# Emotion Recognition During Speech Using Dynamics of Multiple Regions of the Face

YELIN KIM and EMILY MOWER PROVOST, University of Michigan, Ann Arbor

The need for human-centered, affective multimedia interfaces has motivated research in automatic emotion recognition. In this article, we focus on facial emotion recognition. Specifically, we target a domain in which speakers produce emotional facial expressions while speaking. The main challenge of this domain is the presence of modulations due to both emotion and speech. For example, an individual's mouth movement may be similar when he smiles and when he pronounces the phoneme /IY/, as in "cheese". The result of this confusion is a decrease in performance of facial emotion recognition systems. In our previous work, we investigated the joint effects of emotion and speech on facial movement. We found that it is critical to employ proper temporal segmentation and to leverage knowledge of spoken content to improve classification performance. In the current work, we investigate the temporal characteristics of specific regions of the face, such as the forehead, eyebrow, cheek, and mouth. We present methodology that uses the temporal patterns of specific regions of the face in the context of a facial emotion recognition system. We test our proposed approaches on two emotion datasets, the IEMOCAP and SAVEE datasets. Our results demonstrate that the combination of emotion recognition systems based on different facial regions improves overall accuracy compared to systems that do not leverage different characteristics of individual regions.

Categories and Subject Descriptors: I.5.4 [**Computing Methodologies**]: Pattern Recognition—*Applications*

General Terms: Human Factors

Additional Key Words and Phrases: Emotion, emotion recognition, facial movement, segmentation

**ACM Reference Format:**
Yelin Kim and Emily Mower Provost. 2015. Emotion recognition during speech using dynamics of multiple regions of face. ACM Trans. Multimedia Comput. Commun. Appl. 12, 1s, Article 25 (October 2015), 23 pages.
DOI: http://dx.doi.org/10.1145/2808204



## 1. INTRODUCTION

Emotion is a central part of human communication. It governs how we interact with each other, and how we respond to and perceive the outside world. Emotion is expressed across multiple modalities, such as facial, vocal, and bodily expressions. Facial cues include important emotion-related expressions, such as smiling or frowning. However, these movements may occur with and be modulated by speech-related movements [Kim and Mower Provost 2014]. For example, when a person either smiles or says the word "cheese," regions of the face (e.g., the mouth) may be modulated similarly. However, the facial expressions in the latter example may be caused by speech-related movement corresponding to the /IY/ sound. This renders challenging to differentiate between emotions based solely on facial movement. Therefore, it is essential that systems tease apart these two sources of facial movement to improve facial emotion

Authors' addresses: Y. Kim and E. Mower Provost, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109; email: yelinkim@umich.edu.
This article is based on the authors' previous work "Say Cheese vs. Smile: Reducing Speech-Related Variability for Facial Emotion Recognition" in *Proceedings of the ACM International Conference on Multimedia (ACM MM'14)*.

DOI: http://dx.doi.org/10.1145/2808204

recognition performance. In this work, we study how to represent and analyze dynamic facial movement, in the presence of speech, in order to improve the prediction accuracy and interpretability of facial emotion recognition systems.

The majority of previous work in facial emotion recognition has focused on facial expressions produced when an individual is not speaking (e.g., a person smiling) [Shan et al. 2009; Black and Yacoob 1997]. However, an increasing body of literature has analyzed facial movement that cooccurs with speech [Bevacqua and Pelachaud 2004; Metallinou et al. 2010; Mariooryad and Busso 2013]. These studies demonstrated the importance of reducing the effect of speech-related variability on facial movement. The cited works decreased this effect by implementing *phoneme*-dependent emotion classification systems, where a *phoneme* is defined as the smallest unit in a given language. Phoneme-dependent modeling often includes two key modules: (i) segmenting visual cues into phoneme-level units (*phoneme segmentation*) and (ii) classifying emotion within groups of similar *visemes*, visual cues associated with phonemes (*viseme-group classification*). The efficacy of this technique is very clear. However, the challenge is the need for a phonetic transcript to both identify phoneme boundaries and to assign phoneme content. In this work, we explore the first challenge: the identification of phoneme boundaries and ask whether we can find other segmentation strategies for facial movement.

In our previous work, we proposed an unsupervised segmentation strategy to circumvent this requirement [Kim and Mower Provost 2014]. We focused on explicit modeling of facial movements, represented as three-dimensional facial point positions, rather than facial appearance features (e.g., Local Binary Patterns [Shan et al. 2009]). Although the facial appearance features have also been effectively utilized in facial emotion recognition studies [Zhao and Pietikainen 2007; Shan et al. 2009; Dhall et al. 2011], it is difficult to reliably extract meaningful features when frontal face videos/images are not available or occluded. We investigated both the application of sliding windows in addition to segmentation using the natural temporal dynamics of the underlying signal. Sliding-window segmentation is a strategy commonly employed in emotion recognition studies [Mower and Narayanan 2011; Sandbach et al. 2011; Mower Provost 2013; Nicolle et al. 2012]. In this strategy, the facial data are segmented into smaller units, all with the same duration. However, this method is not based on the underlying dynamics of the signal and may miss important patterns in the signal. Further, previous work has demonstrated that the use of segmentation based on fixed length windows performs more poorly than phoneme segmentation [Mariooryad and Busso 2013]. To overcome this limitation, our previous work proposed an automatic, unsupervised segmentation method based on mouth movement, which utilizes a trajectory segmentation algorithm proposed by Lee et al. [2007] for trajectory segmentation and clustering. The algorithm was motivated by Minimum Description Length (MDL) principle, widely used in information theory Our proposed method does not require a phonetic transcript and achieved comparable performance to phoneme segmentation when used as a component of a facial emotion recognition system [Kim and Mower Provost 2014]. However, the limitations of our previous study were two-fold: (i) the proposed methods were tested on a single dataset and (ii) the varying temporal characteristics of different facial regions were not explored.

In the presented work, we assess the utility of unsupervised segmentation approaches by testing our method using an additional database to understand the impact of variable-length segmentation (i.e., unsupervised MDL-based segmentation and phoneme segmentation) and viseme-group classification on facial emotion recognition systems. We discuss the specific effects of the proposed segmentation and classification strategies across two different motion-capture datasets recorded in different settings: read speech (SAVEE) and two-person conversation (IEMOCAP). We found

that when using viseme-group classification it is advantageous to use variable-length segmentation compared to fixed-length segmentation. Further, we analyze the impact of individual facial regions. The results demonstrate that we can increase system-level performance by changing how we integrate information from the facial regions. The results strengthen our argument that both variable-length segmentation and viseme-group classification are critical for facial emotion recognition systems.

## 2. RELATED WORK

### 2.1. Audio-Visual Emotion Recognition in Multimedia

Multimedia emotion recognition studies have gained in popularity due to the growing prevalence of natural human-machine interfaces and the increasing need for automatic behavior assessment. Audio-visual emotion recognition systems utilize a wide range of signals from speech, facial, and body expressions to visual descriptors of multimedia data. Surveys of recent emotion recognition systems can be found in [Narayanan and Georgiou 2013; Hussain et al. 2014; Schuller et al. 2013; Cambria et al. 2013; Calvo and D'Mello 2010; El Ayadi et al. 2011; Gunes et al. 2011; Kleinsmith and Bianchi-Berthouze 2013]. In this article, we highlight the emotion recognition literature focused on speech and facial features. We review emotion recognition systems that use unimodal (i.e., only using speech or facial features) and multimodal (i.e., using both speech and facial features) data.

*2.1.1. Unimodal Emotion Recognition.* Speech is one of the most important methods of human communication [El Ayadi et al. 2011]. The progress made in speech recognition has sparked new research directions into methods to extract and analyze the emotional content from speech [Schuller et al. 2013; Lee and Narayanan 2005; Mower et al. 2009; Schuller et al. 2011; Metallinou et al. 2013]. The studies include investigations into speech features and feature selection methods [Wu et al. 2011; Gharavian et al. 2012], proper units of analysis [Koolagudi et al. 2011; Gold et al. 2011], and classification methods [Lee et al. 2011; Garg et al. 2013].

Emotion has also been modeled using visual cues. The goal is to automatically extract and analyze salient visual information. The earlier work focused on automatic facial expression recognition, since facial expressions arguably contain the most discriminative visual features for emotion recognition [Wan and Aggarwal 2014; Pantic and Bartlett 2007]. Recent studies have focused on inferring emotion based on salient visual cues, not only including facial expression features, but also other types of visual features, such as aesthetic features (introduced by Bhattacharya et al. [2013]), to understand perceived emotions [Jou et al. 2014; Chen et al. 2014].

*2.1.2. Multimodal Emotion Recognition.* Researchers found that the joint use of speech and facial cues can improve overall accuracy in emotion recognition. Many studies have investigated how to combine these two modalities and how to build a classification system that could effectively fuse this information [Kächele et al. 2014; Meng and Bianchi-Berthouze 2011; Sánchez-Lozano et al. 2013; Savran et al. 2012].

Savran et al. [2012] proposed an affect estimation method that combined audio, facial, and lexical information using particle filtering. The authors demonstrated that the predictions from each modality could be effectively combined as measurement variables in particle filter methods. Unlike other Bayesian filtering methods, (e.g., Kalman filtering), the advantages of particle filtering comes from its reduction in assumptions of linearity or Gaussianity. Meng et al. proposed a multistage emotion recognition system using Hidden Markov Models (HMMs) [Meng and Bianchi-Berthouze 2011]. The multistage system showed significant performance gain over a single-stage alternative. Kächele et al. [2014] presented a hierarchical emotion and depression recognition

system that trained ensembles of weak learning algorithms and fuses the audio and facial data using a Kalman filter at the decision level. The International Audio/Visual Emotion Challenge (AVEC) has encouraged the analysis of natural, continuous, emotion prediction. It has also highlighted medical applications, particularly in depression. The challenge participants have demonstrated different methods in multimedia processing and machine learning techniques [Schuller et al. 2011].

### 2.2. Temporal Segmentation of Audio-Visual Data

Audio-visual data are often temporally segmented into smaller units to obtain more meaningful features [Schuller and Rigoll 2006; Koolagudi et al. 2011], to build dynamic classifiers [Mower and Narayanan 2011; Sandbach et al. 2011], and to find semantically meaningful regions [Rui et al. 2000; Bigot et al. 2008; Arons 1994; Koolagudi et al. 2011].

