ABSTRACT

Facial movement is modulated both by emotion and speech articulation. Facial emotion recognition systems aim to discriminate between emotions, while reducing the speech-related variability in facial cues. This aim is often achieved using two key features: (1) phoneme segmentation: facial cues are temporally divided into units with a single phoneme and (2) phoneme-specific classification: systems learn patterns associated with groups of visually similar phonemes (visemes), e.g., /P/, /B/, and /M/. In this work, we empirically compare the effects of different temporal segmentation and classification schemes for facial emotion recognition. We propose an unsupervised segmentation method that does not necessitate costly phonetic transcripts. We show that the proposed method bridges the accuracy gap between a traditional sliding window method and phoneme segmentation, achieving a statistically significant performance gain. We also demonstrate that the segments derived from the proposed unsupervised and phoneme segmentation strategies are similar to each other. This paper provides new insight into unsupervised facial motion segmentation and the impact of speech variability on emotion classification.

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

General Terms
Emotion

Keywords
emotion recognition; dynamics; dynamic time warping; phoneme; segmentation; viseme; facial expression; speech

1. INTRODUCTION

The expression of emotion is complex. It modulates facial behavior, vocal behavior, and body gestures. Emotion recognition systems must decode these modulations in order to gain insight into the underlying emotional message. However, this decoding process is challenging; behavior is often modulated by more than just emotion. Facial expressions are strongly affected by the articulation associated with speech production. Robust emotion recognition systems must differentiate speech-related articulation from emotion variation (e.g., differentiate someone saying "cheese" from smiling). In this paper we explore methods to model the temporal behavior of facial motion with the goal of mitigating speech variability, focusing on temporal segmentation and classification methods. The results suggest that proper segmentation is critical for emotion recognition.

One common method for constraining speech variation is by first segmenting the facial movement into temporal units with consistent patterns. Commonly, this segmentation is accomplished using known phoneme or viseme boundaries. We refer to this process as phoneme segmentation. The resulting segments are then grouped into categories with similar lip movements (e.g., /P/, /B/, /M/, see Table 1). Emotion models are trained for each visually similar phoneme group, a process we refer to as phoneme-specific classification. These two schemes have been used effectively in prior work [2, 14, 19, 20]. However, it is not yet clear how facial emotion recognition systems benefit from each of these components. Moreover, these phoneme-based paradigms are costly due to their reliance on detailed phonetic transcripts. In this paper we explore two unsupervised segmentation methods that do not require phonetic transcripts. We demonstrate that phoneme segmentation is more effective than fixed-length sliding window segmentation. We describe a new unsupervised segmentation strategy that bridges the gap in accuracy between phoneme segmentation and fixed-length sliding window segmentation. We also demonstrate that phoneme-specific classification can still be used given unsupervised segmentation by coarsely approximating the phoneme content present in each of the resulting segments.

Studies on variable-length segmentation and the utility of phoneme-specific classification have recently received attention in the emotion recognition field. Mariooryad and Busso studied lexically constrained facial emotion recogni-

---

1We refer to the process as phoneme segmentation when the vocal signal, rather than the facial movement, is used to segment the data. These audio-derived segmentation boundaries are applied to the facial movement.
In this paper we empirically compare phoneme segmentation and phoneme-specific classification. They introduced feature-level constraints, which normalized the facial cues based on the underlying phoneme, and model-level constraints, phoneme-specific classification [19]. Metallinou et al. [20] also proposed a method to integrate phoneme segmentation and phoneme-specific classification into facial emotion recognition systems. They first segmented the data into groups of visually similar phonemes (visemes) and found that the dynamics of these segments could be accurately captured using Hidden Markov Models. These methods demonstrate the benefits of phoneme segmentation and phoneme-specific classification. However, there remain open questions relating to how similar performance can be achieved absent a detailed phonetic transcript.

