Emotion Spotting: Discovering Regions of Evidence in Audio-Visual Emotion Expressions

Yelin Kim
Department of Computer Engineering
University at Albany, SUNY
1400 Washington Avenue
Albany, NY, USA
yelinkim@albany.edu

Emily Mower Provost
Department of Computer Science
University of Michigan
2260 Hayward Street
Ann Arbor, MI, USA
emilykmp@umich.edu

ABSTRACT

Research has demonstrated that humans require different amounts of information, over time, to accurately perceive emotion expressions. This varies as a function of emotion classes. For example, recognition of happiness requires a longer stimulus than recognition of anger. However, previous automatic emotion recognition systems have often overlooked these differences. In this work, we propose a data-driven framework to explore patterns (timings and durations) of emotion evidence, specific to individual emotion classes. Further, we demonstrate that these patterns vary as a function of which modality (lower face, upper face, or speech) is examined, and consistent patterns emerge across different folds of experiments. We also show similar patterns across emotional corpora (IEMOCAP and MSP-IMPROV). In addition, we show that our proposed method, which uses only a portion of the data (59% for the IEMOCAP), achieves comparable accuracy to a system that uses all of the data within each utterance. Our method has a higher accuracy when compared to a baseline method that randomly chooses a portion of the data. We show that the performance gain of the method is mostly from prototypical emotion expressions (defined as expressions with rater consensus). The innovation in this study comes from its understanding of how multimodal cues reveal emotion over time.

CCS Concepts

•**Information systems** → **Multimedia and multimodal retrieval;** *Specialized information retrieval;*

Keywords

Audio-Visual; Emotion; Temporal Evidence; Emotion Classification; Emotion Spotting

ICMI'16, November 12–16, 2016, Tokyo, Japan
© 2016 ACM. 978-1-4503-4556-9/16/11...$15.00
http://dx.doi.org/10.1145/2993148.2993151

1. INTRODUCTION

Audio-visual emotion recognition systems play a pivotal role in natural and human-centered interactive technology [8, 14, 16, 32]. These systems use the audio-visual data, such as facial movements and vocal changes, to infer emotions. The study of emotion recognition has grown rapidly within the field of multimodal interaction, often focusing on how to combine the emotional information from audio and visual modalities. However, there has been less exploration on how multiple modalities unfold emotion over time and whether partial information may be sufficient for inferring human emotion [12, 13, 34]. For example, a brief, genuine smile may cause interaction partners to perceive happiness even though one person is showing a neutral expression for most of the time [15]. Likewise, a sudden burst of anger can be a significant indicator of underlying anger [12].

The basic assumption behind previous emotion recognition systems is that human emotions are expressed simultaneously with the same duration across multiple modalities [9, 26, 31]. Previous systems have often overlooked the modality-specific temporal characteristics. A complete understanding of these characteristics may allow us to process only the relevant subsets of each modality, rather than all the presented information. In this study, we explore regions within an utterance that contain emotion evidence, that vary for the lower face, upper face and for speech. Throughout this paper, an utterance is defined as a region where a person is actively speaking and is predefined in the emotion datasets we use. We focus on the timings and durations of these regions, which we call 'temporal patterns'. We aim to investigate three important research questions, which we list in the following paragraphs.

The first research question pertains to generalizability and subject-independence of the temporal patterns of emotion. Previous studies have explored the relationship between multiple modalities, however they either neglected temporal patterns [40] or generalizability across multiple human subjects [5]. Human perception studies have found that different durations are required to correctly recognize emotion [33]. This raises the main research question: **(Q1)** Are there consistent temporal patterns of emotion expressions, across subjects, in the lower and upper face and in speech?

We evaluate the efficacy of the temporal patterns in audio-visual emotion recognition systems. We are interested in the following two research questions: **(Q2)** Can we achieve similar accuracy to a traditional method that uses all the data within an utterance ('all-mean method') but using less data

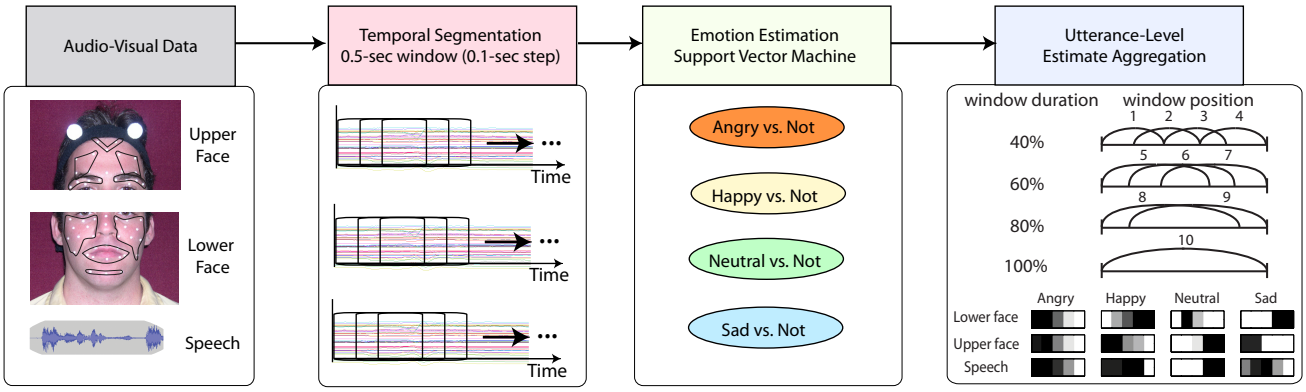

Figure 1: Overview of the proposed method. We first segment lower face, upper face, and speech modalities using fixed-length windows and calculate segment-level emotion estimates using SVMs. We then aggregate the segment-level emotion estimates with different temporal window configurations (Index 1–10).

