A HIERARCHICAL STATIC-DYNAMIC FRAMEWORK FOR EMOTION CLASSIFICATION

Emily Mower and Shrikanth Narayanan

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

ABSTRACT

The goal of emotion classification is to estimate an emotion label, given representative data and discriminative features. Humans are very good at deriving high-level representations of emotion state and integrating this information over time to arrive at a final judgment. However, currently, most emotion classification algorithms do not use this technique. This paper presents a hierarchical static-dynamic emotion classification framework that estimates high-level emotional judgments and locally integrates this information over time to arrive at a final estimate of the affective label. The results suggest that this framework for emotion classification leads to more accurate results than either purely static or purely dynamic strategies.
Index Terms: Emotion Representation, Emotion Classification, Emotion Profiles, Audio-Visual Emotion

1. INTRODUCTION

Emotion classification is the process of extracting affective information from a set of signals and assigning an emotional label based on the feature fluctuations. Currently, computational techniques do not take advantage of the higher-level processing inherent in humans' classification of emotion. These systems often have difficulty arriving at an accurate interpretation of affective state given cross-modal interactions and the influence of context. This paper presents a system that that approximates the higher-level processing of humans by introducing a framework that integrates this high-level information over a locally static and globally dynamic context.

The goal of this work is to understand the trade-offs between dynamic and static modeling and how the strengths of each can be leveraged in an emotion classification framework. In static emotion modeling an utterance is analyzed using statistical functionals extracted over the entire utterance. The features utilized in this type of framework describe the properties over the entire clip. However, utterances with a high-degree of emotional fluctuation or utterances that are long may not be well characterized by this representation scheme. For example, an angry utterance may be mostly unemotional with a single angry outburst. The static utterance-level feature representations of this utterance may render it confusable with neutral speech even though, perceptually, it is clearly angry during the outburst. Therefore, with longer utterances it may be advantageous to utilize dynamic modeling. However, dynamic modeling also presents disadvantages in that it is often difficult to take context into account. In this work we present a hierarchical system that integrates aspects of both static and dynamic modeling to more fully take advantage of the information present affective utterances.

In our earlier work we introduced the concept of Emotion Profiles (EP) as a static modeling technique for representing the affective characteristics of audio-visual utterances [1]. EPs are vector-based characterizations that express the degree of presence or absence of a representative set of basic emotions in an utterance. In this work the set of basic emotions includes anger, happiness, neutrality, and sadness. Schuller and Devillers investigated the utility of chunking test and train utterances in [2]. The authors found that performance gains saturated at utterance chunks of 2 seconds. In [3] the authors utilized bi-directional long-short term modeling and found that the incorporation of within-utterance context improved the overall classification accuracy. In [4] the authors classified facial emotion expressions from video using two approaches: static and dynamic classification. The work presented in this paper builds on the findings of [2–4] and incorporates an EP-representation to create a hierarchical static-dynamic framework for audio-visual emotion classification.

This paper utilizes EPs as a static mid-level representation of emotion serving as input to dynamic Hidden Markov Models (HMM). The goal is to approximate an emotional grammar, describing the emotional transitions within an utterance, that will better capture the emotional modulations inherent in human audio-visual speech. The EPs are used to approximate local context via statistical functionals extracted over sliding windows ranging in length from 0.25 to 2 seconds. The HMMs are then used to integrate the emotional flow, described by the EPs, over the course of the utterances. The hierarchical static-dynamic modeling better captures the affective information in audio-visual utterances when compared to purely static or purely dynamic modeling.

The novelty of this work is in its introduction of an EP-based hierarchical static-dynamic framework for emotion classification. The results demonstrate that this system is effective for capturing the emotional fluctuations of utterances in a speaker-independent fashion. The maximal accuracy of the hierarchical static-dynamic system is 66.89% and the accuracy over 0.25 second window is 64.20%. These accuracies are greater than that of either the static or the dynamic systems, 63.72% and 56.74%, respectively.