In this paper we empirically compare phoneme segmentation with two unsupervised segmentation methods. The first uses fixed-length sliding window segmentation and the second uses variable-length segmentation. In the fixed-length window approach, windows slide over an utterance, segmenting it at regular intervals. The challenge with this approach is that it may inappropriately disrupt important dynamics relating to emotion expression. We propose the second method, based on the TRACLUS algorithm [17], to address these challenges. TRACLUS is an unsupervised technique that segments an utterance into variable-length components based on the dynamics of the signal, mitigating the speech-related variability in facial movements by focusing on short-time patterns that operate at the time-scale of phoneme expression. Figure 1 describes our system: the system segments the facial motion capture data using one of three segmentation methods (phoneme, TRACLUS, fixed window), classifies the emotion content of these segments, and then aggregates the segment-level classification into an utterance-level estimate. We use Dynamic Time Warping (DTW) to compare the similarity between testing segments and held out training segments. We additionally compare the advantages associated with phoneme-specific classification. We perform phoneme-specific classification by comparing a testing segment only to training segments from the same phoneme class. We perform general classification by comparing the testing segment to all training segments. We create a segment-level Motion Similarity Profile (MSP), which is a four-dimensional measure that captures the distribution of the emotion class labels over the closest training segments (class labels are from the set \{angry, happy, neutral, sad\}). We estimate the final emotion label by calculating an average MSP using the set of segment-level MSPs associated with a single utterance.

We present a novel system that explores the importance of variable-length segmentation, segmentation focused on the natural dynamics of the signal, and the benefit derived from phoneme-specific classification, an approach that restricts the impact of the modulations in the facial channel due to associated acoustic variability. Our results demonstrate that variable-length segmentation and phoneme-specific classification provide complementary improvements in classification. We show that variable-length segmentation and phoneme-specific classification can improve performance over fixed-length sliding window segmentation by 4.04%, compared to a system that uses a sliding window and general classification, results supported by prior research [19]. One of the advantages associated with our system is that it captures the dynamics of facial movement, rather than the behavior of static utterance-length features. This is critical for a detailed understanding of

---

**Table 1: Phoneme to viseme mapping.**

<table>
<thead>
<tr>
<th>Group</th>
<th>Phonemes</th>
<th>Group</th>
<th>Phonemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>P, B, M</td>
<td>8</td>
<td>AE, AW, EH, EY</td>
</tr>
<tr>
<td>2</td>
<td>F,V</td>
<td>9</td>
<td>AH, AX, AY</td>
</tr>
<tr>
<td>3</td>
<td>T,D,S,Z,TH,DH</td>
<td>10</td>
<td>AA</td>
</tr>
<tr>
<td>4</td>
<td>W,R</td>
<td>11</td>
<td>AXR, ER</td>
</tr>
<tr>
<td>5</td>
<td>CH,SH,ZH</td>
<td>12</td>
<td>AO, OY, OW</td>
</tr>
<tr>
<td>6</td>
<td>K,G,N,L,HH,NG,Y</td>
<td>13</td>
<td>CH, UW</td>
</tr>
<tr>
<td>7</td>
<td>Y, IH, IX</td>
<td>14</td>
<td>SIL</td>
</tr>
</tbody>
</table>
emotion expression behavior. The performance of our system suggests that our method can be used to provide insight into frame-level movement while providing accurate classification performance. Finally, our results demonstrate that the accuracy derived from a variable-length TRACLUS segmentation approach has higher accuracy than a fixed-length sliding window approach and comparable accuracy to a phoneme segmentation approach, suggesting that segmentation informed by the temporal properties of the signal is crucial for emotion classification systems.

This paper is organized as follows. In Section 2, we provide an overview of previous research. In Section 3, we describe the database. In Section 4, we detail the methodology underlying our approach. In Section 5, we describe experimental setup including the cross-validation scheme and performance measure. In Section 6, we present and analyze the results. We discuss our findings in Section 7. We conclude our work in Section 8.

2. RELATED WORK

Automatic emotion recognition is the process of predicting the high-level affective content of an utterance from the low-level signal cues produced by a speaker. Comprehensive surveys can be found in [9,11,30].

2.1 Segmentation

Temporal segmentation is commonly employed in speech emotion recognition. In the pioneering study of Lee et al., the authors designed and compared the standard emotion-specific Hidden Markov Models (HMM) and HMMs trained on individual phoneme groups for each emotion and found that vowel sounds were the most effective for emotion classification compared to the other four phoneme groups [16]. Batliner et al. treated words as the basic unit of emotion expression. They combined words either into syntactically and semantically meaningful chunks or into sequences of words that belonged to the same emotion class [1]. Jeon et al. investigated different sub-sentence segment units (words, phrases, time-based segments) using a two-level system that focused both on segment-level and utterance-level emotion prediction. They found that time-based segments achieved the best performance [13]. Schuller et al. also investigated different timing patterns for segmentation using absolute and relative time intervals. Utterances were either segmented at fixed time intervals (absolute) or at fixed relative positions such as halves or thirds (relative) [27]. They demonstrated that absolute time intervals of one second achieved the highest accuracy (also demonstrated in [21]). Additionally, they found that systems based on relative time intervals were more accurate than those that used absolute time intervals. Ringeval et al. proposed a speech feature extraction method based on a pseudo-phonetic speech segmentation technique combined with a vowel detector [24]. They compared MFCC acoustic features from these pseudo-phonetic segments (vowels, consonants) with segments created by identifying regions of voiced and unvoiced speech. They showed that the voiced segments could be modeled more accurately than the vowel or consonant segments for emotion recognition.