and with a higher accuracy than a baseline that randomly selects windows? and **(Q3)** What types of emotion expressions are associated with consistent emotion patterns?

In this study, we address the three research questions by proposing a data-driven approach to find consistent temporal patterns of emotion in the lower face, upper face and in speech, varying for four classes of emotion (anger, happiness, neutrality and sadness). Our approach identifies temporal regions within an utterance that lead to the highest emotion recognition rate, varying for different modalities and for different classes of emotion. Figure 1 shows the overview of our system. Firstly, our method segments each of the three modalities using fixed-length windows and estimates the emotion content in each of the segments. Secondly, we create sets of window configurations that cover different regions in time of the utterance (e.g. the beginning vs. the middle) and duration (e.g. 40% vs. 60%). Then finally, we classify the utterance based on each set of window configuration in order to identify the optimal parameters (timing and duration). We compare our proposed methods to two baseline methods: the first uses the all-mean method and the second uses partial data within an utterance, where these regions are randomly selected.

The key innovation of this study comes from our investigation of the three research questions. The first set of experiments using the IEMOCAP dataset [3] demonstrates that there are consistent temporal patterns in the timing and duration that are unique to the upper face, the lower face and speech modalities, respectively. These temporal patterns show similarity, over different folds of experiments, within the same class of emotion and modality. Our proposed system achieves a similar accuracy to traditional systems that use all the available data to infer emotion, while ours uses only 40–80% of the data. It also significantly outperforms the baseline method that uses random temporal regions within an utterance. We also test our system on a second dataset; the MSP-IMPROV dataset [6]. This dataset includes video data instead of the motion-capture data of the IEMOCAP dataset. The findings of our work provide insight into how lower face and upper face and speech modalities reveal emotional evidence over different timings and durations.

2. RELATED WORK

Humans require different amounts of information, over time, to recognize different classes of emotion such as anger, happiness and fear [11, 33]. Pell and Kotz studied human emotion recognition in speech, focusing on how quickly listeners could recognize emotions from five basic classes (anger, disgust, fear, sadness and happiness) [33]. They found that the class "happy" requires more information compared to other classes of emotion such as anger, sadness or fear as well as neutral expressions. Edwards demonstrated that humans are capable of reliably detecting temporal variations of facial movements, even extremely subtle ones [11]. The study also showed that humans assess emotions more accurately at the early stages of an expression.

The field of automatic emotion recognition has a relatively limited number of studies that can provide interpretable descriptions of the timings and durations of emotion evidence. Earlier studies in Action Unit (AU) detection have attempted to model the onset, apex and offset of expressions [36,37]. However, our work differs in that the proposed detection is (i) based on emotion evidence instead of AU intensity and hence, (ii) we do not have explicit boundaries within an utterance (e.g. eyebrow raising). Instead, our proposed approach detects sub-utterance regions automatically, using emotional cues. Studies in the field have found that considering the temporal dynamics of multimodal emotional expressions is useful for improving the performance in emotion classification [10,17,40]. However, an open question remains: can we find temporal patterns of emotional expressions in different modalities?

Previous studies on multimodal emotion expressions have long recognized the importance of understanding these patterns. A recent study by Wagner et al. highlighted this importance, stating that "significant hints from different modalities are not guaranteed to emerge at exactly the same time interval" [40]. The authors explored how to handle missing data in multimodal emotion expressions, however they did not explore the temporal information within these expressions. They suggested future work could be done to improve the accuracy of emotion classification by segmenting different modalities separately and by considering the delays between "occurrences of emotional hints." Amer et al.

also studied the temporal aspects of multimodal emotion expressions and proposed a conditional restricted Boltzmann machine to jointly model multiple modalities [1].

A recent study by Mansoorizadeh et al. attempted to account for asynchrony in multiple modalities [22]. They studied feature-level fusion methods between the upper and lower face and speech signals. They proposed a buffered method to deal with asynchrony across modalities, where they either repeated the last frame value or took the median before fusing these modalities, i.e., filling in the missing values in the data. Their experimental results showed that an asynchronous feature level fusion approach significantly outperforms a synchronous fusion method. However, open questions remain as to whether we can find patterns in the timings across different folds of experiments and leverage these patterns to classify emotion.