Rui et al. studied an automatic method to extract scene highlights for TV baseball games [Rui et al. 2000]. They used speech features to highlight the videos. They identified the 'excited speech' of game announcers and game-specific sounds (e.g., baseball hits) and fused the information using probabilistic frameworks to enhance highlight detection. Bigot et al. proposed a method to find semantically salient regions in audio-visual data using either audio, video, or semantic content [Bigot et al. 2008]. Lee et al. [2007] presented a trajectory segmentation and clustering method based on MDL. The method automatically segments time-series data based on rapid changes, and clusters the segments using density-based clustering. The trajectory segmentation is based on two types of line-segment distance measures: perpendicular and angle distance. To reduce the complexity, they proposed an approximate solution to find rapidly changing points. This algorithm was used to identify emotionally salient regions of audio-visual speech [Mower Provost 2013].

Koolagudi et al. [2011] studied methods to segment speech for emotion recognition based on the prosody of speech segments. They used words and syllables as units of the segments. They found that the system-level performance using prosody-based speech segments was not high, but that the performance significantly improved when combined with spectral features. A Comprehensive survey on phoneme segmentation was conducted in Toledano et al. [2003]. The studies have mainly focused on speech data. In speech recognition, sub-word units such as phonemes are often used since word-level or whole-word models are challenging to build due to large vocabulary sizes in natural language [Gold et al. 2011]. Several recent works have approached phoneme segmentation problems as well [Kalinli 2012; Qiao et al. 2008; Keshet et al. 2005]. The work of Keshet et al. [2005] tackles phoneme alignment based on discriminative learning, similar to SVM. The work of Qiao et al. [2008] presents an unsupervised method for phoneme segmentation.

### 2.3. Phoneme or Viseme-Dependent Modeling

Systems that employ phoneme or viseme-dependent modeling seek to disentangle how speech and emotion modulate speech and facial movement. These studies are predicated on the knowledge that groups of phonemes can be modulated differently by the same emotion [Lee et al. 2004; Busso and Narayanan 2007]. These systems have been demonstrated effective for emotion classification [Mariooryad and Busso 2013; Metallinou et al. 2010]. Phoneme-based modeling has also been demonstrated to be effective in cross-corpora experiments. For instance, Vlasenko et al. [2014] found that emotion classification systems that use phonetic transcripts obtained from a phoneme-level bigram language model can increase the cross-corpora classification accuracy.

In particular, researchers found that facial cues are difficult to model when facial movement is modulated by both emotion and speech production [Kim and

Mower Provost 2014; Mariooryad and Busso 2013; Metallinou et al. 2010]. These studies have approached these challenges by building emotion classification systems that train classifiers on specific groups of phonemes with similar facial movement. This construction allows for a focus on modulations due to emotion, rather than due to articulation and emotion. Metallinou et al. first conducted phoneme-dependent modeling on the IEMOCAP database for facial emotion recognition [Metallinou et al. 2010]. They presented an emotion classification system, based on Hidden Markov Models (HMM), that separates the classifiers into 14 similar viseme groups, the groups also used in our study. The highest unweighted accuracy they achieved was 55.74%, when using viseme-specific HMMs. Mariooryad and Busso [2013] studied two types of methods to reduce or compensate for speech variability in facial emotion recognition: feature-level and model-level compensation. The feature-level method normalizes phoneme-dependent patterns in facial movement using the whitening transformation to compensate for the difference in phoneme-dependent patterns in the features. The model-level method separates emotion classifiers into viseme-dependent groups. The study found that both the feature and model-level compensation methods improve overall performance. In particular, their results showed a larger performance gain for the model-level method, compared to the feature-level method. The previous studies demonstrated the benefits of phoneme segmentation and viseme-group classification. However, an open question remains in how similar levels of accuracy can be achieved without segmenting based on phoneme transcript and whether phonemes are the correct unit for segmentation.

## 3. AUDIO-VISUAL DATABASES

In this work we use the IEMOCAP audio-visual emotion database [Busso et al. 2008], also used in [Mariooryad and Busso 2013; Metallinou et al. 2010; Kim and Mower Provost 2014], and the SAVEE dataset [Haq and Jackson 2010]. Both of the databases contain audio, visual, and motion capture data. Further, both provide phoneme-level transcripts, obtained by force aligning the transcript to the acoustic data.

### 3.1. IEMOCAP Database

The IEMOCAP database [Busso et al. 2008] contains approximately 12 hours of dyadic conversations between five pairs of actors (each pair contains one male and one female). This database has been widely used in the field of automatic audio-visual emotion recognition [Metallinou et al. 2012; Kipp and Martin 2009; Mower et al. 2011; Lee et al. 2009]. Each session contains both acted and improvised scenarios. The data are captured using audio-visual cameras and a nine-camera Vicon recoding system, providing 3-D marker coordinates at 120 frames per second (fps). The data include 53 motion capture markers on the actor's face. We use a subset of 46 markers, as shown in Figure 1. The five nose markers are excluded due to their limited movement, and the two eyelid markers are also excluded due to their frequent occlusions, as in [Metallinou et al. 2010; Kim and Mower Provost 2014].

The data were evaluated per utterance (i.e., a turn that an actor is actively speaking) by human evaluators using both categorical and dimensional labeling schemes. The dimensional attributes include valence (positive vs. negative), activation (calm vs. excited), and dominance (passive vs. dominant). They were evaluated by at least two evaluators. The categorical labels include *Anger, Happiness, Neutrality, Sadness, Excitement, Surprise, Frustration, Fear, Disgust, Other*. They were evaluated by at least three evaluators. We use utterances with majority voted categorical labels from the set: *Anger, Happiness+Excitement, Neutrality, Sadness*, in line with previous studies [Mariooryad and Busso 2013; Metallinou et al. 2010]. There are $43.0 \pm 26.2$ angry,

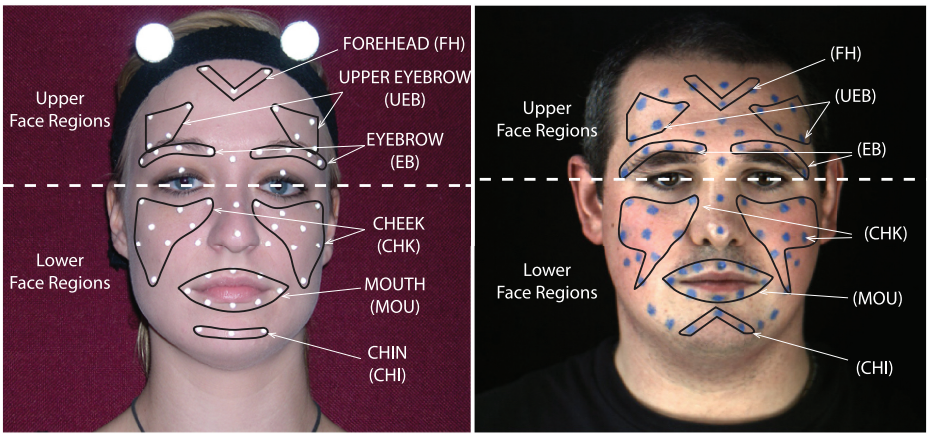


Fig. 1. Positions of face markers and six face regions: chin (CHI), forehead (FH), cheek (CHK), upper eyebrow (UEB), eyebrow (EB), and mouth. The images are from the IEMOCAP (left, [Busso et al. 2008]) and SAVEE datasets (right, [Haq and Jackson 2010]).

$91.5 \pm 37.5$ happy, $44.6 \pm 27.4$ neutral, and $51.8 \pm 28.1$ sad utterances per speaker, totaling 3,060 utterances over all speakers. The mean length of an utterance is $4.73 \pm 3.34$ seconds. Utterances have an average of 0.75 seconds of silence at the beginning of an utterance and 0.86 seconds at the end of an utterance.

### 3.2. SAVEE Database

The SAVEE dataset contains read speech of four male British English speakers, eliciting six emotions: *Anger, Disgust, Fear, Happiness, Sadness,* and *surprise*. Each emotion was expressed in 15 phonetically balanced sentences, and *Neutrality* in 30 sentences. This results in 480 utterances in total. In our work, we use four classes for consistency with the IEMOCAP database: *Anger, Happiness, Neutrality, and Sadness*, resulting in 300 utterances in total. The average length of an utterance within the subset is $3.85 \pm 0.33$ seconds. The utterances have an average of 0.51 seconds of silence at the beginning of an utterance and 0.55 seconds at the end of an utterance. The facial data include 2D coordinates of 60 markers on the forehead, eyebrows, cheeks, lips and jaw (Figure 1). The sampling rate was 44.1 kHz for audio, and 60 fps for video.

The provided emotion labels of the SAVEE dataset are the labels given to the actors, rather than the intended target emotion (annotated labels were not available). This is different from IEMOCAP, in which we use emotion labels derived from perceptual evaluations. However, the authors presented a high level of agreement between the intended target emotion and perceived emotion: 441 out of 480 total sentences in the data were perceived as the intended target emotion by at least 8 out of 10 annotators, indicating good agreement between the actor's intended emotion and the annotator's perception [Haq and Jackson 2010]. Additional differences between the SAVEE and IEMOCAP databases include: (i) 2-dimensional vs. 3-dimensional motion capture data, (ii) motion-capture frame rate of 60 fps vs. 120 fps, (iii) four speakers, each with scripted utterances, vs. ten speakers, each embedded within a dyadic interaction.

We use a subset of the 60 motion capture markers to have a configuration similar to the IEMOCAP database. The subset totals 46 markers (Figure 1). Further, we address the difference in fps between the two databases by interpolating the SAVEE motion capture data using cubic spline interpolation (described in Section 4.1) to increase the

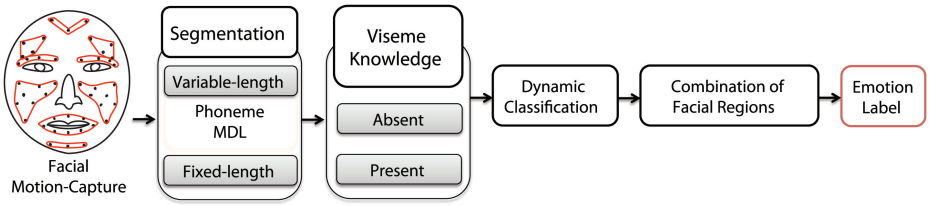


Fig. 2. Our system uses facial motion-capture data. It investigates three segmentation methods and explores the benefit of using knowledge of the spoken content of each segment. The system estimates the emotion label by estimating the similarity of the movement in each segment to movement observed in specific emotion classes. Finally, it combines the emotion estimates provided by the individual facial regions to infer a final estimated emotion label.