There have also been research efforts in temporal segmentation for facial expression recognition. As seen in audio modeling, these methods include phoneme-based segmentation [7,12] and the standard fixed-length and multiple fixed-length segmentation [21,23,26]. Cohen et al. proposed a multi-level HMM for the automatic segmentation and classification of facial expressions [7]. The proposed method automatically segments and classifies a video consisting of six sequences that display each of the six basic emotions (anger, disgust, fear, happiness, sadness, and surprise). The multi-level HMM makes an assumption that the transitions between emotions pass through the neutral state. They compare this with the standard emotion-specific HMM, where the input video is pre-segmented for each emotion. They found that the accuracy of their proposed system was similar to that of the standard emotion-specific HMM. Hoey used a manual segmentation of facial cues into segments presenting a subject’s underlying emotion [12]. He proposed a multi-level weakly supervised dynamic Bayesian network that learns the high-level dynamics of facial expressions.

There has also been work focused on unsupervised segmentation of facial expression data. Zhou et al. presented an unsupervised segmentation and clustering method of facial events [31]. They used $k$-means clustering with a Dynamic Time Alignment Kernel [29] to segment and cluster posed and unposed facial events. They found moderate inter-system agreement with the Facial Action Coding System (FACS). However, most of the previous works used facial data not modulated by spoken content, rendering it challenging to understand the impact of speech-related variability (a notable exception includes [31]). Our work is differentiated from the previous studies, by its focus on how we can better estimate emotion class by reducing the variability of facial movements caused by speech, while using unsupervised techniques.

2.2 Phoneme-Specific Classification

Phoneme-specific classification systems have been used to decrease speech-related variability and capture emotion-specific variability. These systems segment facial cues into phoneme segments and build phoneme-specific emotion classifiers over each phoneme group [19,20]. Marinho and Busso presented two different types of phoneme-specific classification: feature-level (each phoneme segment was normalized based on the underlying phoneme content) and model-level (phoneme-specific classification) [19]. They estimated utterance-level emotion content by aggregating the posterior probabilities of the phoneme-level estimates. They found that the model-level constrained system performed more accurately than the feature-level system and that both systems outperformed the phoneme-independent baseline. The model-level system used static features, statistics extracted within phoneme segments, in contrast to dynamic features, features that capture the dynamics of frame-level cues, as used in our paper. Metallinou et al. modeled the frame-level dynamics of facial motion-capture data using HMMs and estimated the phoneme-level emotion as a single hard label [20]. They estimated utterance-level emotion using a majority vote over the phoneme-level labels. The previous studies demonstrated the benefits of phoneme segmentation and phoneme-specific classification. However, an open question remains in how similar levels of accuracy can be achieved without knowledge of phoneme transcript.
3. DATABASE

In this work we use the IEMOCAP audio-visual emotion database [3], also used in [19, 20]. The IEMOCAP database contains audio-visual and motion capture recordings of interactions between five male-female pairs of actors, performing from plays and improvised scenarios. The motion capture data includes the three-dimensional motion of fifty-three markers on the face with sample rate of 120 Hz (Figure 2). Rotation effects were reduced by transforming the facial coordinate space such that the nose tip is always at the origin. The dataset also includes a detailed phonemic transcript.

We assign an emotion class to each utterance, or speech turn, as the majority vote over the three human annotators. The human annotators assign labels from the set of anger, sadness, happiness, disgust, fear and surprise, frustration, excited neutral, and other. For purposes of consistency with the literature [19, 20], we use data annotated as anger, happiness, sadness, and neutrality and merge the classes of excitement and happiness. There are an average of 61.1 ± 27.53 angry utterances, 122.4 ± 24.86 happy utterances, 58.1 ± 21.61 neutral utterances, and 64.4 ± 25.18 sad utterances per speaker. The mean length of an utterance is 4.73 ± 3.34 seconds.