3. DATASETS

3.1 IEMOCAP

The IEMOCAP dataset, an established dataset in the field of emotion recognition [3], includes audio, video and three-dimensional motion capture recordings of dyadic interactions between ten speakers who were given both improvised and scripted prompts.

The facial motion-capture markers are tracked at 120 frames per second. We group the markers into the upper and lower regions of the face, similar to prior work [25]. The upper facial region includes the forehead, upper eyebrows and eyebrows. The lower facial region includes the cheeks, mouth and chin. There are three markers on the forehead, eight markers on the upper eyebrow region and eight markers on the eyebrow region. There are 16 markers on the cheek regions, eight markers on the mouth and three markers on the chin. We use a subset of the 46 motion capture markers (Figure 1) to be consistent with previous work [18, 19]. The audio is recorded with a sample rate of 48kHz.

The recordings are segmented at utterance level or a turn when a speaker is actively speaking. The average duration of an utterance is 4.72 ± 3.36 seconds. Each utterance is annotated by at least three human annotators. Some of the data have no majority vote agreement. As in previous work [26, 27], we neglect these data and use only utterances with majority vote agreement for the classes of angry, happy, neutral and sad. The mean and standard deviation of number of utterances over the ten speakers for each class are: 61.00 ± 27.30 for angry, 122.60 ± 25.11 for happy, 58.00 ± 21.62 for neutral, and 63.80 ± 23.56 for sad.

3.2 MSP-IMPROV

We also test our system on the MSP-IMPROV [6]. This dataset contains 12 speakers in total (six males and six females), talking in pairs. The difference between the MSP-IMPROV dataset and the IEMO-CAP dataset is that it contains constrained improvisational sessions, where a speaker is asked to speak a specific sentence ("target sentence") in each session. There are four sub-sets in the dataset: (i) target-improvised: target sentences during improvised sessions, (ii) target-read: a read speech version of the target sentences, (iii) other-improvised: utterances in the improvisational sessions other than the target sentences, and (iv) natural interaction: utterances made during the breaks between the improvisational sessions (i.e., while the actors are not acting). We use a subset of the dataset that is labeled as anger, happiness, neutrality and sadness, to be consistent with the IEMOCAP dataset. We use 7,323 utterances, in total, which includes 2,273 natural interaction utterances, 578 target-read utterances, 3,836 other-improvised utterances and 636 target-improvised utterances. The mean and standard deviation of the number of utterances of each emotion class over the twelve speakers are: 179.58 ± 42.73 angry, 163.00 ± 40.27 happy, 119.50 ± 29.43 neutral, and 148.17 ± 37.76 sad utterances. There are 2,501 prototypical utterances and 5,937 non-prototypical utterances.

We extract Action Unit (AU) features using CERT [20] (motion capture data was not collected). We use the same audio feature set, pitch, energy and MFB features, as in the IEMOCAP dataset. This results in a smaller feature dimensionality in the MSP-IMRPOV dataset: 8 AU features for the upper face and 14 AU features for the lower face at the frame level, and 64 features for the upper face and 112 features for the lower face regions at the segment level.

4. PROPOSED SYSTEM

Our system is composed of four main modules: feature extraction, segment-level emotion estimation, window-based averaging and final emotion classification. We first extract audio-visual features from the lower and upper face as well as speech modalities, using the features that have shown to be effective in previous emotion recognition studies [4,27,38,39]. Next, we segment the audio-visual features into fixed-length windows and estimate segment-level emotions, using Support Vector Machines (SVMs). Then, we apply various window configurations with different window timings and durations over the segment-level emotion estimates, to discover the best window configuration for each modality and each class of emotion. Finally, we use the emotion estimates that are aggregated by the identified windows to infer the utterance-level emotional class.

To train and test our proposed system, we employ leave-one-speaker-out cross-validation. Since the IEMOCAP data includes ten speakers in total, we conduct ten-fold experiments. In each of the ten experiments, we use nine speakers to "train" the system and a left-out speaker to test the emotion classification performance of the trained system. In order to choose the best-performing window configuration, we do leave-one-speaker-out cross-validation on each of the nine training speakers. This means that for each of the ten speakers, we compute the validation accuracy of nine training speakers, when eight speakers are used to train the system and a left-out training speaker is used for validation.

4.1 Feature Extraction

The feature used in this study are divided into three modalities: (1) lower face, (2) upper face, and (3) speech.

The lower face includes three facial regions (the chin, mouth and cheeks); the upper face includes forehead, eyebrows and the upper eyebrow area. We extract the (x, y, z)-coordinates of the motion capture features to track the movements of the lower and upper face. The chosen origin is the tip of the nose and the facial features are rotated to compensate for head rotation. We pre-process the markers by first translating to make the tip of the nose the starting point and then rotating to compensate for head movement. In order to reduce any subject variations in facial configurations, we also mean-normalize each marker position, by

making the mean of each dimension of marker position per subject the global mean over all subjects, as in previous work [18, 19, 24]. We also exclude segments with less than seven frames (0.058 seconds) because of insufficient temporal information, as in [18, 19].