Table I. Summary of the Abbreviations Associated with the Six Approaches Tested in This Work

| | | Segmentation | | |
|---|---|---|---|---|
| | | Phon | MDL | Win |
| Classification | General | Gen/Phon | Gen/MDL | Gen/Win |
| | Viseme-group | VG/Phon | VG/MDL | VG/Win |

frame rate to 120 fps. We discuss the impact of this interpolation in Section 4.1. Finally, we scale the SAVEE motion capture data to have the same minimum and maximum values as in the IEMOCAP database.

## 4. METHODOLOGY

The overview of our proposed method is shown in Figure 2. We first separate the tracked marker positions into six facial regions to capture the facial region-specific characteristics in emotion expression (Section 4.1). We then temporally segment the data using three segmentation methods (fixed-length, phoneme, and MDL-based; Section 4.2) and measure the time-series similarity between the identified segments using Dynamic Time Warping (DTW). We calculate the distribution of emotion classes over each segment and use this information to estimate the emotion class of the segment. We aggregate the segment-level emotion estimates over the utterance to estimate the utterance-level emotion, described in detail in Section 4.4. During classification, we explore the benefit of using viseme-group classification, given each of the three segmentation strategies. This allows us to understand the impact of using knowledge of the viseme group in classification. We refer to classification as *general* (contrasted with *viseme-group*) when we do not take the knowledge of viseme information into account, described in detail in Section 4.3. Finally, we investigate different methods to combine the emotion evidence derived from the individual facial regions, described in Section 4.5.

In our experiments, we test six approaches that use combinations of different temporal segmentation and classification methods, originally proposed in Kim and Mower Provost [2014]. The two rows in Table I describe the classification scheme: general and viseme-group classification, and the three columns describe the segmentation scheme: phoneme, MDL, and fixed-length sliding window.

### 4.1. Motion Capture Preprocessing

Both the IEMOCAP and SAVEE datasets provide facial markers that are (1) translated so that a nose tip becomes the origin of each frame, and (2) rotated to compensate for head movement. In addition, we perform mean-normalization on the facial data of individual speakers to mitigate their different facial configurations. The

mean-normalization method was suggested in Metallinou et al. [2010] and used in our previous work [Kim and Mower Provost 2014]. We compute the global mean value over all speakers for each marker coordinate and scale each individual speaker's data to make the mean of each speaker to be the same as the global mean.

We divide the facial motion capture data into six facial regions to study region-specific facial movements, including: chin, forehead, cheek, upper eyebrow, eyebrows, and mouth, as in Kim and Mower Provost [2014]. As shown in Figure 1, there are three markers are in the *Chin* and *Forehead regions*, 16 markers in the *Cheek region*, and eight markers in the *Upper Eyebrow, Eyebrow*, and *Mouth regions*. We track the region-specific marker positions and represent each as a multidimensional trajectory. For instance, given a data segment with $N$ motion-capture frames and $M$ marker coordinates (3-D for IEMOCAP and 2-D for SAVEE), the final data are an $N \times M$ trajectory.

As in our previous work [Kim and Mower Provost 2014], we exclude segments with fewer than seven frames or approximately 0.058 seconds. Our preliminary work demonstrated that the exclusion of segments with short durations does not make significant changes in emotion classification accuracy, which may due to insufficient temporal information within the segments. The computation time during DTW calculation can be considerably improved by excluding such segments, since 43.5%, 0.67%, and 0.86% of all phoneme, MDL, and fixed-length segments in the IEMOCAP dataset has duration less than seven frames, respectively. The high percentage of excluded phoneme segments occurs because many of the phonemes in the data have very short durations. Further, we drop segments with any missing values in the 46 markers we use. This results in different sets of utterances for each segmentation scheme. We use the set of 3,060 intersecting utterances. This number is slightly higher than in Metallinou et al. [2010] and similar to Mariooryad and Busso [2013]. In the SAVEE dataset, 34.86% of the phonemes are rejected, 2.23% of the MDL segments are rejected, and 1.44% of the window segments are rejected. The number of utterances remains the same after the exclusion process.

Our preliminary experiments showed that the difference between the SAVEE and IEMOCAP datasets in terms of frame rate (60 fps and 120 fps, respectively) impacted the overall accuracy. We mitigate this effect by increasing the SAVEE frame rate to 120 fps using cubic spline interpolation. This interpolation allows us to apply the same pre-processing steps to SAVEE as applied to the IEMOCAP (e.g., excluding of segments less than seven frames). In addition, our preliminary investigations showed that the SAVEE dataset had marker coordinates had a smaller range than IEMOCAP. This difference in range affected the MDL segmentation process. To mitigate this effect, for each marker coordinate, we scaled the SAVEE data to have the same minimum and maximum value as the IEMOCAP data. After MDL segmentation, we used the original marker values without scaling for the remainder of the classification framework to retain the original characteristics of the SAVEE dataset.

### 4.2. Segmentation

*4.2.1. Sliding Window Segmentation ("Win").* The Win segmentation method segments each utterance into fixed-length windows. We use window segments without overlapping to enable comparisons with the phoneme and MDL segmentation methods, which do not have overlapping windows. We retain all windows, including segments at the end of an utterance that are shorter than the standard window size. For instance, consider an $N \times M$ trajectory of the eyebrow region over an utterance, where $N$ is the number of frames and $M$ is the number of marker coordinates ($N = 128$ and $M = 24$ for the IEMOCAP data). If we segment this trajectory using 0.1-second window there will be 12 frames per window. The resulting segments of this utterance are ten trajectories, each of size $12 \times 24$ and one trajectory of size $8 \times 24$.

Table II. Visually Similar Viseme Groups (VG)

| Group | Phonemes | Group | Phonemes |
|---|---|---|---|
| V1 | P, B, M | V8 | AE, AW, EH, EY |
| V2 | F,V | V9 | AH,AX,AY |
| V3 | T,D,S,Z,TH,DH | V10 | AA |
| V4 | W,R | V11 | AXR,ER |
| V5 | CH,SH,ZH | V12 | AO,OY,OW |
| V6 | K,G,N,L,HH,NG,Y | V13 | UH,UW |
| V7 | IY,IH, IX | V14 | SIL |

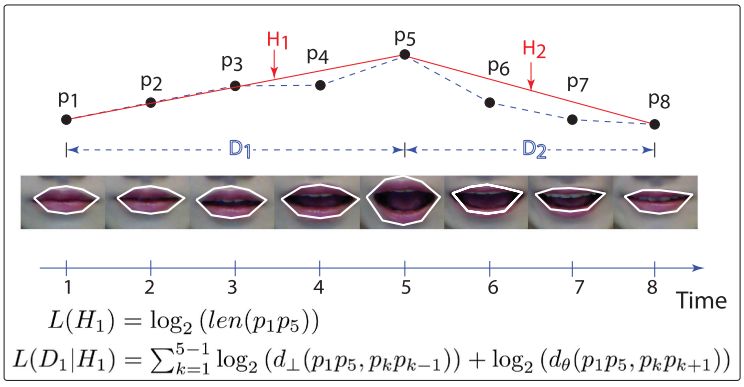


Fig. 3. An example of the MDL segmentation method for visualization. The x-axis is time and y-axis is the position of the mouth marker at the top of the lip over part of an utterance. The blue dashed lines are the marker position, the red lines are the approximated movement using the MDL approach. The proposed MDL segmentation method finds regions where the dynamics of the movement are consistent. In this example, the mouth opens widely and then starts to close at frame $p_5$. Therefore, MDL uses $\{c_1 = 1, c_2 = 5, c_3 = 8\}$ (including the starting and end point of each utterance) as characteristic points. The hypothesis $H_1$ and $H_2$ correspond to the segmentation based on the characteristic points, lines between $p_1$ and $p_5$ and $p_5$ and $p_8$. The data $D_1$ and $D_2$ are the original mouth movement $\{p_1p_2, p_2p_3, p_3p_4, p_4p_5\}$, and $\{p_5p_6, p_6p_7, p_7p_8\}$.

*4.2.2. Phoneme Segmentation ("Phon").* The Phon segmentation method segments the facial data within an utterance based on the temporal phoneme boundaries. For instance, if a speaker is saying "hello," we segment the facial trajectories using the phoneme boundaries between /SIL/, /HH/, /AH/, /L/, /OW/, and /SIL/ phonemes. The set of phonemes that we use in this study is in Table II. The boundaries for these phonemes were obtained by force aligning the audio to the known transcript. The average length of phoneme segments is $0.17 \pm 0.01$ seconds for the IEMOCAP data, and $0.14 \pm 0.01$ seconds for the SAVEE data.

*4.2.3. MDL Segmentation (MDL).* In this work, we describe an unsupervised variable-length segmentation that does not require a phonetic transcript, proposed in our earlier work [Kim and Mower Provost 2014]. We segment the data using the movement of the mouth. This allows us to capture the facial cues that are most highly related to speech production, important due to the focus on viseme-group classification. The segmentation algorithm was originally proposed in the context of a trajectory segmentation and clustering algorithm, called TRACLUS [Lee et al. 2007]. It automatically finds points that should be used to segment regions of the data with different temporal characteristics. The application of this algorithm in the context of facial movement allows us to segment the facial data based on the natural dynamics of the mouth. A mouth-based example is presented in Figure 3.