4. PROPOSED METHODS

We present a system operating at two levels: segment-level and utterance-level. We test six classification schemes for segment-level classification that combine segmentation approaches (phoneme, TRACLUS, sliding window) and classification approaches (phoneme-specific or general). Phoneme-specific classification compares test segments only to training segments with the same phoneme content. General classification compares test segments to all training segments.

The classification schemes are as follows (Figure 3):

1. General phoneme ("Gen/Phon"): phoneme segmentation and general classification.
2. General TRACLUS ("Gen/TRA"): TRACLUS segmentation and general classification.
5. Phoneme-specific TRACLUS ("PS/TRA"): TRACLUS segmentation and phoneme-specific classification. A phoneme label is assigned by identifying the phoneme with the longest duration within each TRACLUS segment.
6. Phoneme-specific window ("PS/Win"): fixed-length window segmentation and phoneme-specific classification. Again, a phoneme label is assigned by identifying the phoneme with the longest duration within each window segment.

We aggregate the segment-level estimates to form an utterance-level estimate of emotion content. In the following sections, we first introduce the preprocessing that we apply to the motion capture data that serve as our input for segmentation (Section 4.1). We then describe our segmentation approaches (Section 4.2). Finally, we provide a description for how we estimate segment-level emotion and aggregate these estimates into utterance-level emotion labels (Section 4.3).

4.1 Motion Capture Preprocessing

We group the motion capture data points into six facial regions: chin, forehead, cheek, upper eyebrow, eyebrows, and mouth. There are 3 markers in both the Chin and Forehead regions, 16 markers in the Cheek region, and 8 markers in the Upper Eyebrow, Eyebrow, and Mouth regions as shown in Figure 2. This results in 9 (x, y, z)-coordinates for the Chin and Forehead regions, 48 coordinates for the Cheek region, and 24 coordinates for the Upper Eyebrow, Eyebrow, and Mouth regions, for a total of 138 marker coordinates. We perform median-filter smoothing using a window size of three over each of the six regions.

Each segment is represented by six different trajectories, one for each facial region. Each trajectory can be described by a N × K motion-capture trajectory, where N is the number of motion-capture frames within the segment and K is the number of 3-D marker coordinates in the face region. We exclude the five nose markers because of their limited movement and the two eyelid markers because they are often occluded, as in [20]. We also exclude data with less than seven frames (threshold number of frames were chosen empirically) or 0.058 seconds, 43.5% of all phoneme segments, 0.67% of all TRACLUS segments, and 0.86% of all fixed-length segments (the last fixed-length segments within an utterance is less than 0.25 seconds). We drop segments with any missing values in the 46 markers we use. This results in different sets of utterances for each segmentation scheme. Hence, we use the set of intersecting utterances for all experiments; fixed-length, TRACLUS, phoneme segmentation schemes. In total, we test our system using 3,060 utterances.

4.2 Segmentation

In this section we present the three temporal segmentation schemes that we use to segment the facial motion capture data. The sliding window and TRACLUS segmentation strategies do not require transcripts, while the phoneme segmentation strategy does require transcripts.
4.2.1 Sliding Window Segmentation (“Win”)

We obtain fixed-length segments using a sliding window of length 0.25 seconds without overlapping. We also include the last remaining segment at the end of an utterance, even if that segment is less than 0.25 seconds. In our initial experiments, we tested windows of length \{0.15, 0.25, 0.5, 1, 1.5, 2\} seconds, window sizes demonstrated to be effective in previous work [15,21]. We found that the differences in performance between window lengths of 0.15 seconds and 1.5 seconds were not statistically significantly different. Performance using the 2-second windows was significantly lower.

4.2.2 Phoneme Segmentation (“Phon”)

We obtain phoneme segments by force aligning the audio to the known transcript. We use the phoneme time boundaries to segment the motion-capture data into phoneme segments. We refer to the segments as phoneme segments, rather than viseme segments, to indicate that we are using the features of the audio to segment the data, rather than the features of the video. We group the phoneme segments into categories based on similar articulatory configurations (Table 1), as in [19,20].