The speech features contain spectral and prosodic features that have been shown to be useful in emotion recognition [27]. This includes pitch, energy, and mel-filterbank coefficients (MFBs), extracted using Praat [2]. This results in 29-dimensional speech feature vectors.

4.2 Segment-Level Emotion Estimation

4.2.1 Temporal Segmentation

Based on findings of previous work [19], we choose to use 0.5-second windows, moving with 0.1-second time steps. We use all of fixed-length windows, including segments at the end of an utterance that are shorter than 0.5 seconds. The segment-level features are computed using the mean standard deviation, upper quantile, lower quantile, quantile range and 3-degree polynomial regression coefficients within each segment. This results in 648 features for the lower face, 456 features for the upper face and 232 features for speech.

4.2.2 Emotion Estimation

We estimate segment-level emotion evidence using the Emotion Profile (EP) technique proposed by Mower et al. [27, 29]. We first train four binary emotion classifiers, using the utterance-level data of the lower and upper face and speech, for the classes of anger, happiness, neutrality and sadness. Each classifier is a radial basis function kernel Support Vector Machine (SVM), where the soft margin parameter c is chosen as 1 as in [6, 23]. We set the gamma in the kernel function as the inverse of the number of input features, in order to be consistent with the default value as suggested in [7].

For each test utterance, we use the segmented data of the utterance as an input to the trained emotion classifiers. The output of each classifier is a signed distance to the hyperplane, as in previous work [27, 29]. We crease segment-level EPs by combining the distances into a single vector. Finally, we convert these SVM outputs into probabilistic values, by first applying z-normalization for each test fold and by taking a sigmoid function. Then we normalize the values of each emotion component, so that the sum of the four emotion components sums to 1.

4.3 Window-Based Averaging

A traditional method of aggregating segment-level emotion estimates is to take the mean over all the segment outputs within an utterance [19]. In this paper, we argue that salient information is embedded within an utterance. Thus, we only calculate mean values over regions within an utterance. We investigate which regions and durations of the EP are useful for final emotion inference across the upper face, lower face, and speech.

The last module of Figure 1 shows ten different configurations with different window durations and positions, each denoted as indices 1 to 10. We use cross-validation over each training fold to choose one of these ten window configurations for each modality, individually. The indices are as follows:

- Index 1–4: divide an utterance into four regions, each with 40% duration of an utterance
- Index 5–7: divide an utterance into three regions, each with 60% duration of an utterance
- Index 8, 9 : divide an utterance into two regions, each with 80% duration of an utterance
- Index 10 : all data

We compute validation accuracy of per-angry, per-happy, per-neutral and per-sad emotion classes, when one of the three modalities (lower face, upper face or speech) is used in classification.

4.4 Emotion Classification

We use the chosen window configurations to estimate the utterance-level emotion estimates, from segment-level emotion estimates. We have four emotion estimates: angry, happy, neutral and sad, for each of the three modalities: lower face, upper face and speech. For each modality and emotion pair, we take the average of segment-level estimates individually within the region of an utterance, based on the chosen window configuration. For instance, if the cross-validated parameters for a test speaker (Figure 1) are 5, 1, and 8 for an angry classification, using lower face, upper face and speech, then we take a mean of the angry components of EPs using 60% of the beginning of an utterance, 40% of the beginning of an utterance, and 80% of the beginning of an utterance for the test speaker for the segment-level emotion estimates from the lower face, upper face, and speech modalities, respectively.

After we apply different window configurations for each modality and each emotion component, we take an average over the different modalities to get a four-dimensional EP. Each dimension of the EPs is the averaged emotion component of angry, happy, neutral and sad classes. As in [27], we choose the final emotion label that is the maximum component among the four emotion components. For instance, if the outputs of angry, happy, neutral, and sad classifiers are $[0.17, -0.63, -0.23, 0.80]$, then we choose sadness as the emotion label of the test data.

5. RESULTS AND DISCUSSIONS

We design and perform experiments to address our three research questions (Q1-Q3).

To address Q1, we first investigate the chosen window timings and durations across ten training folds (Section 5.1). We show that consistent patterns exist that vary depending on the different classes of emotion and the individual modalities. We provide insights into how these temporal patterns match the findings of human perceptual studies.

To address Q2, we evaluate the performance of our proposed audio-visual emotion recognition systems by comparing it with a baseline method. The baseline method selects temporal regions using a uniformly distributed randomization of the ten different window configurations. We randomly select the parameters for each emotion component, and for each modality, of the lower and upper face and speech. We make 50 runs to obtain our classification results. We then calculate an average of these 50 runs, for each speaker, in order to compare our proposed method with the baseline.

We compare the performance of our system with a traditional system that uses all the information within an utterance: the all-mean method. This method takes a mean of the evidence of emotion within an utterance to infer the final emotion.