The segmentation algorithm finds a hypothesis, $H$, that describes the original data $D$ trading off between conciseness and preciseness of the description. It aims to minimize the sum of $L(H) + L(D|H)$, where $L(H)$, length of the hypothesis, computes the conciseness of the hypothesis, and $L(D|H)$, the distance between the hypothesis and the data, measures the preciseness of the hypothesis. By minimizing $L(H) + L(D|H)$, the algorithm finds regions of consistent dynamics. For instance, Figure 3 shows the mouth trajectory example where, after frame $p_5$, the trajectory changes. In this case, MDL would identify $\{c_1 = 1, c_2 = 5, c_3 = 8\}$ (including the starting and end point of the trajectory) as *characteristic points*. Characteristic points mark the beginning and end of regions with consistent dynamics. In Lee et al. [2007], the authors proposed to measure $L(H)$ as the length of the proposed segmentation (e.g., in Figure 3, $L(H_1)$ is measured as the log of the length of a line connecting $p_1$ to $p_5$). The quantity $L(D|H)$ captures the difference between the original line segments, $D$ and the proposed segmentation, $H$. For example, in Figure 3, $L(D_1|H_1)$ is the log of the summation of differences between each of the blue dashed lines $p_1p_2$, $p_2p_3$, $p_3p_4$, $p_4p_5$, and the red line $p_1p_5$. The segmentation can be formulated as an optimization problem: Equation (1).

$$\begin{aligned} &\arg\min_{H} && L(H) + L(D|H), \\ &\text{where} && L(H) = \sum_{j=1}^{n-1} \log_2(len(p_{c_j} p_{c_{j+1}})), \\ & && L(D|H) = \sum_{j=1}^{n-1} \sum_{k=c_j}^{c_{j+1}-1} \log_2(d_\perp(p_{c_j} p_{c_{j+1}}, p_k p_{k+1})) + \log_2(d_\theta(p_{c_j} p_{c_{j+1}}, p_k p_{k+1})). \end{aligned} \tag{1}$$

In Equation (1), $d_\perp$ is the perpendicular distance between the line segments and $d_\theta$ is the angular distance [Chen et al. 2003].

As an approximate solution, the TRACLUS algorithm [Lee et al. 2007] compares the cost of partitioning, $cost_{par}$, and nonpartitioning, $cost_{nopar}$, at each data point, $p$, Equation (2).

$$cost_{par} = L(H) + L(D|H), \quad cost_{nopar} = L(D) = \sum_{j=1}^{p-1} \log_2(len(p_j p_{j+1}). \tag{2}$$

The algorithm advances through the trajectory and estimates whether the data should be segmented at each point. The algorithm makes a segmentation decision based on the equation: $cost_{par} \geq cost_{nopar} + MDL_{Advantage}$. When this equation is true, the algorithm identifies the characteristic point as the previous point, marking the end of a segment. The characteristic point is the point prior to the one where the cost of partitioning is suddenly higher than the cost of not partitioning. The point at which the inequality is true then forms the beginning of the next segment. It is important to note that the parameter $MDL_{Advantage}$ controls the granularity of the segmentation and hence the average of segment length. We describe the method that we use to choose $MDL_{Advantage}$ in Section 4.4.1. Additional details can be found in Lee et al. [2007].

In our work, the input to MDL segmentation is the mouth trajectory (24-dimensional for IEMOCAP and 18-dimensional for SAVEE) smoothed using a median filter with a window size of three (window size chosen empirically), to smooth the 3D-captured mouth movement trajectory.

### 4.3. Knowledge of Viseme Information

Studies of visual speech production have indicated that there are groups of visemes with similar facial movements (Table II) [Lucey et al. 2004]. Recent research has found

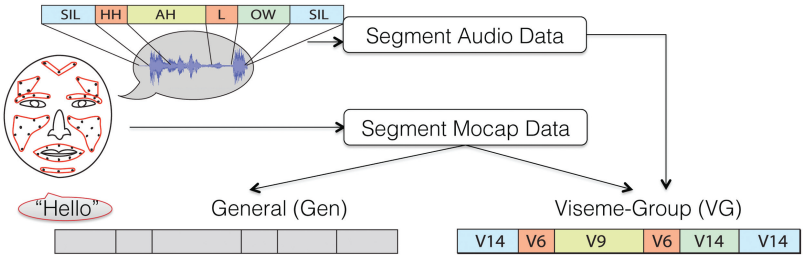


Fig. 4. A comparison between the general and viseme-group classification methods for an example in which a speaker is saying "hello." In viseme-group classification, we use knowledge of what was said, assigning the viseme group that occupies the longest duration within each segment. We separate the segments into different emotion classifiers based on their assigned viseme group. In general classification, this information is not used.

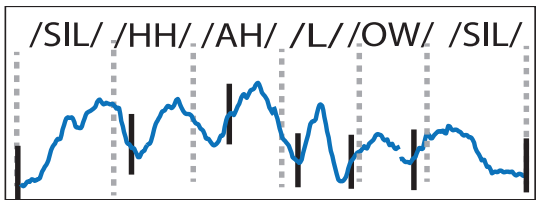


Fig. 5. VG/MDL example for describing how to assign phoneme content to MDL segments. The gray dashed lines show the phoneme boundaries and the black hash marks represent different segment boundaries. Notice the potential mismatch between the hash marks and dashed lines in MDL segmentation. For example, /SIL/ and /HH/ are both present in the second segment. The /HH/ phoneme occupies a longer duration for this segment and the viseme group associated with /HH/ is assigned to this segment.

that it is beneficial to separate emotion classifiers into 14 similar viseme groups, so that each classifier has less speech-related variation [Metallinou et al. 2010; Marioryad and Busso 2013]. We add to this knowledge by understanding how segmentation affects the utility of viseme-group classification.

We use two classification schemes: viseme-group and general classification. In viseme-group classification (VG), it is assumed that the classifier knows which viseme group the segment belongs to. We implement this by assigning a viseme group label to each segment based on the phoneme content. For example, if the speaker says "hello", we have /SIL/ (silence), /HH/, /AH/, /L/, /OW/, and /SIL/ phonemes. The two /SIL/ phonemes will be compared in emotion classifier 14, and /HH/ and /L/ in classifier 6, etc (Figure 6 and Table II). In general classification (Gen), it is assumed that this knowledge is absent. This results in a single emotion classifier that has data from all viseme groups (Figure 4).

For MDL and Win segmentation, the segment boundaries may not line up with the phoneme boundaries. To estimate the corresponding viseme group of MDL and Win segments, we assign a viseme group label to a segment based on the phoneme that occupies the longest duration within each segment, for VG/MDL and VG/Win. For instance, in Figure 5 we consider a VG/MDL example. Note the mismatch between the phonetic transcript (dashed line in the figure) and the MDL segmentation result (hash marks). If the first MDL segment is 85% /SIL/ and 15% /HH/, we assign the phoneme content of the segment to the /SIL/ group, and apply emotion classifier 14.

### 4.4. Emotion classification

*4.4.1. Cross Validation.* Our proposed methods have two hyperparameters: $MDL_{Advantage}$ (MDL segmentation) and window length (Win segmentation). We choose $MDL_{Advantage}$

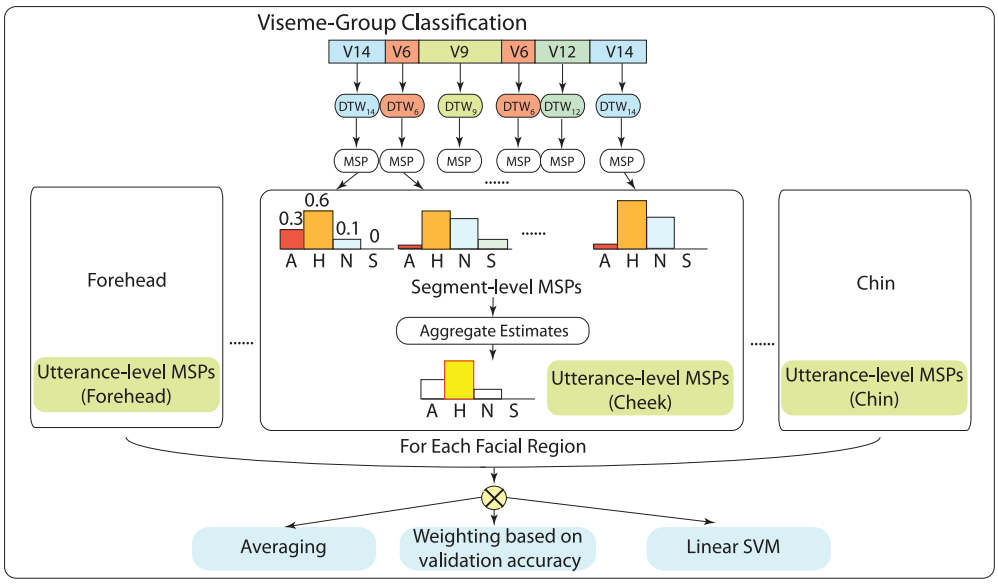


Fig. 6. MSP calculation and facial-region combination example for viseme-group classification.

from the set {0, 6, 10, 20}. We choose the window length from the set of {0.1, 0.25, 0.5, 1, 1.5, 2} seconds.

We build speaker-independent emotion recognition systems using leave-one-speaker-out cross validation, and tune the parameters ($MDL_{advantage}$ and window length) using leave-one-training-speaker-out cross validation. For each speaker in the training set, we hold out a speaker as a validation speaker and train the model using the rest of the training speakers. We repeat this process over each training speaker and calculate the average of the validation accuracy. We choose the value of the parameter that maximizes performance over the set of validation speakers. For the SAVEE dataset, we also do lexical-independent classification to ensure that the same sentence does not appear in both the training and testing sets. This is because SAVEE is a read emotional speech database that has emotion-specific sentences (12 of 15 sentences were emotion-specific for each emotion class: Angry, Happy, and Sad).

*4.4.2. DTW-Motion Similarity Profile Emotion Classification.* We use the DTW-Motion Similarity Profile (MSP) method proposed in Kim and Mower Provost [2014] to infer utterance-level labels based on the temporal similarity between segments, as shown in Figure 6. The DTW method is computationally costly in the inference stages since it compares a test data to all training data. However, the method can provide interpretable descriptions about how two facial movements are similar. This method has two steps for emotion classification: (1) segment-level DTW calculation and (2) utterance-level emotion inference.

