EmoShapelets: Capturing Local Dynamics of Audio-visual Affective Speech

Yuan Shangguan
University of Michigan
Ann Arbor, Michigan 48109
Email: juneysg@umich.edu

Emily Mower Provost
University of Michigan
Ann Arbor, Michigan 48109
Email: emilykmp@umich.edu

***Abstract*—Automatic recognition of emotion in speech is an active area of research. One of the important open challenges relates to how the emotional characteristics of speech change in time. Past research has demonstrated the importance of capturing global dynamics (across an entire utterance) and local dynamics (within segments of an utterance). In this paper, we propose a novel concept, EmoShapelets, to capture the local dynamics in speech. EmoShapelets capture changes in emotion that occur within utterances. We propose a framework to generate, update, and select EmoShapelets. We also demonstrate the discriminative power of EmoShapelets by using them with various classifiers to achieve comparable results with the state-of-the-art systems on the IEMOCAP dataset. EmoShapelets can serve as basic units of emotion expression and provide additional evidence supporting the existence of local patterns of emotion underlying human communication.**

Keywords*—*emotion classification; time series; emotion representation; audio-visual multi-modal; local dynamics; emotion unit

I. Introduction and Related Works

Recent decades have witnessed an increase in the volume of research in classifying affective speech. The challenges of accurately recognizing the emotional content of speech lie not only in the complex relationships between the audio-visual speech features [1] [2], but also in the difficulty of modeling the changes of those features in time [3] [4]. The dynamics of emotion have been explored globally (across the entire utterance) [5] [6] and locally (within segments of the utterance) [7] [8]. A computational model that captures these dynamics of emotion expressions will contribute to both the advancement of automatic emotion recognition methods and the understanding of the dynamic patterns of human communication.

Past research efforts have developed approaches to model emotion dynamics both locally and globally in speech. Kim and Mower Provost [9] demonstrated that dynamic time warping (DTW) could be used to capture emotionally-relevant global dynamics when used in conjunction with k-Nearest Neighbors (kNN). Another popular global dynamic model is Hidden Markov Models (HMMs) [10] [11] [12]. Metallinou et al. [13] use Bidirectional Long Short-Term Memory (BLSTM), HMMs and hybrid HMMs/BLSTM classifiers to model the dynamics of emotion both within an utterance, and between utterances of a dialogue. Their work focused on capturing the dynamics of the dimensional properties of emotion (i.e., valence and activation). They showed that the classification of an utterance within a dialog could be improved by modeling the dynamic evolution of emotion over that dialogue. In [14], Mower Provost demonstrated that local dynamics could be used to discriminate between emotional classes. This work constructed time series estimates of emotion, which were then automatically segmented by identifying regions with consistent dynamics, approximated as straight-lines. These approximations were clustered and the emotional salience of each cluster was calculated. Class labels were assigned to test data by identifying salient evidence associated with each of the classes and choosing the class with the largest amount of salient evidence. However, there remain open questions regarding additional methods for capturing local dynamics, methods that do not restrict the dynamics to linear patterns.

In this paper, we develop a novel framework for extracting local emotional speech patterns. These representative local patterns are referred to as EmoShapelets. EmoShapelets not only provide insights into how emotions vary temporally, but also can be used as input to classifiers to achieve results comparable to the state-of-the-art. The EmoShapelets method makes the essential assumptions that: (1) there exist localized, speaker independent, dynamic emotion patterns inside each utterance and (2) these patterns accumulate to affect the perceived emotion of the entire utterance.

Our method builds on three main concepts, the first concept is called Emotograms (EGs) [15]. EGs are time series estimates of basic emotion content within an utterance. They are typically represented by a 4-dimensional time series, where each dimension represents a classifier-derived estimate of the presence or absence of happiness, anger, sadness or neutrality [15]. EGs can represent different shades of emotions inside the utterance, and how their presence changes over time. EmoShapelets are constructed from segments of the EGs in this work. The second concept is Shapelet [16]. A Shapelet is a sub-sequence in a time series data that is discriminative with respect to target labels. Since the introduction of Shapelets [16], researches have shown success in using Shapelets to mine and classify time series data, typically with Nearest-neighbor [17] [18] or decision tree classifiers [19] [20]. We adopt the Shapelets algorithms to

discover representative segments of EGs. The third concept is Dynamic Time Warping Barycenter Averaging (DBA) [21] [22]. DBA is a technique that generates the most representative time series in clusters of time series data. It seeks to create an average time series for neighbors identified by their DTW distances, without losing the general representative shape of the average time series. DBA is used to update the Shapelets in our algorithm to maximize their emotional salience. Each of these components will be further explained in Section IV.