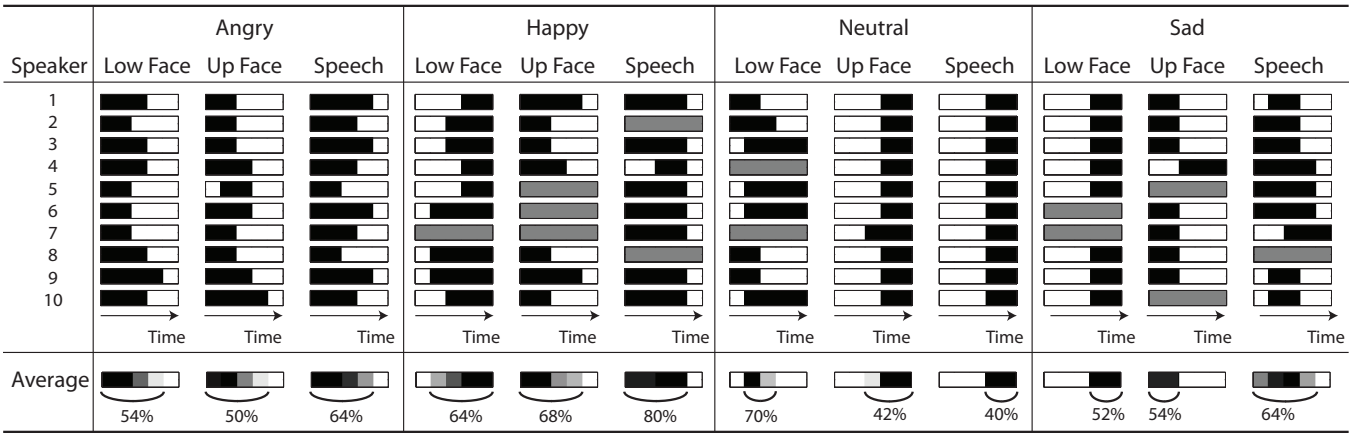

Figure 2: Temporal window configurations for each speaker, for individual modalities (LowOpt: lower face, UpOpt: upper face, AudOpt: speech) and for each emotion component (Angry, Happy, Neutral, Sad). The last row is an averaged window configuration over ten speakers. For each speaker, black regions are the chosen regions used for emotion classification. The darker regions in the last row show overlapping windows from the ten speakers. We also show the average percentage of an utterance used over the ten speakers.

To address Q3, we compare the performance gain of our system based on the inter-rater agreement of each utterance. We divide the utterances into prototypical (defined as rater consensus) and non-prototypical (defined as no rater consensus, but in the presence of majority vote) utterances.

The performance measurement of our system is unweighted (UW) recall, to be consistent with previous work [23,30]. We perform paired t-tests to test the significance of comparisons of accuracy between our proposed method and the baseline, and between our proposed method and the all-mean method. This paired t-test has been used previously, in an emotion recognition on the IEMOCAP data [19].

5.1 Temporal Evidence Analysis

The chosen window configurations show that there is consistency, across speakers, with respect to the timings and durations of emotion (Figure 2). The first ten rows show the chosen window for individual modalities and for each class of emotion, while the last row shows an average over the ten speakers. For each speaker, the chosen regions are represented in black and for the averaged regions, the darker ones represent the overlapping regions of the ten speakers. The areas are consistent across multiple folds with training speakers for different modalities and classes of emotion.

The table and figure show that the lower facial region is consistently chosen, for anger–at the beginning of an utterance, for happiness at the end and for sadness generally at the end. The neutral class is a mixture of different window configurations. This finding agrees with the historical difficulties experienced in classifying neutrality [21, 28, 35]. The upper face is generally chosen at the beginning of an utterance for anger, happiness and sadness while neutrality uses information from the end of an utterance. As shown in the table, speech requires longer durations to identify the classes of emotion for angry, happy and sad. The three classes of emotions are chosen at the beginning of an utterance in cross-validation. Neutrality is chosen at the end of an utterance.

Table 2: IEMOCAP experimental results on the proposed, all-mean, and baseline methods. The accuracies are unweighted recall ('UW'), per-class accuracy for angry ('A'), happy ('H'), neutral ('N'), and sad ('S') emotion classes, and weighted accuracy ('W').

Method	UW	A	H	N	S	W
Proposed	**65.60**	72.88	72.02	40.96	76.53	66.38
All-mean	**65.59**	71.05	73.99	38.15	79.16	66.92
Baseline	**63.70**	68.49	73.10	36.91	76.29	64.87

The window configurations also reveal similar findings as those from previous studies. In a recent speech emotion recognition study [33], the authors found that angry, sad, fearful and neutral emotion expressions are more accurately recognized when given shorter data, compared to happy emotion expressions. As shown in Figure 2, the percentage of an utterance used for recognizing happiness is 80% for speech, which is higher than the other classes of emotion: 64% for anger, 40% for neutral, and 64% for sad classes of emotion. This finding may indicate that our proposed data-driven window configurations can provide insight into how humans perceive emotion expressed over time.

5.2 Evaluation of Emotion Recognition

We compare the performance to our baseline, where windows are chosen at random. We also compare our results to a method that uses all the segment-level evidence within an utterance, instead of partial information, as in our proposed method. The results in this section will help us to answer Q2. Overall, our proposed method significantly outperforms the baseline method and it achieves a comparable accuracy to the all-mean method.