First, we calculate the segment-level similarity in facial movement between the training and test segments. For instance, if we have two $K$-dimensional facial movement trajectories of length $M_1$ and $M_2$, i.e., $T_1 \in \mathbb{R}^{M_1 \times K}$ and $T_2 \in \mathbb{R}^{M_2 \times K}$, we compute the similarity between the two trajectories using the multidimensional the algorithm. It computes the $M_1$-by-$M_2$ local cost matrix $Q$ as follows, where $i$ and $j$ denote the frame-level temporal point of $T_1$ ($1 \leq i \leq M_1$) and $T_2$ ($1 \leq j \leq M_2$), respectively,

$$Q(i, j) = \sum_{k=1}^{K} (T_1(i, k) - T_2(j, k))^2, \quad (3)$$

Then, for each facial region, we calculate the emotional distribution of the $k$ closest training segments, called the segment-level MSP, where $k$ is 20, as in our previous work [Kim and Mower Provost 2014]. For instance, the first segment-level MSP in Figure 6 represents a four-dimensional vector $\{0.3, 0.6, 0.1, 0\}$, calculated using the labels of the 20 closest training segments: 6 angry, 12 happy, 2 neutral, and 0 sad.

Once we have segment-level MSPs for each segment, we average these to create an utterance-level MSP, a single four-dimensional emotion estimate for each facial region. We combine individual face regions with different methods (described in Section 4.5) to obtain the final utterance-level MSP. We normalize each of the four dimensions using speaker-specific z-normalization, to mitigate the imbalance in the emotion class distribution, e.g., there are approximately twice as many happy utterances compared to the other emotions. We assign the final utterance-level label based on the maximum component of the aggregated MSP, e.g., *happiness* in Figure 6.

### 4.5. Combination methods of Facial Regions

We investigate three types of decision-level combination methods of individual face regions: (i) simple averaging, (ii) weighted averaging, and (iii) SVM-based aggregation methods. The last stage of Figure 6 demonstrates that we combine utterance-level MSPs of individual face regions and explore the three combination methods.

*4.5.1. Averaging.* For the simple averaging method, we use ten different types of experiments to aggregate the MSPs from the individual facial regions, extended from our previous work [Kim and Mower Provost 2014]. In our previous work, we combined the facial region-specific MSPs to obtain the final MSP, representing the emotion estimates over the entire face. We report the ten AV('AVeraged faces') experimental results of (i) AV6 (averaged over all 6 face regions), (ii) AV4 (averaged over Chin, Cheek, Upper eyebrow, Mouth), (iii) AV Up (averaged over Forehead, Upper eyebrow, and Eyebrow) and (iv) AV Low (averaged over Chin, Cheek, and Mouth), and six individual facial regions of (v) Chin, (vi) Forehead, (vii) Cheek, (viii) Upper eyebrow, (ix) Eyebrow, and (x) Mouth. Unlike the previous work where we used segments with the same parameters (e.g., windows of the same fixed length, segments found using the same $MDL_{Advantage}$ parameter) over all AV experiments, we use the parameters chosen for individual AV experiments based on the cross validation accuracy (as described in 4.4.1). Different {speaker, classification (Gen or VG) methods, segmentation (Win, MDL, or Phon) methods} sets have different parameters chosen for each of the AV experiments. For each AV experiment, the individual face regions use the same parameter and are combined to calculate the final MSP.

*4.5.2. Weighting Based on Validation Accuracy.* In the second experiment, we aggregate the emotional evidence using a weighted average. This allows us to more strongly weight information from emotionally expressive areas of the face, compared to less emotionally expressive areas. We first identify the parameters that are associated with the highest performance for each facial region using cross validation (described in Section 4.4.1). We calculate the accuracy over the validation speakers and use these accuracies as the initial weights: $Val_i$. We sum the weights over the six facial regions and normalize each of the weights to ensure that they sum to 1. Then, rather than aggregating MSPs by averaging, we compute a weighted average using the learned weights.

*4.5.3. Linear-Support Vector Machine.* We investigate a third aggregation method, which allows for adaptation based on estimated emotional expressivity of the individual facial

regions. We use linear-kernel Support Vector Machine (SVM), in order to find the weighted linear combination of the MSPs that are associated with the individual facial regions. The input to the SVM is the six four-dimensional MSP estimates (associated with each of the six regions of the face). The goal of the SVM is to estimate the emotion class label. We select the parameter C ($10^k$) through cross validation, selecting over the set: $k = \{-6, -5, -4, -3, -2, -1, 0, 1\}$.

## 5. EXPERIMENTAL RESULTS

We present results for each database (IEMOCAP and SAVEE). We describe the results in terms of the three combination methods: (i) detailed experiments for each of the 10 averaging methods, where each of the ten methods use the best parameters chosen by cross validation, (ii) weighting of individual facial regions based on cross validation accuracy, and (iii) linear-SVM based weighting. In addition to the three segmentation methods of Win, MDL, and Phon segmentation, we present the utterance-level ('Utt') performance for general classification, where utterances are used without any segmentation. To be consistent with previous multiclass emotion recognition research [Mower et al. 2011; Lee et al. 2009], we use unweighted accuracy, or averaged recall, to calculate the average accuracy.

### 5.1. SAVEE Experiments

*Significance Tests.* To the best of our knowledge, previous work on the SAVEE dataset did not employ significance tests [Haq and Jackson 2010]. Since the SAVEE dataset has four speakers, each speaking the same set of utterances, we develop Generalized Linear Mixed Models (GLMM) with binomial link function that predicts the correctness of each utterance and speaker, similar to Boston et al. [2008]. The GLMM use mixed-effects models that incorporate both random and fixed-effects parameters. We develop the models that treat both test speakers and utterance IDs as the random effects. We then compare MDL and window segmentation, as well as Phon and window segmentation, each separately within VG and Gen classification. Hence, fixed effects of the GLM models are classification (Gen or VG), segmentation (Win, MDL, or Phon), and the interaction between classification and segmentation. For the random effects, we use both test speakers and utterance IDs. The response of our models is correctness of the emotion inference given each segmentation and classification methods, where the conditional distribution of the response given the random effects is assumed as the binomial distribution. We fit the models using `glmer` function, implemented in `R` [Bates et al. 2007]. In each experimental result, we claim significance in the accuracy between the MDL and window, as well as Phon and window segments when $p < 0.05$.

*5.1.1. Averaging.* Table III shows the results of the SAVEE dataset when the MSPs of individual face regions are averaged. We tested the system using the parameter sets chosen over the set of {0.1, 0.25, 0.5, 1, 1.5, 2} seconds for window lengths and over the set of {0, 6, 10, 20} for $MDL_{advantage}$ of MDL segments. We present the parameter sets averaged over the all four test speakers in Table VII and will discuss the interpretation of these results in Section 6. We also tested the system using phoneme segments (average segment length of 0.14 seconds), and utterance-length segments (average utterance length of 3.84 seconds).

In the SAVEE dataset, AV 6 outperforms the other methods of averaging different facial regions. For the AV 6 experiment results, we found that VG classification is more accurate than Gen classification for variable-length segmentation. The performance increases, comparing Gen and VG classification, for both MDL segmentation (75.62% to 80.00%, $p < 0.05$) and Phoneme segmentation (75.42% to 79.59%, $p < 0.05$). The

Table III.
SAVEE average classification accuracy (%) using six schemes of: (1) VG/Win, (2) VG/MDL, (3) VG/Phon, (4) Gen/Win, (5) Gen/MDL, (6) Gen/Phon, and (7) Gen/Utt (utterance-length) segments, using the averaging method. The results are presented as mean over the 10 speakers. '*' indicates a significant increase compared to the baseline window segmentation method.

| Cla | Seg | AV4 | AV6 | AV up | AV low | Chin | FH | CHK | U.EYE | EB | MOU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VG | Win | 71.67 | **77.29** | 72.92 | 63.96 | 39.79 | 59.79 | 64.38 | 55.21 | 72.92 | 65.42 |
| | MDL | 69.79 | **80.00*** | 75.00* | 63.13 | 37.50 | 58.13 | 63.33 | 56.46 | 77.08 | 65.63 |
| | Phon | 69.59 | **79.59** | 74.79 | 64.59 | 32.08 | 62.08 | 61.67 | 57.50 | 73.13 | 62.71 |
| Gen | Win | 71.04 | **77.29** | 68.54 | 66.25 | 38.33 | 55.00 | 65.83 | 54.79 | 64.79 | 63.96 |
| | MDL | 71.46 | **75.62** | 62.29 | 65.21 | 43.75* | 54.58 | 61.67 | 56.04 | 63.96 | 62.08 |
| | Phon | 72.71 | **75.42** | 61.88 | 68.34 | 44.80 | 56.67 | 63.55 | 55.42 | 63.54 | 66.88 |
| | Utt | 69.17 | 79.38 | 76.46 | 61.04 | 35.21 | 56.04 | 65.63 | 56.46 | **79.58*** | 58.33 |

window segmentation does not demonstrate any improvement (both methods demonstrate an accuracy of 77.29%).

The results demonstrate that the accuracy increases when variable-length segmentation is used in place of fixed-length segmentation in viseme-group classification, also shown in our previous work [Kim and Mower Provost 2014]. In particular, for the AV 6 experiment, the MDL (80.00%) and phoneme (79.59%) segments outperform the window segments (77.29%). The performance improvement of MDL over window segments is statistically significant ($p < 0.02$), whereas phoneme over window segments is not ($p = 0.29$). Moreover, we achieve comparable accuracy ($p = 0.190$) between our proposed MDL segmentation and phoneme segmentation. MDL significantly outperforms window segments in the AV up experiment, achieving 75.00% compared to 72.92% for window segmentation ($p < 0.05$). The VG/MDL method also achieves improvement compared to VG/Win for the eyebrow (EB), achieving 4.16% improvement ($p = 0.07$). The results provide evidence that MDL segmentation can be effectively used in emotion classification.

In Gen classification, the best results of MDL segments (75.62%, $p = 0.071$) and phoneme segments (75.42%, $p = 0.071$) are lower than that of window segments (77.29%), although the results are not significantly different. The highest accuracy is achieved with utterance-length segments (79.38%) for the AV 6 experiment. However, this phenomenon is not consistent over the different facial regions. For instance, in the mouth region, phoneme segments (66.88%) outperform window (63.96%) and utterance-length segments (58.33%), whereas MDL segments (62.08%) work slightly worse than window segments. For the chin, both variable-length segmentation strategies, MDL (43.75%) and phoneme (44.80%), outperform fixed-length segments, both window (38.33%) and utterance-length (35.21%) segments. The difference between the MDL and window segments was significant ($p < 0.05$). For the eyebrow, the utterance-length segments achieve significant increase compared to the other segmentation methods, achieving 79.58% accuracy. Overall, phoneme segments perform well for the lower facial regions, mouth and chin, whereas utterance-length segments perform well for regions less modulated by speech, such as the eyebrow.