2. DESCRIPTION OF DATA

The data utilized in this study comes from the USC IEMOCAP database [5], which contains dyadic acted data from five male-female pairs of actors (ten actors total). The data were recorded using microphone, video, and motion capture cameras and contain over 12-hours of affective expression segmented into over 10,000 utterances (approximately half of which have motion capture recordings). The data were labeled using two strategies, dimensional (describing utterances using affective properties) and categorical (describing emotion using semantic labels) evaluation protocols. The work presented in this paper only utilizes the categorical evaluations. Each utterance was evaluated by at least three categorical evaluators (out of six total). The utterances modeled in this paper have a majority-voted ground truth from the set of angry, happy, neutral, and sad. Please see [5] for additional database details.

3. EMOTION PROFILES

Emotion Profiles are a method for characterizing the affective properties of an utterance in terms of a set of emotion components. The affective components represent the degree of presence or absence of each of the emotion components within the utterance. For example in an utterance with a ground truth of "angry" the EP may characterize the utterance as: +1.4 angry, -0.3 happy, -1.2 neutral, -0.4 sad. This can be interpreted as confidence in the presence of anger and in the absence of neutrality and lack of confidence in the either the presence or absence of happiness and sadness (Figure 1). In this paper, the EP components are a subset of the common "basic" emotions of angry, happy, and sad with an additional component of neutral.

3.1. Construction of an EP

The EPs employed in this study are implemented using Support Vector Machines (SVM). SVMs have been shown to be effective for emotion classification [6, 7]. SVMs are binary maximum margin classifiers that find a separating hyperplane maximizing the distance from the hyperplane to the points closest to the hyperplane.

EPs are generated by first training self vs. other classifiers over the angry, happy, neutral, and sad utterances in the training data. The testing data is classified using these trained SVMs and the final EPs are constructed by weighting the binary membership (± 1) by the distance from the hyperplane for each of the four emotion classifiers. The motivation is that points further away from the hyperplane are less confusable in the feature space (or projected feature space) and are therefore clearer examples of a given class. For example, a point that received a +1 (membership) and was close to the hyperplane would be an unsure example of the given class. If a different point received a -1 (lack of membership) and was far from the hyperplane, this point could be interpreted as a confident example of an utterance not in the target class. This measure also has been shown to be perceptually relevant; points that lie in the corresponding stereotypical regions of the emotional space are generally confident examples of the semantic classes [8]. The four estimates are combined into a single vector (for two graphical examples of profiles, see Figure 1).

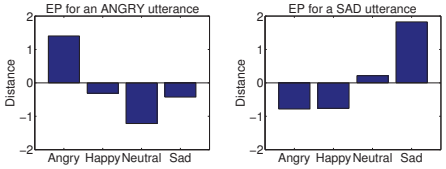


Fig. 1. Example emotion profiles of an angry and a sad utterance.

4. FEATURE EXTRACTION AND SELECTION

The features are extracted from the audio and motion capture data. The audio features include pitch, energy, and Mel Filterbank (MFB) coefficients. Pitch and energy are commonly used in emotion classification [9]. MFBs are less common in emotion classification than Mel Frequency Cepstral Coefficients (MFCC). Previous research has suggested that MFBs may be more useful than MFCCs for emotion classification [10].

The video features utilized in this study are based on Facial Animation Parameters (FAP), part of the MPEG-4 standard. FAPs are distances between specific points on the face and can be used to provide emotionally realistic animations. The configuration of the motion-capture points used in this study is similar to the configuration of the points used to calculate the FAPs in [11] but adapted to the IEMOCAP maker configuration. The subset was chosen to include emotionally meaningful movements (e.g., eye squint, smile, etc.).

The feature set is composed of statistical functionals of the extracted audio and motion-capture features across various window lengths and include: mean, quantile minimum, quantile maximum, and quantile range for a total of 685 features. The statistical functionals were extracted only over speech regions. These regions were approximated by the first and last nonzero pitch value (extracted using Praat [12]). The functionals were extracted over windows of length 0.25, 0.5, 1, 1.5, and 2 seconds without overlap.

4.1. Feature Selection

The initial feature set included 685 features, too high of a dimensionality for the size of the considered datasets. The feature dimensionality was reduced using Principal Feature Analysis (PFA) [13]. PFA is an extension of PCA. PFA projects the original features into the PCA space. This space is then clustered using k-means. The features that are closest to the centers of the k clusters are returned as the final feature set instead of returning features that are linear combinations of the original space, resulting in features that are perceptually meaningful. Our previous work utilized PFA on the USC IEMOCAP data [8]. The final feature set contains 180 features (pilot results indicated the efficacy of retaining a large number of features for the EP calculation).