Table 1 shows the UW recall of unimodal and multimodal experiments for our proposed window method, the all-mean method and the baseline method, using randomized window configurations. We test different combinations of modali-

Table 1: Unweighted recall of unimodal and multimodal experiments for our proposed window method, the all-mean method, and the baseline method using randomized window configurations. The symbol "[*]" indicates statistical significance levels ($p < 0.05$) between our proposed method and baseline. All the results between the proposed method and the all-mean method are statistically comparable to each other ($p > 0.05$).

	L.F+U.F+Aud	L.F	U.F	Aud	L.F+Aud	U.F+Aud	L.F+U.F
Proposed	65.60[*]	59.76[*]	52.85[*]	54.56	64.50	61.24	60.19[*]
all-mean	65.59	60.65	52.62	55.47	65.83	62.08	60.57
Baseline	63.70	57.92	51.21	53.41	62.79	60.02	58.78

ties and each column represents all the modalities that combine the lower face (LF), the upper face (UF) and speech (Aud), the lower face, the upper face, the audio, the lower face with audio, the upper face with audio, and the lower and upper face. For the IEMOCAP, our proposed window-mean method achieves 65.60% UW and the all-mean method achieves 65.59% UW. The difference is not significant. Our proposed method is significantly higher than the average UW accuracy of the randomized window method, achieving 65.60% vs. 63.70%, respectively (1.90% higher, $p < 0.05$), when all modalities are used.

Our proposed method also outperforms the baseline for all types of modality combinations. For the lower face, our proposed method achieves a 1.84% higher UW accuracy, significantly improving the baseline ($p < 0.05$). The upper face region also significantly outperforms the baseline, achieving a 1.63% higher UW accuracy ($p < 0.05$). Speech is also higher when we use our proposed method, but not significantly (1.14% improvement, $p = 0.06$). Speech with the lower face and speech with the upper face both achieve higher UW recall than the baseline, but this difference is not significant (1.71% improvement with $p = 0.06$ and 1.23% improvement with $p = 0.08$, respectively). The lower and upper face combination achieves significantly higher performance, showing 1.42% improvement with $p < 0.05$.

The bottom row of Figure 2 shows the percentage of an utterance used for each emotion class, averaged over ten speakers. The results indicate that the system only uses 40% to 80% of an utterance. This highlights the benefit of our system, as it is capable of spotting a region within an utterance, and reasoning only over that region, while achieving comparable accuracy with an experiment that uses the full information of an utterance. The results also demonstrate that, on average, speech requires more regions of an utterance than lower and upper face regions for emotion classes, i.e., angry, happy and sad, while the lower face is used more for neutral recognition with the IEMOCAP data.

Table 2 shows the comparison of the results of emotion classification between our proposed window method, the all-mean method and the baseline using randomized window configurations, where all modalities are used. As shown in Table 1, the use of all modalities shows the highest accuracy compared to pairs of modalities. Each column on the tables shows the UW recall, per-emotion class accuracy for angry, happy, neutral and sad classes of emotion and weighted accuracy.

Finally, to address Q3, we investigate the performance of our proposed system compared to the all-mean method. The performance gain of our system is mostly from the prototypical expressions (defined as rater consensus). The results are presented in Table 3. For prototypical utterances, our method gets 77.49% compared to all-mean's 75.84% (1.65% difference. p-value=0.53). For non-prototypical utterances (defined as no rater consensus), our method gets 57.29% compared to all-mean's 57.33% (0.05% difference, p-value = 0.93). This may indicate that the temporal patterns of emotion and emotion spotting are more useful when the expression is more explicit and prototypical to human evaluators.

Table 3: IEMOCAP dataset: Emotion classification unweighted recall (%) for prototypical and non-prototypical utterances. The accuracies are unweighted recall ('UW'); and per-class accuracy for angry ('A'), happy ('H'), neutral ('N'), and sad ('S') emotion classes.

Type	Method	UW	A	H	N	S
Prot	Proposed	77.49	83.50	81.65	62.73	80.66
	All-mean	75.85	80.77	83.47	53.26	83.26
Non-prot	Proposed	57.29	64.15	56.54	36.15	72.33
	All-mean	57.33	62.23	58.60	33.56	74.95

5.3 Validation With MSP-IMPROV Data

We repeat the IEMOCAP experiments on the MSP-IMRPOV dataset. Figure 3 shows the per-fold and averaged window configurations chosen for the MSP-IMPROV dataset. As seen in the IEMOCAP dataset, there are consistent patterns across multiple training folds. In addition, the patterns are similar to the IEMOCAP dataset. These similar patterns include: the lower face and speech modalities of anger, the lower and upper face, and speech modalities of happiness, the upper face and speech of neutrality and the lower face modality of sadness.