*5.1.2. Weighting Based on Cross-Validation Accuracy.* Table IV demonstrates the SAVEE results when we weight the face region-specific MSPs based on validation accuracy. We found that weighting face region-specific MSPs lowered the accuracy of SAVEE (the opposite trend can be observed for IEMOCAP), although the decrease is not significant. VG/Win remains the same 77.29% accuracy, whereas VG/MDL accuracy decreases from 80.00% to 76.46%. For Gen classification, Gen/Win decreases from 77.29% to 75.00% and Gen/MDL decreases from 75.62% to 75.00%. We hypothesize that this is due to the high variability between speakers (e.g., one speaker has a significantly

Table IV.
SAVEE dataset results found by weighting the individual facial region based on validation accuracy (left) and (2) based on a linear SVM (right). We report the average recall for each of the four emotion classes and the overall average recall ("average").

| | | Weighting Based on Validation Accuracy | | | | | SVM Weighting | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cla | Seg | Average | Ang | Hap | Neu | Sad | Average | Ang | Hap | Neu | Sad |
| VG | Win | **77.29** | 81.67 | 76.67 | 76.67 | 88.33 | **88.75** | 80.00 | 95.00 | 88.34 | 91.67 |
| | MDL | **76.46** | 83.33 | 71.67 | 71.67 | 85.00 | **92.08*** | 93.33 | 90.00 | 90.00 | 95.00 |
| | Phon | **77.71** | 83.33 | 76.67 | 62.50 | 88.33 | **93.12*** | 95.00 | 93.33 | 89.17 | 95.00 |
| Gen | Win | **75.00** | 86.67 | 73.33 | 73.33 | 88.34 | **88.75** | 88.34 | 90.00 | 80.00 | 96.67 |
| | MDL | **75.00** | 93.34 | 66.67 | 66.67 | 90.00 | **88.13** | 95.00 | 81.67 | 79.17 | 96.67 |
| | Phon | **77.08** | 91.67 | 70.00 | 55.00 | 91.67 | **88.96** | 91.67 | 88.33 | 79.17 | 96.67 |

lower recognition rate, with a relative difference of about 20% from the other three speakers), and the lack of training speakers when calculating validation accuracy (i.e., only two speakers for training in cross validation). The per-emotion class accuracies demonstrate that *Anger* ($p < 0.05$, significant) and *Sadness* (not significant, $p = 0.052$) are well recognized compared to *Happiness* and *Neutrality*. The phoneme segments perform well in both VG (77.71%) and Gen (77.08%) classification, showing the highest performance among the three segmentation methods.

*5.1.3. SVM-Based Weighting Method.* Table IV demonstrates the results of linear-SVM based MSP combination. The hyper-parameter $C$ of the SVM is chosen as $10^{-4}$ using cross validation. It is shown that the results are improved using linear-SVM, achieving up to 92.08% accuracy for VG/MDL, improving from 80.00% of the simple averaging method. This is a significant improvement in accuracy over VG/Win ($p < 0.03$). VG/Win, Gen/Win, and Gen/MDL also improve from 77.29% to 88.75%, 77.29% to 88.75%, and 75.62% to 88.13%, respectively. The phoneme segments perform the best for both VG (93.12%) and Gen (88.96%) classification. VG/Phon outperforms VG/Win significantly ($p < 0.007$). The per-emotion class accuracies show improved performance for *Happiness* and *Neutrality*. We present the learned SVM weights in Figure 7 and will discuss the corresponding findings in the previous psychology studies on emotion perception in Section 6.

We hypothesize that the SVM-based weighting method more reliably captures the region-specific temporal characteristics compared to the weighting based on validation accuracy, since the SVM learns more general patterns across training speakers that are associated with emotion prediction compared to the direct validation accuracy. We discuss the learned SVM weights for each emotion prediction task in more detail in Section 6.

### 5.2. IEMOCAP

*Significance Tests.* For the IEMOCAP dataset, we use paired t-tests to be consistent with previous work on this dataset [Marioooryad and Busso 2013; Kim and Mower Provost 2014]. The paired t-test for leave-one-speaker-out cross validation has shown to be useful to test the significance of the difference [Dietterich 1998; Kim and Mower Provost 2014]. We claim significance when the p-value is less than 0.05.

*5.2.1. Averaging Method.* Table V summarizes the average accuracy for each of the 10 different experiments. As in the SAVEE dataset, the two parameters, window length and $MDL_{Advantage}$, are chosen over the set of $\{0.1, 0.25, 0.5, 1, 1.5, 2\}$ seconds and $\{0, 6, 10, 20\}$. In VG classification, variable-length segments outperform window segments in the AV 4 experiment and AV low experiments. MDL segments outperform window segments in most of the experiments except for the upper face regions

Table V.
IEMOCAP average classification accuracy (%) using six schemes of: (1) VG/Win, (2) VG/MDL, (3) VG/Phon, (4) Gen/Win, (5) Gen/MDL, (6) Gen/Phon, and (7) Gen/Utt (utterance-length) segments, using the averaging method. The results are presented as mean over the 10 speakers. '*' indicates a significant increase compared to the baseline window segmentation method.

| Cla | Seg | AV4 | AV6 | AV up | AV low | Chin | FH | CHK | U.EYE | EB | MOU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VG | Win | 54.30 | **54.93** | 47.14 | 52.44 | 50.36 | 42.92 | 43.24 | 45.21 | 42.93 | 53.86 |
| | MDL | 55.31 | 55.07 | 45.65 | 55.08* | 53.28* | 42.43 | 44.53* | 46.04 | 42.16 | **55.46** |
| | Phon | **56.14** | 55.60 | 45.85 | 54.05 | 49.96 | 40.43 | 44.92* | 45.07 | 40.14 | 56.07 |
| Gen | Win | 54.20 | **54.24** | 47.01 | 52.54 | 49.70 | 39.43 | 42.49 | 44.15 | 41.95 | 53.04 |
| | MDL | **53.51** | 51.92 | 42.39 | 51.56 | 46.94 | 38.97 | 42.56 | 42.33 | 38.33 | 52.04 |
| | Phon | **54.04** | 52.05 | 42.30 | 51.77 | 46.94 | 39.00 | 42.71 | 43.28 | 38.58 | 51.13 |
| | Utt | 40.02 | **40.83** | 39.70 | 40.02 | 35.51 | 39.70 | 25.00 | 25.00 | 25.00 | 37.10 |

Table VI.
IEMOCAP dataset results found by weighting the individual facial region based on validation accuracy (left) and (2) based on a linear SVM (right). We report the average recall for each of the four emotion classes and the overall average recall ("average").

| | | Weighting Based on Validation Accuracy | | | | | SVM Weighting | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cla | Seg | Average | Ang | Hap | Neu | Sad | Average | Ang | Hap | Neu | Sad |
| VG | Win | **56.66** | 57.59 | 69.83 | 33.19 | 66.04 | **56.06** | 66.16 | 77.28 | 16.58 | 64.23 |
| | MDL | **57.57** | 63.03 | 69.71 | 34.41 | 63.13 | **56.63** | 68.30 | 76.18 | 15.35 | 66.68 |
| | Phon | **57.18** | 61.76 | 68.69 | 40.39 | 57.90 | **55.06** | 67.66 | 74.94 | 17.48 | 60.16 |
| Gen | Win | **56.02** | 55.70 | 70.23 | 33.51 | 64.62 | **53.98** | 63.47 | 77.74 | 13.75 | 60.97 |
| | MDL | **55.00** | 59.93 | 69.48 | 34.08 | 56.48 | **53.96** | 62.12 | 76.52 | 15.85 | 61.35 |
| | Phon | **53.76** | 58.14 | 69.96 | 37.71 | 49.25 | **52.22** | 62.49 | 76.97 | 16.12 | 53.28 |

(the exceptions in upper face regions are not significant). In particular, variable-length segments outperform window segments significantly particularly in the lower face regions. In the AV low experiment, the performance gain when using VG/MDL compared to VG/Win is significant, achieving 2.64% ($p < 0.05$). The chin and cheek regions also showed significant increase compared to the window segments. For the chin, VG/MDL significantly outperforms VG/Win by 2.92% ($p < 0.05$). For the cheek, both VG/MDL and VG/Phon significantly outperform VG/Win by 1.29% and 1.68%, respectively (both $p < 0.05$). The significant improvement in the lower regions of the face when using the MDL segmentation may indicate that the mouth-based segmentation strategy of MDL performs well for the regions that are modulated by speech [Chandrasekaran et al. 2009]. VG/Phon outperforms VG/Win by 1.84% in the AV4 experiment. VG/MDL also outperforms VG/Win by 0.74%. However, these differences are not statistically significant. Moreover, the mouth region achieves higher accuracy (55.46%) than the AV6 method for window segments (54.93%), although not significant.

In Gen classification, the accuracy between different segmentation methods was similar. Also, the utterance-length segmentation performed poorly (40.83%), unlike in the SAVEE dataset.