4.2.3 TRACLUS Segmentation (“TRA”)

We obtain unsupervised variable-length segments using a trajectory partitioning algorithm called TRACLUS (Trajectory Clustering) [17]. TRACLUS takes a trajectory as input, automatically recognizes regions of consistent dynamics, and identifies characteristic points as the end-points for these regions (Figure 4). We treat the characteristic points as segment boundaries. Although the TRACLUS algorithm both partitions and clusters time series data, we only use the partition portion of the algorithm.

The input to TRACLUS segmentation is the smoothed 24-dimensional mouth trajectory using a median filter with a window size of three. We use the resulting characteristic points to segment the data from all six regions. We segment based only on the mouth trajectory because our preliminary results demonstrated that the accuracy of the classifiers operating on the mouth-derived segmentation boundaries were more accurate than those operating on boundaries derived from the remaining five regions.

TRACLUS identifies characteristic points using Minimum Description Length (MDL) coding. MDL is a widely used information theoretic concept positing that the best hypothesis \(H\) for given data \(D\) is the one that derives the best compression of that data, mitigating both the complexity associated with the compressed representation and the loss between the compression and the original data. The encoding of the data is a line connecting the first and last points (the characteristic points) of the region. The implication is that the variations between the first and last points represent noise around a period of consistent dynamics. The TRACLUS segments are formed by segmenting the motion capture data at the characteristic points.

The cost of approximating \(D\) as \(H\) is calculated in two parts: \(L(H)\) and \(L(D|H)\). \(L(H)\) is the length, in bits, of the description of the hypothesis, and \(L(D|H)\) is the length, in bits, of the description of the data when encoded by the hypothesis. The goal is to find the best hypothesis \(H\) to explain \(D\) that minimizes \(L(H) + L(D|H)\). Given an \(M\)-length trajectory \(TR = p_1p_2\cdots p_M\) and a set of \(n\) characteristic points \(\{c_1, c_2, \ldots, c_n\}\) (where \(c_1 = 1, c_n = M\)), let \(\text{len}(p_{c_j}p_{c_{j+1}})\) be the length of a line segment \(p_{c_j}, p_{c_{j+1}}\), i.e. the Euclidean distance between \(p_{c_j}\) and \(p_{c_{j+1}}\). The optimization method formula is described in equation 1. The values \(d_\perp\) and \(d_\parallel\) are the perpendicular distance and angular distance, adapted from a similarity measure using line segments [5]. Since our trajectory is multi-dimensional, we use multi-dimensional Euclidean distance between each trajectory point. Please see [17] for more details.
Figure 4: TRACLUS segmentation example. Each point \( \{p_1, p_2, \ldots, p_8\} \) represents frame-level mouth motion capture (illustrated as eight 2-D marker points for visualization purposes). In this example, TRACLUS identifies characteristic points \( \{c_1 = 1, c_2 = 5, c_3 = 8\} \). \( H_1 \) and \( H_2 \) are the lines connecting \( p_1 \) and \( p_8 \), and \( p_5 \) and \( p_8 \), respectively. The components of the cost associated with this segmentation can be calculated as the length of \( H_1 \) and the cost of approximating \( D_1 \) as \( H_1 \) and the length of \( H_2 \) and the cost of approximating \( D_2 \) as \( H_2 \).

### 4.3 Emotion Classification

#### 4.3.1 Emotion Similarity Calculation: DTW

The segment-level similarity between the motion capture data within each segment is calculated using DTW, a widely used time-series similarity measure [28]. Unlike HMM, DTW does not make any statistical assumptions about the temporal data, instead explicitly measuring temporal similarity, rather than probabilistic state transitions [15]. Multi-dimensional DTW measures the time-series similarity between two multi-dimensional trajectories with the same dimensionality (the lengths of the trajectories can be different). All of our calculations are multidimensional and we will refer to multi-dimensional DTW as DTW. DTW uses dynamic programming to calculate an optimal match between two time-series [25]. We compute a local cost matrix \( Q \), that contains distances between the two trajectories at every instance in time. For instance, as we described in Section 4.1, let two \( K \)-dimensional face movement trajectories of length \( M_1 \) and \( M_2 \) be \( T_1 \in \mathbb{R}^{M_1 \times K} \) and \( T_2 \in \mathbb{R}^{M_2 \times K} \). The multi-dimensional DTW algorithm populates the \( M_1 \)-by-\( M_2 \) local cost matrix \( Q \) according to the following equation:

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

where \( i \) and \( j \) represent the frame-level point of \( T_1 \) \( (1 \leq i \leq M_1) \) and \( T_2 \) \( (1 \leq j \leq M_2) \), respectively. Each local cost element of \( Q \) is an aggregate of the individual multi-dimensional distances between points in the trajectory calculated using the 2-norm, the sum of the squared differences across all dimensions. DTW then finds the optimal warping path that minimizes the accumulated distances in \( Q \). The accumulated distance along the optimal path is the final DTW distance. We use Sakoe-Chiba normalization, which adds 2 weights to the diagonal path when calculating the local distance cost and divides the final distance by the sum of the lengths of two trajectories [25].