The accuracy of emotional classification, using our proposed window method, shows a slightly higher accuracy than the baseline method (ours: 56.30%, baseline: 56.21%, not significantly different), and a lower accuracy than the all-mean method over the four subsets together (all-mean: 58.07%, p-value<0.05). For subset-specific classification results, the difference between our proposed method and the all-mean method are not significantly different, and our method achieves a significantly higher accuracy than the baseline for the target-improvised subset and is not significantly different for the other subsets:

- Target-improvised: ours (63.32%), all-mean (61.06%), baseline (59.60%)
- Target-read: ours (49.69%), all-mean (50.85%), baseline (49.47%)
- Other-improvised: ours (53.42%), all-mean (55.05%), baseline (53.47%)
- Natural Interaction: ours (46.81%), all-mean (49.00%), baseline (46.30%)

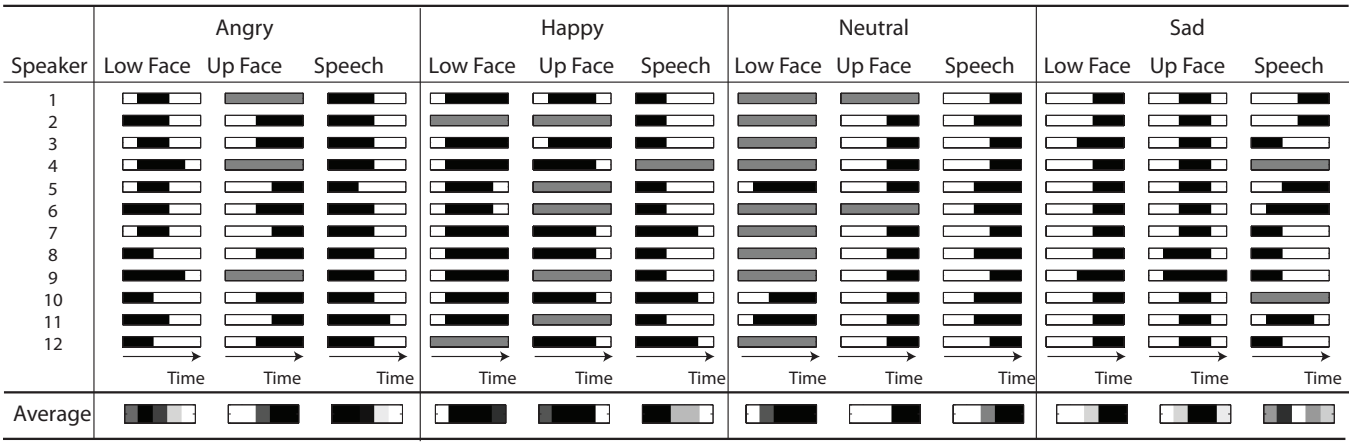

Figure 3: Temporal window configurations for each speaker, for individual modalities (LowFace, Up Face, Speech) and for each emotion component (Angry, Happy, Neutral, Sad) of the MSP-IMPROV dataset. The last row is an averaged window configurations over ten speakers. For each speaker, black regions are the chosen regions used for emotion classification, and the darker regions in the last row show overlapping windows over the 12 speakers.

Table 4: MSP-IMPROV dataset: Emotion classification unweighted recall (%) for prototypical and non-prototypical utterances. The accuracies are unweighted recall ('UW'); and per-class accuracy for angry ('A'), happy ('H'), neutral ('N'), and sad ('S') emotion classes.

Type	Method	UW	A	H	N	S
Prot	Proposed	64.54	60.77	86.23	59.61	51.56
	All-mean	64.45	60.72	87.36	61.36	48.35
Non-prot	Proposed	52.29	43.58	73.71	48.70	43.19
	All-mean	54.50	48.46	75.56	49.31	44.67

The findings from the MSP-IMPROV dataset align with the Q1 from the IEMOCAP dataset. We show that there are consistent patterns across speakers. The findings for Q2 are mixed. Our proposed method achieves a slightly higher accuracy compared to the baseline, however, it is lower than the all-mean method. The findings on Q3 (Table 4) on the MSP-IMPROV dataset suggest that non-prototypical expressions, often subtle and ambiguous to human evaluators, make it difficult to identify the patterns in windows that lead to higher emotion recognition rate.

6. CONCLUSIONS

In this work, we explore whether a subset of an utterance can be used for emotion inference and how the sub-set varies by classes of emotion and modalities. We propose a windowing method that identifies window configurations, window duration and timing, for aggregating segment-level information for utterance-level emotion inference. The experimental results using the IEMOCAP and MSP-IMPROV datasets show that the identified temporal window configurations demonstrate consistent patterns across speakers, specific to different classes of emotion and modalities.

We compare our proposed windowing method to a baseline method that randomly selects window configurations and a traditional all-mean method that uses the full information within an utterance. Our proposed method shows a significantly higher performance in emotion recognition than the baseline method, achieving a 65.60% UW accuracy, 1.90% higher than the baseline). Our method also achieves similar performance to the traditional all-mean method (65.59%, statistically insignificant difference), while our method only uses 40–80% of information within each utterance.

The identified windows also show consistency across speakers, demonstrating how multimodal cues reveal emotion over time. We demonstrate that these patterns also align with psychological findings. The innovation that our work brings is that we discover consistent temporal patterns of emotion expressions across subjects in the lower face and upper face and speech modalities.