*5.2.2. Weighting Using Validation Accuracy .* The weighting method that combines the individual facial regions based on cross validation accuracy improves the performance, up to 57.57% when using the VG/MDL method. This is the highest accuracy in the IEMOCAP dataset and it outperforms the simple averaging method in the AV6 experiment by 1.74% (not significant). This is higher than VG/Win method (56.66%), however the difference is not significant. The VG/MDL result is higher than both of the previous work [Mariooryad and Busso 2013; Metallinou et al. 2010]. VG/MDL and VG/Win perform significantly better using the validation accuracy-based weighting method

Table VII.
A description of the selected parameters for the SAVEE dataset based on leave-one-training-speaker-out cross validation and averaged over all folds. For the Win segmentation method, the parameter is segment length of each window and the for MDL segmentation method, the parameter is $MDL_{advantage}$, described in Section 4.2. A larger $MDL_{advantage}$ corresponds to longer average segment length. Note that phoneme segmentation methods do not have any parameters that control granularity.

| Cla | Seg | AV4 | AV6 | AV up | AV low | Chin | FH | CHK | U.EYE | EB | MOU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| VG | Win | 0.18 | 0.21 | 0.14 | 0.34 | 0.18 | 0.33 | 0.14 | 0.10 | 0.53 | 0.18 |
| | MDL | 6 | 5 | 6.5 | 6.5 | 11.5 | 2.5 | 10 | 6 | 5 | 10 |
| Gen | Win | 0.28 | 0.81 | 1.28 | 0.61 | 0.43 | 0.59 | 0.10 | 0.68 | 1.15 | 0.14 |
| | MDL | 4.5 | 14 | 15 | 11.5 | 5 | 6.5 | 11.5 | 9 | 10.5 | 12.5 |

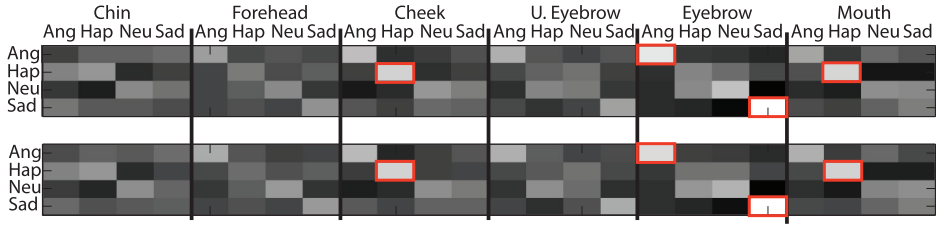


Fig. 7. The averaged learned SVM weights for the SAVEE dataset describing how to combine the MSPs of the individual facial regions (top: weights for Gen/Win and bottom: weights for VG/Win). Darker boxes correspond to smaller weights and brighter boxes correspond to larger weights. The boxes that are highlighted in red are the four highest weights.

compared to the simple averaging method (55.07% to 57.57%, $p < 0.05$; and 54.93% to 56.66%, $p < 0.05$). The Gen classification is also improved compared to the simple averaging method, particularly for Gen/MDL. Gen/MDL improves from 51.92% to 55.00% in the AV6 experiment. The average accuracies (Gen/Win 56.02% and Gen/MDL 55.00%) are smaller than seen in the VG classification results. The phoneme segments perform better for VG classification (57.18%) compared to Gen classification (53.76%), as in the other experiments. Gen/Win performs significantly better than Gen/Phon ($p < 0.05$).

*5.2.3. SVM-Based Weighting Method.* The hyper-parameter of SVM chosen based on cross validation was $10^{-5}$. Linear-SVM slightly improves the accuracy for VG/MDL compared to AV6 in the simple averaging method (55.07%). It also slightly improves the VG/Win accuracy in the AV6 experiment (54.93%). For Gen classification, the accuracy was slightly lower than the averaging method. The differences are not significant.

For the VG classification, MDL segmentation achieves significantly higher accuracy compared to the simple averaging method (55.07% to 56.63%, $p < 0.05$). VG/Win also performs significantly better using the SVM weighting method compared to the simple averaging method (54.93% to 56.05%, $p < 0.05$).

## 6. DISCUSSION

Table VII shows the parameters selected for the SAVEE dataset, averaged over all four test speakers {1, 2, 3, 4}. Note that larger values of the $MDL_{advantage}$ parameter corresponds to longer average segment lengths. The parameters chosen for each facial region demonstrate that in general classification, the upper facial regions such as the eyebrow and forehead perform well with longer segments. This trend may indicate that the upper facial regions may be characterized by longer-term dynamic patterns.

Figure 7 shows the trained SVM weights based on the 24-dimensional features for each face/emotion set for the SAVEE dataset. In the figure, there are six different facial regions and four different emotion classes. We investigate the contribution of

Table VIII.
Accuracy result comparisons of Gen/Win (top) and VG/Win (bottom) with different window sizes from the set {0.1, 0.25, 0.5, 1, 1.5, 2} seconds (IEMOCAP).

| | Gen/Win | | | PS/Win | | |
|---|---|---|---|---|---|---|
| Window Size (sec) | AV Mou | AV4 | AV6 | AV Mou | AV4 | AV6 |
| 0.1 | 52.41 | 55.51 | 52.93 | 54.24 | 55.06 | 55.03 |
| 0.25 | 52.04 | 55.04 | 52.49 | 52.63 | 54.83 | 54.92 |
| 0.5 | 52.09 | 55.24 | 53.78 | 52.41 | 55.04 | 55.19 |
| 1 | 50.79 | 53.54 | 54.29 | 52.12 | 53.41 | 53.98 |
| 1.5 | 50.96 | 54.36 | 54.13 | 51.45 | 53.95 | 53.8 |
| 2 | 50.12 | 53.11 | 52.11 | 50.49 | 51.72 | 51.75 |

each facial region to the final emotion inference. We estimate the contribution based on the SVM weights, e.g., $w_1 face_1 + w_2 face_2 + w_3 face_3 + w_4 face_4 + w_5 face_5 + w_6 face_6$. The weights $w_i$ of each face region $i = \{1, 2, \ldots, 6\}$ are averaged over the four test speakers. We find that in both Gen and VG classification, the mouth regions have higher weights on the happiness component of the four-dimensional MSPs, whereas eyebrow regions have higher weights on the anger and sadness components. This finding corresponds to the previous emotion perception studies that certain facial regions contribute more to specific emotion perception [Shah et al. 2013]. The studies on facial Action Units (AUs, [Ekman and Friesen 1977]) have shown that certain anatomical regions of the face, or action units, are strongly related to specific emotions. These studies have shown that happiness is strongly related to action units on the mouth (including action unit 6: cheek raiser and action unit 12: lip corner puller), and anger is strongly related to action units on the eyebrow (including action unit 4: brow lowered and action unit 7: lid tightener).

For the IEMOCAP dataset, we observe similar performance between Gen/Win, Gen/MDL, and Gen/Phon. This may imply that without any additional phoneme information it is important to capture longer-term dynamics to understand emotion expression. In addition, we compare the Gen and VG classification accuracies of the fixed-length segments with different window sizes, {0.1, 0.25, 0.5, 1, 1.5, 2} seconds. Table VIII summarizes Gen/Win (top) and VG/Win (bottom) accuracy for different window sizes. The Gen/Win accuracy shows statistically insignificant changes across different window sizes. However, the VG/Win accuracy shows significant increase in AV 6 accuracy between 1 and 2 seconds (2.18%, $p < 0.02$) and between 1.5 and 2 seconds (2.02%, $p < 0.03$). Both VG/Win and Gen/Win perform poorly with 2-second windows. This decrease in performance of the 2-second windows compared to smaller window sizes is higher for VG/Win compared to Gen/Win, which may suggest that in VG classification it is critical to use segments that have similar lengths to phoneme segments.

We also compare how many phonemes in each window segment with different sizes. For each window size of {0.1, 0.25, 0.5, 1, 1.5, 2} seconds, the average number of phonemes are $1.71 \pm 0.80$, $2.67 \pm 1.49$, $3.87 \pm 2.23$, $5.47 \pm 2.93$, $6.54 \pm 3.24$, and $7.19 \pm 3.47$.

The consistent increase in accuracy of VG/Win associated with an increase in window length may imply that the window segments that contain phoneme at the closest (i.e. most overlap with phoneme segment) will perform better than the others.

## 7. CONCLUSIONS

In this study, we investigate an unsupervised, variable-length segmentation method for compensating for facial movement due to speech, to improve the performance of facial emotion recognition systems. We present detailed results on two different datasets and propose a combination strategy that can account for different temporal characteristics of different facial regions. Our segmentation method is based on the MDL principle. We

demonstrated that a hyper-parameter $MDL_{Advantage}$ can change the average segment lengths and how this impacts the system-level performance. Based on this finding, we show how we can combine different hyper-parameters chosen per face regions using cross validation for final emotion inference. We use linear-kernel SVMs to combine facial region-specific emotion evidence and investigate the weights between facial regions and emotions to explore the different contributions of individual facial regions for inference of specific emotion.

Our experimental results on the two IEMOCAP and SAVEE datasets demonstrate that the two variable-length segmentation methods, MDL and phoneme, achieve higher emotion classification rates compared to fixed-length window segmentation in VG classification. We also find that methods to combine estimated emotion from individual face regions, can increase the accuracy significantly.

In our future work, we will investigate the efficacy of MDL segmentation based on different facial regions. In our preliminary study [Kim and Mower Provost 2014], we found that it is more beneficial to use the mouth region for MDL segmentation, compared to other facial regions. However, it is not yet clear whether this is true for other datasets such as SAVEE. For instance, we found that the SAVEE dataset shows high accuracy using utterance-level eyebrow segments, unlike the chance-level accuracy for the IEMOCAP dataset. This may indicate that the difference between the read speech (SAVEE) and more natural dynamic conversation (IEMOCAP) may have different facial movement characteristics. We will investigate the use of other facial regions for the other datasets in MDL segmentation.

Further, our results indicate that different segmentation strategies per different face regions, for instance, MDL segmentation for mouth region and emphasis-based segmentation for eyebrow region, may benefit the overall facial emotion recognition systems. In addition, we plan to combine the facial emotion recognition system that we developed with audio emotion recognition systems.

The presented work utilizes three-dimensional facial point positions to explicitly model physical facial movement. Previous studies have shown that facial appearance features are useful in emotion classification tasks. However, the IEMOCAP dataset does not provide frontal face videos and the faces in the SAVEE dataset are marked with blue measurement points. Action Units (AU) are also commonly used in facial emotion modeling research. AUs capture the presence of specific types of movement associated with groups of muscles, rather than movement of facial points in a three-dimensional space. Future work will explore modeling the temporal patterns associated with emotionally-relevant AUs and the use of appearance features.

## REFERENCES

Barry Arons. 1994. Pitch-based emphasis detection for segmenting speech recordings. In *Proceedings of the International Conference on Spoken Language Processing*. 1931–1934.

Douglas Bates, Martin Maechler, and Ben Bolker. 2007. lme4: Linear mixed-effects models using S4 classes (R package version 0.9975-11).

Elisabetta Bevacqua and Catherine Pelachaud. 2004. Expressive audio-visual speech. *Comput. Anim. Virtual Worlds* 15, 3–4, 297–304.

Subhabrata Bhattacharya, Behnaz Nojavanasghari, Tao Chen, Dong Liu, Shih-Fu Chang, and Mubarak Shah. 2013. Towards a comprehensive computational model for aesthetic assessment of videos. In *Proceedings of the 21st ACM International Conference on Multimedia*. ACM, 361–364.