4.2. Data Handling

The data were split into four utterance length categories: 0.5-1.5 seconds, 1.5-3 seconds, 3-6 seconds, and over 6 seconds. This work will assess the utility of using a hierarchical static-dynamic framework for utterances with a majority ground truth of angry, happy, neutral, or sad over the four utterance lengths (Table 1).

5. METHODS

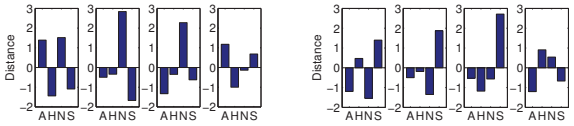
5.1. Static Modeling via Emotion Profiles

Emotion Profiles are extracted in a speaker independent fashion on a disjoint set of speakers. During training, the data are partitioned into the semantic clusters of angry, happy, neutral, and sad. Utterances not in these affective categories are not considered in the current work. The data in each semantic cluster is used to train one of the four binary Support Vector Machines (SVM) using the self vs. other paradigm. These classifiers will be used to characterize test utterances by estimating the presence or absence of the affective class in a given utterance. The SVMs utilized in this work are trained with Radial Basis Function kernels with a sigma of 9.5. The computation time is minimized using sequential minimal optimization (SMO) [14]. We have demonstrated that this method is effective for representing affective utterances [1].

During the testing stage, EPs are generated for a held-out speaker using the four binary classifiers. The binary classifiers are used to generate an EP trajectory over each test utterance using a sliding window of lengths: 0.25. 0.5, 1.0, 1.5, and 2.0 seconds with no overlap. Thus, each test utterance is represented with five sets of EPs. The sets all have different lengths (more EPs are extracted for shorter windows). These sets will be referred to as trajectories

Data Type	Angry	Happy	Neutral	Sad	Total
6 - inf	107	191	73	152	523
3 - 6	183	257	144	149	733
1.5 - 3	184	226	181	154	745
0.5 - 1.5	107	127	146	141	521

Table 1. The distribution of the emotion classes considered.



(a) Example trajectory for an angry utterance with a 2 second window.

(b) Example trajectory for an sad utterance with a 2 second window.

Fig. 2. Example EP trajectories for an angry and a sad utterance. Note that in both cases, neither emotion has a uniform expression under this representation (AHNS describe the angry, happy, neutral, and sad confidences).

(Figure 2). Utterances that are shorter than a given window length have a trajectory composed of a single EPs.

EPs are also trained for the set of training speakers using a leave-one-train-speaker-out paradigm. This creates training EP trajectories that do not incorporate data from the held-out test speaker.

5.2. Dynamic Modeling of Emotion Profiles

The second stage in the hierarchical static-dynamic classification is the dynamic HMM modeling. The HMMs model the relationship between the four-dimensional EP fluctuations and the high-level emotional ground truth. There are four HMMs trained, one for each emotion class (angry, happy, neutral, sad). Each HMM has three-states and two mixture components per state.

5.3. Comparison to Static and Dynamic Modeling

The results from the hierarchical static-dynamic framework will be compared to classification results using only static and only dynamic classification. In the static classification the EPs are generated using features extracted over the entirety of each utterance. The final emotion label is assigned using k-Nearest Neighbors (k = 45, determined empirically). This comparison will provide evidence supporting the need to incorporate the dynamics of emotion flow in an emotion classification framework.

In the dynamic classification, HMMs model the feature fluctuation of the utterances (three states, two mixtures per state). The features utilized in this modeling are the same statistical functionals used in the static-dynamic modeling (180 features). The dimensionality is high but was maintained to demonstrate the differences in results given the same starting data. These features are still extracted over windows ranging from 0.25 to 2 seconds but this time are not used to generate EPs. This study will demonstrate the importance of locally static modeling in the context of the dynamic tracking of emotion.

5.4. Validation

All systems presented in this work are speaker independent. In all cases both the training and testing EPs are are generated using a disjoint speaker set. The results are averaged over the ten speakers in the IEMOCAP dataset.