#### 4.3.2 Use of Phoneme Knowledge

Phoneme-specific and general classification refer to whether or not knowledge of the phoneme content of a segment is used in the similarity calculation (Figure 3). In phoneme-specific classification, knowledge of phoneme content is used in the modeling process to constrain the number of training segments to which each testing segment is compared. Phoneme content is defined as the dominant phoneme in the segment. We group the phonemes into the fourteen groups outlined in Table 1, as used in previous work [18, 20]. For a given testing segment, we calculate the DTW distance from this testing segment to all other training segments within the phoneme group (Section 4.3.1). This allows us to reduce the lexical variability of the facial movement (Figure 5). In general classification, the testing segments are compared to all training segments, independent of phoneme group.
4.3.3 Utterance-Level Classification

The DTW calculation allows us to identify a set of training segments to which the testing segments are most similar. We retain the $c$ closest neighbors, using the MSP representation to describe the distribution of the emotion labels amongst the $c$ neighbors. This allows us to estimate the emotion content of the testing segment. For instance, if $c=10$ and the segment-level labels of the $c$ closest training segments include 3 angry, 6 happy, 1 neutral, and 0 sadness segments, then the MSP is $[0.3, 0.6, 0.1, 0]$. In our experiments, we set $c=20$, based on preliminary analyses.

For each facial region, we calculate the segment-level MSPs (e.g., $MSP_{mouth,1} \ldots MSP_{mouth,L}$). We create an utterance-level MSP by averaging the segment-level MSPs (e.g., $MSP_{mouth} = mean(MSP_{mouth,1} \ldots MSP_{mouth,L})$). This results in a single four-dimensional emotion estimate for each facial region. We average the six facial region MSPs to obtain the final utterance-level MSP. We normalize each of the four dimensions using speaker-specific z-normalization. We assign an utterance-level label based on the maximum of the aggregated MSP. This is based on the assumption that the largest component in the MSP corresponds to the class to which the utterance is most similar.

5. EXPERIMENTAL SETUP

Speaker normalization: We use the speaker normalization method used in [20] to mitigate the intrinsic differences in the facial configuration of the speakers. For each speaker $i$, we calculate the mean of each marker coordinate $j$ of a speaker ($m_{i,j}$). We then calculate $M_{j}$, the mean of each marker coordinate $j$ across all speakers. We then calculate the normalization coefficient, $c_{i,j}$, where $c_{i,j} = M_{j}/m_{i,j}$. In the calculation of $c_{i,j}$, $i$ represents each speaker from 1 to 10, and $j$ represents each marker coordinate from 1 to 138. We multiply each $i$-th speaker’s $j$-th marker coordinate by $c_{i,j}$ to make the mean of each speaker’s facial marker positions equal.

Cross-validation: We use a subject-independent training paradigm, leave-one-speaker-out cross validation (one speaker is used for testing, the other nine for training).

Performance measure: We include three experiments to understand the contribution and utility of each of the six facial regions: (1) AV mouth: the single mouth region, (2) AV 4 faces: four facial regions of chin, cheek, upper eyebrow, and mouth, excluding forehead and eyebrow, and (3) AV 6 faces: all six facial regions. Note that AV represents “average.”

In all cases, the reported accuracy is unweighted recall (UW). UW is an average of the per-emotion class recall. It is used to offset the class imbalance observed in the data. We calculate the UW for each speaker. We then average the UW over speakers and report the mean and standard deviation as the final UW.

6. RESULTS

The results demonstrate that phoneme-specific classification is more accurate than general classification for both the phoneme and TRACLUS segmentation strategies across all three experiments: AV mouth, AV 4, and AV 6. The AV mouth and AV 4 experiments have similar levels of accuracy. Both outperform the AV 6 experiment. General classification is more accurate for fixed-length sliding window segmentation in the AV 4 experiment.