7. ACKNOWLEDGEMENT

This material is based in part upon work supported by IBM under contract 4915012629. Any opinions, findings, conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of IBM.

8. REFERENCES

[1] M. R. Amer, B. Siddiquie, S. Khan, A. Divakaran, and H. Sawhney. Multimodal fusion using dynamic hybrid models. In *IEEE Winter Conference on Applications of Computer Vision*, pages 556–563. IEEE, 2014.

[2] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 6.0.17)[computer program]. retrieved 21 april 2016 from http://www.praat.org/.

[3] 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. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[4] C. Busso, S. Lee, and S. S. Narayanan. Using neutral speech models for emotional speech analysis. In *Interspeech*, pages 2225–2228, 2007.

[5] C. Busso and S. S. Narayanan. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Transactions on Audio, Speech, and Language Processing*, pages 2331–2347, 2007.

[6] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 2015.

[7] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction. *Signal Processing Magazine, IEEE*, 18(1):32–80, 2001.

[9] E. Cvejic, J. Kim, and C. Davis. Temporal relationship between auditory and visual prosodic cues. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[10] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474. ACM, 2015.

[11] K. Edwards. The face of time: Temporal cues in facial expressions of emotion. *Psychological science*, 9(4):270–276, 1998.

[12] A. Freitas-Magalhães. Microexpression and macroexpression. *Encyclopedia of human behavior*, 2:173–183, 2012.

[13] E. A. Haggard and K. S. Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy. In *Methods of research in psychotherapy*, pages 154–165. Springer, 1966.

[14] M. G. Helander. *Handbook of human-computer interaction*. Elsevier, 2014.

[15] A. Ito, X. Wang, M. Suzuki, and S. Makino. Smile and laughter recognition using speech processing and face recognition from conversation video. In *Cyberworlds. International Conference on*, pages 8–pp. IEEE, 2005.

[16] J. A. Jacko. *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. CRC press, 2012.

[17] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels. Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 9–16. ACM, 2015.

[18] Y. Kim and E. Mower Provost. Say cheese vs. smile: Reducing speech-related variability for facial emotion recognition. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 27–36. ACM, 2014.

[19] Y. Kim and E. Mower Provost. Emotion recognition during speech using dynamics of multiple regions of the face. *ACM Trans. Multimedia Comput. Commun. Appl.*, 12(1s):25:1–25:23, Oct. 2015.

[20] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG)*, pages 298–305, 2011.

[21] M. Lugger, M.-E. Janoir, and B. Yang. Combining classifiers with diverse feature sets for robust speaker independent emotion recognition. In *European Signal Processing Conference*, pages 1225–1229. IEEE, 2009.

[22] M. Mansoorizadeh and N. M. Charkari. Multimodal information fusion application to human emotion recognition from face and speech. *Multimedia Tools and Applications*, 49(2):277–297, 2010.

[23] S. Mariooryad and C. Busso. Feature and model level compensation of lexical content for facial emotion recognition. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*.

[24] A. Metallinou, C. Busso, S. Lee, and S. Narayanan. Visual emotion recognition using compact facial representations and viseme information. In *ICASSP*.

[25] A. Metallinou, S. Lee, and S. Narayanan. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *International Conference on Acoustics Speech and Signal Processing*, pages 2462–2465. IEEE, 2010.

[26] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan. Context-sensitive learning for enhanced audiovisual emotion classification. *Affective Computing, IEEE Transactions on*, 3(2):184–198, 2012.

[27] E. Mower, M. J. Matarić, and S. Narayanan. A framework for automatic human emotion classification using emotion profiles. *Transactions on Audio, Speech, and Language Processing*.

[28] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan. Interpreting ambiguous emotional expressions. In *International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–8. IEEE, 2009.

[29] E. Mower and S. Narayanan. A hierarchical static-dynamic framework for emotion classification. In *ICASSP*, pages 2372–2375. IEEE, 2011.

[30] E. Mower Provost. Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3682–3686. IEEE, 2013.

[31] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled hmm for audio-visual speech recognition.

[32] M. Pantic and L. J. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[33] M. D. Pell and S. A. Kotz. On the time course of vocal emotion recognition. *PLoS One*, 6(11):e27256, 2011.

[34] S. Polikovsky, Y. Kameda, and Y. Ohta. Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor. In *International Conference on Crime Detection and Prevention (ICDP)*, pages 1–6. IET, 2009.

[35] K. S. Rao and S. G. Koolagudi. *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer Science & Business Media, 2013.

[36] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive conditional ordinal random fields for facial action intensity estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 492–499, 2013.

[37] A. Savran, B. Sankur, and M. T. Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774–784, 2012.

[38] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, et al. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In *INTERSPEECH*, pages 2253–2256. Citeseer, 2007.

[39] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. "of all things the measure is man": Automatic classification of emotions and inter-labeler consistency. In *ICASSP*, pages 317–320. Citeseer, 2005.

[40] J. Wagner, E. Andre, F. Lingenfelser, and J. Kim. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 2(4):206–218, 2011.