6. RESULTS

6.1. Static Modeling

The results demonstrate that as the length of an utterance increases, the classification accuracy of static modeling also increases (Table 2). This result suggests that in this modeling framework the emotion content of longer utterances is more easily modeled than that of shorter utterances. The classification results over utterance lengths of 6+ seconds, 3-6 seconds, 1.5-3 seconds, and 0.5-1.5 seconds are: 66.73%, 64.80%, 62.28%, and 61.23%, respectively. The accuracy over all sentence lengths is 63.72%.

Length	Number	Window	Static	Dynamic	Static-Dynamic
6 - inf	523	2.0	66.73	60.61	72.85
		1.5		61.76	74.57
		1.0		63.48	74.38
		0.5		59.66	70.75
		0.25		57.57	70.55
3 - 6	733	1.0	64.80	56.48	68.35
		0.5		57.71	67.97
		0.25		50.48	63.85
1.5 - 3	745	0.5	62.28	52.48	62.82
		0.25		53.96	60.94
0.5 - 1.5	521	0.25	61.23	54.70	62.96

Table 2. Accuracy of the static-dynamic modeling, static modeling, and dynamic modeling over utterances of four lengths.

6.2. Dynamic Modeling

The results of the dynamic classification also demonstrate that the accuracy of classification is higher for longer sentences than for shorter sentences (for sentences longer than 3 seconds). The maximal classification accuracies over utterance lengths of 6+ seconds, 3-6 seconds, 1.5-3 seconds, and 0.5-1.5 seconds are: 63.48%, 57.71%, 52.48%, and 54.70% respectively (Table 2). The classification accuracies over the 0.25 second window are: 57.57%, 50.48%, 53.96%, and 54.70%. The maximal accuracy over all sentence lengths is 56.74%. The accuracy over all sentence lengths for features extracted with the 0.25 second window is 53.85%.

6.3. Static-Dynamic Modeling

The static-dynamic classification results suggest that all utterances are more accurately modeled using the hierarchical static-dynamic modeling than either static or dynamic modeling (Table 2). Overall, longer utterances are more accurately classified than shorter utterances, again for utterances over 3 seconds. The maximal classification accuracies over utterance lengths of 6+ seconds, 3-6 seconds, 1.5-3 seconds, and 0.5-1.5 seconds are: 74.57%, 68.35%, 62.82%, 62.96%. The accuracies over windows of 0.25 are: 70.55%, 63.85%, 60.94%, and 62.96%, respectively (Table 2). The overall maximal accuracy is 66.89% (this is statistically significantly different from the static and dynamic modeling results, difference of proportions, $\alpha = 0.05$). The overall accuracy with a 0.25 second window is 64.20% (this is statistically significantly different from the dynamic modeling result, difference of proportions, $\alpha = 0.05$). The classification accuracy does not increase significantly for window lengths greater than 1 second (for utterances longer than 3 seconds). This result is similar to the findings in [2]; in general test data extracted at windows greater than 1 second in length did not provide additional performance gain (full saturation was at 2 seconds). There are many factors that influence the accuracy of this modeling. A subset of these factors include: affective characterization ability of the EP over the window lengths and the type of data considered. We will assess the contribution of both factors.

6.4. Factors Influencing Accuracy

The results demonstrate that longer utterances are more accurately modeled than shorter utterances using a static classification framework. This suggests that either EPs extracted over longer windows are more representative than EPs extracted over shorter windows or that EPs extracted from longer utterances are more representative than EPs extracted from shorter utterances. This will be tested by classifying the utterance-level tag using the EP trajectory components individually. The problem setup is as follows: for every EP trajectory component in every trajectory, estimate the utterance-level label of that trajectory using a k-Nearest Neighbor classifier (k = 45)

trained using individual EP trajectory components from a disjoint speaker set. Then, calculate the accuracy by merging the classification results of all the EP trajectory components estimated from each time window (e.g., all 2-second components, all 1-second components, etc.). The results are obtained using leave-one-speaker-out cross-validation. The accuracies of the individual EP trajectory component classification over components extracted from windows of 2 seconds, 1.5 seconds, 1 second, 0.5 seconds, and 0.25 seconds were 62.54%, 61.68%, 60.82%, 58.60%, and 56.18% respectively. This result suggests that data extracted from longer windows can be more accurately modeled than data extracted over shorter windows. However, these results are similar over all sentences lengths (Figure 3) suggesting that the length of the sentence does not strongly influence the representative nature of a segment of its data. Instead, the emotion content can be more accurately modeled given knowledge of the dynamics explained by the high-level emotion assessment trajectories. This suggests that the higher classification accuracy observed for the longer utterances cannot be explained only by the higher fidelity of the EPs extracted over longer windows. Instead, the advantage comes from the integration of the individual EPs over time. Further, although large segments are modeled more accurately than small segments (Figure 3), this accuracy does not lead to higher classification results when the trajectories are dynamically modeled (Table 2). This suggests that, for the sentence length considered, there is not a benefit to utilizing windows that are longer than 2 seconds.