Given phoneme-specific classification, variable-length segmentation (phoneme and TRACLUS) outperforms fixed-length sliding window segmentation for the AV mouth and AV 4 experiments, and the performance is comparable for the AV 6 experiment. This trend was not seen in general classification. In fact, in general classification, fixed-length sliding window segmentation outperforms variable-length segmentation in the AV mouth and AV 4 experiments.

The overall UW comparison is summarized in Table 2. The results are presented as the average unweighted recall over the ten subjects. We test the significance of the difference between the classification schemes using paired t-tests over the 10 subjects, suggested in [8]. In the remainder of this section, we provide additional details regarding the effect of segmentation, phoneme-specific classification, and statistical significance. The results will be discussed in terms of the AV mouth and AV 4 experiments.
6.1 Variable-length Segmentation

In this section we compare the UW of the system using fixed-length segments and variable-length segments (phoneme and TRACLUS) to understand the contribution of the segmentation.

The UW for PS/Phon is 56.14%, for PS/TRA is 56.18%, and PS/Win is 54.83%, all in the AV 4 experiment. This trend suggests that phoneme-specific classification benefits from meaningful segmentation. In the AV mouth experiment, the UW for PS/Phon is 56.07%, for PS/TRA is 56.08%, and for PS/Win is 56.63%. In this experiment, both PS/Phon and PS/TRA performed statistically significantly more accurately than PS/Win at p<0.05, providing additional evidence that phoneme-specific classification benefits from segmentation tied to the dynamics of the signal. PS/Phon uses the dynamics of the speech signal during segmentation and and PS/TRA uses the dynamics of the mouth movement signal during segmentation. The unweighted recall associated with the PS/Phon and PS/TRA schemes are not statistically significantly different (p<0.05). The importance of variable-length segmentation was also demonstrated in previous work [22]. However, this is the first demonstration of the efficacy of unsupervised variable-length segmentation for raw facial movements.

The highest UW for Gen/Phon is 54.04%, for Gen/TRA is 53.38%, and Gen/Win is 55.04%, all within the AV 4 experiment. In the AV mouth experiment, the UW for Gen/Phon is 51.13%, for Gen/TRA is 51.79%, and for Gen/Win is 52.04%. The differences between experiments and within experiments are not statistically significantly different. The trend in the results demonstrates that when using general classification (i.e., the phoneme content is not known) it is advantageous to model the dynamics of more than just the mouth region. It further demonstrates that absent knowledge of phoneme content, it may be beneficial to model longer-term dynamics (captured by windows of 0.25 seconds, compared to 0.17 and 0.16 second windows for phoneme and TRACLUS segmentation, respectively, see Section 6.3).

6.2 Phoneme-specific Classification

In this section we compare the UW of the system using phoneme-specific and general classification. This analysis allows us to understand the importance of constraining variability by limiting the content to which we compare testing segments.

We present the results in terms of improvements in performance, moving from general classification to phoneme-specific classification for each of the three segmentation schemes. We analyze the improvement in performance in terms of AV mouth and AV 4, respectively. The improvement for phoneme segmentation was 4.94% and 2.10%, for TRACLUS segments was 4.28% and 2.80%, and for sliding window segments was 0.59% and -0.21%. The improvements for both variable-length segmentation schemes are statistically significant at p<0.05 (for both AV mouth and AV 4). The performance was not statistically significantly different for fixed-length sliding window segmentation.

These results demonstrate that the knowledge of phoneme content within a segment contributes less when segmentation is not tied to the dynamics of the signal (compare fixed-length segments to variable-length phoneme and TRACLUS segments). Consider the results of the AV mouth experiment, the experiment in which the phoneme segmentation and TRACLUS segmentation strategies saw the greatest increase in performance. The experiment is unique, because it operates only on the mouth, the region of the face most influenced by lexical content. Phoneme segmentation explicitly controls for the dynamics introduced by lexical variability by segmenting at phoneme boundaries. TRACLUS segmentation identifies points at which dynamics change, capturing an approximation of this variability. Therefore, the improvement observed for variable-length segmentation emphasizes the importance of proper segmentation when integrating higher-level knowledge (phoneme content) that could account for fluctuations in the signal. It further suggests that given effective segmentation, emotion can be estimated from a smaller set of facial motion capture data (AV mouth vs. AV 4/6).