Our proposed method first constructs EGs from audio and facial expression features. We extract a set of local time series segments from each dimension of the training EGs. We iteratively cluster the segments and generate representative segments using DBA. We collect the representative segments (*EmoShapelets*) and check for redundancy. The final set of non-redundant EmoShapelets is then used as a set of templates to estimate the emotion content of the test utterances.

The contribution of our work includes novel adaptation of the Shapelet classification method into the field of emotion recognition. We modify the original Shapelet algorithm to capture the multi-dimensional, multi-class local dynamics in speech, and use DBA to update EmoShapelets. We experiment with the use of EmoShapelets in speech emotion classification. Our EmoShapelet classification method achieves 64.9% accuracy, comparable to the state-of-the-art dynamic models (63.95% [9] and 63.80% [14]) on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [23].

II. Dataset

We use the IEMOCAP database collected at University of Southern California [23]. The database contains audio, video, and motion capture data of 10 speakers (5 males and 5 females). This work uses the audio and facial motion capture data. The emotion content of each utterance is evaluated by at least three evaluators. We use categorical emotion labels that include happiness, anger, sadness, and neutrality. The emotion label of excitement is merged with happiness, as in previous work [9] [15] [24]. We exclude utterances with only breathing or silence. The shortest duration of utterances used is 0.5 seconds [14]. This work uses a total of 3158 utterances.

III. Feature Extraction and Selection

We use both audio and motion capture ("visual") features. We extract the audio features using openSMILE [25], an open-source feature extractor, including: intensity, loudness, 12 Mel-frequency cepstral coefficients (MFCCs), pitch, Line Spectral Frequencies (LSFs), zero crossing rate, voice activity and energy, with the delta, delta-delta, and regression coefficients. We extract visual features by defining distances between the facial motion capture points, consistent with [9], [14], and [15]. We use a sliding window of length 0.25 seconds with half a window overlap in order to compare to other methods on the same dataset [9] [14]. The final feature set is composed of statistics of the audio and video features, over each window. These statistics include the mean, 25th and 75th percentile, maximum, standard deviation, skewness and kurtosis of the frame-based data in the window. There are 1703 audio features and 620 facial expression features.

We perform feature selection on the training data set using minimal-redundancy-maximal-relevance (mRMR) supervised feature selection method [26]. mRMR selects features that have the maximum mutual information with the classification label (maximal relevance) and are mutually distinct from each other (minimal redundancy). We sweep through input feature sizes from 100 to 500 with a step size of 50. We find that 200 features gives the highest average validation accuracy over all speakers.

IV. Methods

A. Emotograms

Emotograms (EGs) were introduced in [14] as a framework to analyze speech emotion dynamics. EGs are a high-level representation of emotion salient features. They capture the estimated presence or absence of basic emotion classes (e.g., anger, happiness, neutrality, and sadness) over time. EGs are constructed by applying a suite of binary classifiers (one for each basic emotion) to windowed utterances. Emotion dynamics can then be studied by analyzing how the estimates of emotion change over time. For example, local dynamics of EGs are demonstrated to be effective in speech emotion classification task [14]. In this work, EmoShapelets primitives are automatically extracted from segments of the EGs.

In this paper, we define EGs of each utterance as a 4-dimensional time series of emotion estimates (Equation 1).

$$EG_{i,t} = [c_{t,happy}; c_{t,angry}; c_{t,sad}; c_{t,neutral}] \quad (1)$$

Each c_{t,e_k} indicates the classifier-derived confidence in the presence or absence of emotion e_k at window t in utterance i [14].