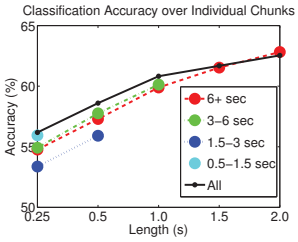


Fig. 3. The results of individual EP trajectory component classification over various time windows.

The enhanced accuracy of the longer utterances in the hierarchical static-dynamic system can also not be solely explained by the clarity of the affective information. The proportion of prototypical (data with complete evaluator agreement) and nonprototypical (data with only a majority-vote evaluator agreement) differs between the utterance length groups; as the length of an utterance decreases, the percentage of utterances without total evaluator agreement increases (Table 3). This may explain the differences in the accuracies of the static modeling. However, it does not explain the enhanced performance gain when classifying with the hierarchical static-dynamic system, seen in the longer utterances (utterances greater than 3 seconds). This suggests that although the emotional clarity of the utterances may affect a baseline classification, it cannot fully explain the added performance of the proposed system.

Length	6 - inf	3 - 6	1.5 - 3	0.5 - 1.5
Prototypical	58.13	54.30	49.93	44.34
Nonprototypical	41.87	45.70	50.07	55.66

Table 3. The evaluator agreement across the four sentence lengths and emotion clarity classes.

7. CONCLUSIONS

This paper presented a novel hierarchical static-dynamic emotion classification framework. The results were presented across four different sentence lengths and in all contexts excepting the dynamic modeling of utterances less than 3 seconds, longer utterances were more accurately classified than shorter utterances. Additionally. the performance gain between static modeling, dynamic modeling and static-dynamic modeling was higher for the longer sentences. This result suggests that when modeling longer data it may be beneficial to estimate higher-level emotional fluctuations.

The enhanced performance of the static-dynamic system, when compared to the static classification is encouraging. In human-machine interaction it may not be possible to work with fully segmented data. In these cases static modeling will not be possible because the boundaries of the utterances are not known. This paper demonstrates that local context can be integrated with global dynamics to either meet or exceed the classification accuracy of the static system. This suggests that such a method should be investigated for online emotion recognition.

One of the limitations of this work is its consideration of a single database. In future work it will be beneficial to apply these techniques to additional data collected in alternative settings. This will provide further evidence of the efficacy of this technique.

The enhanced accuracy of the classification of the longer utterances suggests that the system may be approximating an emotional grammar, which describes how emotions change over time within single utterances. Future work will investigate the presence of such a grammar and how it can be modeled using a combination of static and dynamic modeling. Such a finding would provide performance gains for both emotion classification and emotion recognition.

8. ACKNOWLEDGEMENTS

This research was supported in part by funds from the National Science Foundation and the United States Army.

References

[1] 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*, Accepted for Publication.

[2] B. Schuller and L. Devillers, "Incremental acoustic valence recognition: an inter-corpus perspective on features, matching, and performance in a gating paradigm," in *InterSpeech*, Makuhari, Japan, Sept. 2010, pp. 801–804.

[3] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *InterSpeech*, Makuhari, Japan, Sept. 2010, pp. 2362–2365.

[4] I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang, "Facial expression recognition from video sequences: Temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 160–187, 2003.

[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] 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 CVPR*, vol. 2, pp. 568–573, 2005.

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

[8] E. Mower, M. Matarić, and S. Narayanan, "Robust representations for out-of-domain emotions using emotion profiles," in *SLT*, Berkeley, California, Dec. 2010.

[9] F. Eyben, M. Wöllmer, and B. Schuller, "openEAR - introducing the munich open-source emotion and affect recognition toolkit," in *ACII*, Amsterdam, The Netherlands, Sept. 2009.

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

[11] 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.

[12] P. P. G. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

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

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