmer, 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)*, 2008.

[18] 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.

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

[20] 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.

[21] 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.