The improvement in classification, going from general to phoneme-specific classification (observed in all cases excepting general classification with window segmentation, AV 4 experiment), demonstrates the importance of removing variation due to speech-related articulation when considering the dynamics of frame-level facial movement.

6.3 Details of Segments

Both the fixed-length window and TRACLUS segments contain more than one phoneme on average. The fixed-length segments contain 2.67 ± 1.49 phonemes. The TRACLUS segments contain 2.13 ± 1.01 phonemes. The phoneme
segments are $0.17 \pm 0.01$ seconds in length, the TRACLUS segments are $0.16 \pm 0.01$ seconds in length, and the fixed-length sliding windows are 0.25 seconds in length.

7. DISCUSSION

We find that phoneme-specific classification with variable-length segmentation (phoneme and TRACLUS segmentation) performs more accurately than phoneme-specific classification with fixed-length sliding window segmentation. This suggests that when modeling emotion expressions it is critical to focus both on the lexical content of the signal and the underlying dynamics of the signal. Phoneme-specific classification with variable-length segmentation significantly outperforms general classification with sliding window segmentation in the AV mouth experiment (improvement of 4.03% and 4.04% for phoneme segmentation and TRACLUS segmentation, respectively, $p < 0.05$).

We find that phoneme-specific classification using the lower face regions (mouth, cheek, chin) performs more accurately than systems using the upper face regions (forehead, eyebrow, upper eyebrow) (Table 3). This finding aligns with the previous findings [19]. Further, we are able to extract the emotion information from the mouth region more accurately than from any other single face region. We hypothesize that the relative accuracy of the mouth-region classification is due to the lexical constraint that is placed on the modeling, we explicitly control for the variability that affects the mouth region (i.e., speech). The segmentation and phoneme-specific classification mitigates a large degree of the variability associated with speech. Therefore, the majority of the differences observed in the trajectories of the mouth region are tied to emotion or speaker variability. The other regions are not temporally constrained in the same manner. Research has demonstrated that eyebrow movements are related to prosodic patterns [4, 19]. However, these patterns are not generally time-synchronous with the production of phonemes.

In this paper we explicitly model the dynamics of the raw facial movement. This reliance on raw facial movement will provide new insight into emotionally salient facial behaviors and facial dynamics. Our current method performs comparably to the best result of Metallinou et al. [20], the previous state-of-the-art in raw facial movement modeling (we demonstrate an increase in performance of 0.44%). The presented PS/TRA scheme (phoneme-specific classification with TRACLUS segmentation) achieves comparable accuracy to their PS/Phon scheme. This suggests that frame-level dynamics can be comparably measured with unsupervised segmentation and a coarse knowledge of phoneme content, which has the potential to decrease the cost of transcription.

8. CONCLUSIONS

In this work, we investigate the advantages of variable-length segmentation and phoneme-specific classification for facial emotion recognition. We found that both variable-length segmentation and phoneme-specific classification improve overall performance. This highlights the importance of carefully considering segmentation when preprocessing for emotion classification.

However, phoneme segmentation requires the use of a transcript, which can be costly to obtain. The results demonstrated that absent knowledge of phoneme timing we could still improve performance over a fixed-length window baseline using TRACLUS, an unsupervised, automatic variable-length segmentation method. In phoneme-specific classification, the variable-length segmentation significantly outperforms the fixed-length segmentation when modeling the mouth data. This is promising because in real world feature extraction the mouth region can be extracted and tracked reliably (for example, [6, 10]).

In the future, we will investigate the utility of multiple variable-length segmentation techniques. The relatively low accuracy of the upper face regions suggests that the upper face regions may require longer time frames to capture the emotion-specific modulations (e.g., a frown of disapproval may last longer than phoneme-level time frames). Finally, our focus on the movement of raw feature points precludes the effects of facial structure. Future work will address modeling strategies that incorporate structural modeling. The presented study will be a building block for our future investigations on feature interaction at the different levels, such as sub-utterance-level, utterance-level, and dialog-level. We will also explore the inclusion of audio data based on pitch, energy, or frequency variations. Finally, we will explore the performance of our system on additional facial motion capture databases.

9. ACKNOWLEDGMENTS

We thank Dr. Donald Winsor and Laura Fink for offering computing resources. We thank Chansoo Lee and Yuan (June) Shangguan for their many helpful comments on this paper. We also wish to thank our anonymous reviewers for their insightful feedback.

10. REFERENCES