Benjamin Bigot, Isabelle Ferrane, and Z. Ibrahim. 2008. Towards the detection and the characterization of conversational speech zones in audiovisual documents. In *Proceedings of the International Workshop on Content-Based Multimedia Indexing (CBMI'08)*. IEEE, 162–169.

Michael J. Black and Yaser Yacoob. 1997. Recognizing facial expressions in image sequences using local parameterized models of image motion. *Int. J. Comput. Vision* 25, 1, 23–48.

Marisa Boston, John Hale, Reinhold Kliegl, Umesh Patil, and Shravan Vasishth. 2008. Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. The Mind Research Repository (beta) 1.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resources Eval.* 42, 4, 335–359.

Carlos Busso and Shrikanth S. Narayanan. 2007. Interrelation between speech and facial gestures in emotional utterances: a single subject study. *IEEE Trans. Audio Speech Lang. Process.* 15, 8, 2331–2347.

Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Trans. Affective Computing* 1, 1, 18–37.

Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* 1.

Chandramouli Chandrasekaran, Andrea Trubanova, Sébastien Stillittano, Alice Caplier, and Asif A. Ghazanfar. 2009. The natural statistics of audiovisual speech. *PLoS Comput. Biol.* 5, 7, 1000436.

Jingying Chen, Maylor K. Leung, and Yongsheng Gao. 2003. Noisy logo recognition using line segment Hausdorff distance. *Pattern Recog.* 36, 4, 943–955.

Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. 2014. Object-based visual sentiment concept analysis and application. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 367–376.

Abhinav Dhall, Akshay Asthana, Roland Goecke, and Tom Gedeon. 2011. Emotion recognition using PHOG and LPQ features. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG'11)*. IEEE, 878–883.

Thomas G. Dietterich. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput.* 10, 7, 1895–1923.

Paul Ekman and Wallace V. Friesen. 1977. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA.

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recog.* 44, 3, 572–587.

Vipul Garg, Harsh Kumar, and Rohit Sinha. 2013. Speech based Emotion Recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers. In *Proceedings of the National Conference on Communications*. IEEE, 1–5.

Davood Gharavian, Mansour Sheikhan, Alireza Nazerieh, and Sahar Garoucy. 2012. Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network. *Neural Comput. Appl.* 21, 8, 2115–2126.

Ben Gold, Nelson Morgan, and Dan Ellis. 2011. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. John Wiley & Sons.

Hatice Gunes, Björn Schuller, Maja Pantic, and Roddy Cowie. 2011. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*. IEEE, 827–834.

Sanaul Haq and Philip J. B. Jackson. 2010. *Machine Audition: Principles, Algorithms and Systems*. IGI Global, Hershey PA, Chapter Multimodal emotion recognition, 398–423.

M. Sazzad Hussain, Sidney K. D'Mello, and Rafael A. Calvo. 2014. 25 Research and development tools in affective computing. In *The Oxford Handbook of Affective Computing*, 349.

Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. 2014. Predicting viewer perceived emotions in animated GIFs. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 213–216.

Markus Kächele, Michael Glodek, Dimitrij Zharkov, Sascha Meudt, and Friedhelm Schwenker. 2014. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. *Depression* 1, 1.

Ozlem Kalinli. 2012. Automatic phoneme segmentation using auditory attention features. In *Proceedings of INTERSPEECH*.

Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2005. Phoneme alignment based on discriminative learning. http://u.cs.biu.ac.il/~jkeshet/papers/KeshetShSiCh05.pdf.

Yelin Kim and Emily Mower Provost. 2014. Say Cheese vs. smile: Reducing speech-related variability for facial emotion recognition. In *Proceedings of the ACM International Conference on Multimedia (ACM MM'14)*.

Michael Kipp and J.-C. Martin. 2009. Gesture and emotion: Can basic gestural form features discriminate emotions? In *Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII'09)*. IEEE, 1–8.

Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2013. Affective body expression perception and recognition: A survey. *IEEE Trans. Affective Computing* 4, 1, 15–33.

Shashidhar G. Koolagudi, Nitin Kumar, and K. Sreenivasa Rao. 2011. Speech emotion recognition using segmental level prosodic analysis. In *Proceedings of the International Conference on Devices and Communications*. IEEE, 1–5.

Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Emotion recognition using a hierarchical binary decision tree approach. In *Proceedings of INTERSPEECH*. 320–323.

Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2011. Emotion recognition using a hierarchical binary decision tree approach. *Speech Commun.* 53, 9, 1162–1171.

Chul Min Lee and Shrikanth S. Narayanan. 2005. Toward detecting emotions in spoken dialogs. *IEEE Trans. Speech Audio Process.* 13, 2, 293–303.

Chul Min Lee, Serdar Yildirim, Murtaza Bulut, Abe Kazemzadeh, Carlos Busso, Zhigang Deng, Sungbok Lee, and Shrikanth Narayanan. 2004. Emotion recognition based on phoneme classes. In *Proceedings of INTERSPEECH*. 205–211.

Jae-Gil Lee, Jiawei Han, and Kyu-Young Whang. 2007. Trajectory clustering: a partition-and-group framework. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 593–604.

Patrick Lucey, Terrence Martin, and Sridha Sridharan. 2004. Confusability of phonemes grouped according to their viseme classes in noisy environments. In *Proceedings of the Australian International Conference on Speech Science & Technology*. 265–270.

Soroosh Mariooryad and Carlos Busso. 2013. Feature and model level compensation of lexical content for facial emotion recognition. In *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition (FG'13)*. DOI:http://dx.doi.org/10.1109/FG.2013.6553752

Hongying Meng and Nadia Bianchi-Berthouze. 2011. Naturalistic affective expression classification by a multi-stage approach based on hidden Markov models. In *Affective Computing and Intelligent Interaction*, Springer, 378–387.

Angeliki Metallinou, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2010. Visual emotion recognition using compact facial representations and viseme information. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE, 2474–2477.

Angeliki Metallinou, Athanasios Katsamanis, and Shrikanth Narayanan. 2013. Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image Vision Comput.* 31, 2, 137–152.

Angeliki Metallinou, Martin Wollmer, Athanasios Katsamanis, Florian Eyben, Björn Schuller, and Shrikanth Narayanan. 2012. Context-sensitive learning for enhanced audiovisual emotion classification. *IEEE Trans. Affective Computing* 3, 2, 184–198.

Emily Mower, Maja J. Mataric, and Shrikanth Narayanan. 2009. Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. *IEEE Trans. Multimedia* 11, 5, 843–855.

Emily Mower, Maja J. Mataric, and Shrikanth Narayanan. 2011. A framework for automatic human emotion classification using emotion profiles. *IEEE Trans. Audio Speech Lang. Process.* 19, 5 (2011), 1057–1070.

Emily Mower and Shrikanth Narayanan. 2011. A hierarchical static-dynamic framework for emotion classification. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 2372–2375.

Emily Mower Provost. 2013. Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. 3682–3686.

Shrikanth Narayanan and Panayiotis G. Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proc. IEEE* 101, 5, 1203.

Jérémie Nicolle, Vincent Rapp, Kévin Bailly, Lionel Prevost, and Mohamed Chetouani. 2012. Robust continuous prediction of human emotions using multiscale dynamic cues. In *Proceedings of the ACM International Conference on Multimodal Interaction*. ACM, 501–508.

Maja Pantic and Marian Stewart Bartlett. 2007. Machine analysis of facial expressions. In *Face Recognition*, I-Tech Education and Publishing, Vienna, Austria, 377–416.

Yu Qiao, Naoya Shimomura, and Nobuaki Minematsu. 2008. Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 3989–3992.

Yong Rui, Anoop Gupta, and Alex Acero. 2000. Automatically extracting highlights for TV baseball programs. In *Proceedings of the 8th ACM International Conference on Multimedia*. ACM, 105–115.

Enrique Sánchez-Lozano, Paula Lopez-Otero, Laura Docio-Fernandez, Enrique Argones-Rúa, and José Luis Alba-Castro. 2013. Audiovisual three-level fusion for continuous estimation of Russell's emotion circumplex. In *Proceedings of the 3rd ACM International Workshop on Audio / Visual Emotion Challenge*. ACM, 31–40.

Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Daniel Rueckert. 2011. A dynamic approach to the recognition of 3d facial expressions and their temporal models. In *Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition and Workshops*. 406–413.

Arman Savran, Houwei Cao, Miraj Shah, Ani Nenkova, and Ragini Verma. 2012. Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering. In *Proceedings of the ACM International Conference on Multimodal Interaction*. 485–492.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Commun.* 53, 9, 1062–1087.

Björn Schuller and Gerhard Rigoll. 2006. Timing levels in segment-based speech emotion recognition. In *Proceedings of INTERSPEECH*. 1818–1821.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth Narayanan. 2013. Paralinguistics in speech and language: State-of-the-art and the challenge. *Computer Speech Lang.* 27, 1, 4–39.

Miraj Shah, David G. Cooper, Houwei Cao, Ruben C. Gur, Ani Nenkova, and Ragini Verma. 2013. Action Unit Models of Facial Expression of Emotion in the Presence of Speech. In *Proceedings of the Conference on Affective Computing and Intelligent Interaction*. IEEE, 49–54.

Caifeng Shan, Shaogang Gong, and Peter W. McOwan. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. *Image Vision Comput.* 27, 6, 803–816.

Doroteo Torre Toledano, Luis A. Hernández Gómez, and Luis Villarrubia Grande. 2003. Automatic phonetic segmentation. *IEEE Trans. Speech Audio Process.* 11, 6, 617–625.

Bogdan Vlasenko, Dmytro Prylipko, Ronald Böck, and Andreas Wendemuth. 2014. Modeling phonetic pattern variability in favor of the creation of robust emotion classifiers for real-life applications. *Computer Speech Lang.* 28, 2, 483–500.

Shaohua Wan and J. K. Aggarwal. 2014. Spontaneous facial expression recognition: A robust metric learning approach. *Pattern Recog.* 47, 5, 1859–1868.

Siqing Wu, Tiago H. Falk, and Wai-Yip Chan. 2011. Automatic speech emotion recognition using modulation spectral features. *Speech Commun.* 53, 5, 768–785.

Guoying Zhao and Matti Pietikainen. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 6, 915–928.

Received February 2015; revised March 2015, July 2015; accepted July 2015