We extract EGs by training binary emotion classifiers for each basic emotion class. The input to the classifier is the set of audio-visual features (Section III). We use binary SVM models (e.g., angry vs. not angry) to estimate c_{t,e_k}. The SVM binary models are implemented with LIBSVM [27] using a radial basis function (RBF) kernel with gamma coefficient ranging from 2^{-10} to 2^{-6}. Each c_{t,e_k} is measured by the signed distance to the decision hyperplane [14]. See Figure 1 for an example of an Emotogram time series in a happy utterance.

B. EmoShapelet

We formally define an EmoShapelet as a time series vector that captures the sub-utterance dynamics in one of the four dimensions of the EG. The EmoShapelet algorithm involves five steps: generate candidate EmoShapelets from segments of EGs, identify representative EmoShapelets, assess EmoShapelet quality, identify a threshold to note EmoShapelet presence, and finally, remove redundant EmoShapelets.

1) Generate Candidate EmoShapelet: We generate initial EmoShapelet candidates by segmenting EGs into segments of length l, with $l-1$ overlap (Algorithm 1). We sweep through a range of lengths for l from 5 windows to 40 windows in EGs (see Equation 1). This corresponds to segment lengths of 0.625 seconds to 5 seconds.

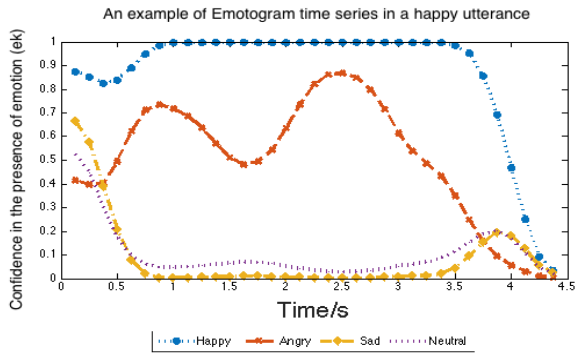

Fig. 1. An example of an EGs in a happy utterance. Four dimensions of the Emotogram are blue:happy, red:angry, yellow:sad, purple:neutral.

Algorithm 1 Shapelet Y, D, l

Require: Y (target labels), D (time series Emotograms), l (intended EmoShapelet length)
Ensure: Set of EmoShapelets LL();
  for each segment T in data D **do**
    for $i = 1$ to $|T| - l + 1$ **do**
      Use segment $T_{i:i+l}$ as initial EmoShapelet S;
      $quality \leftarrow assessCandidate(Y, D, S)$;
      $newS \leftarrow EmoShapeletUpdate(Y, D, S)$;
      Remove $newS$ if it is redundant to set LL;
      Otherwise append $newS$ and its $quality$ to LL;
    end for
  end for
  sort LL by the quality of the EmoShapelets;
  return LL;

2) Identify Representative EmoShapelets: We propose a method to identify representative EmoShapelets using DBA. This approach is in contrast to collecting *all* candidate EmoShapelets, allowing us to reduce the number of EmoShapelets and to control for the redundancy between extracted EmoShapelets.

Given an EmoShapelet candidate, S, extracted from the training data, we calculate the distance to training segments using DTW. We rank the segments from each utterance by distance to S and collect the closest segment, given that the segment is within a threshold distance d (see Section IV-B4 for a description of d). The collection of segments form a cluster.

We apply DBA to the cluster and S to find a representative EmoShapelet, $newS$. DBA first uses DTW to align S with all N segments in the cluster, $\{D_j\}_{j=1}^{N}$ using $DTWmatchpoints$ [16] [28]. At each point in S, denoted S_t, it replaces the value at S_t with an average of the magnitudes of the point in each segment aligned with S_t (Equation 2). The detailed algorithm for DBA can be found in [21] [22].

$$S_i = mean\left(\sum_{j=1}^{|D|} DTWmatchpoints(S_i, D_j)\right) \quad (2)$$

3) Assess EmoShapelet Quality: It is important to retain candidates that are emotionally-relevant. In this section, we establish how emotional relevance is calculated.

Algorithm 2 assessCandidate Y, D, S

Require: D (time series data), Y (target emotion labels), S (EmoShapelet candidate);
Ensure: EmoShapelet S with $S.IG$ (maximum information gain), $S.L$(distance to each utterance), $S.d$ (threshold of distance);
  for each time seriers D_k in data D **do**
    $(L_k, N_k) \leftarrow findDistandNeighborsUCRSuite(S_{i,l}, D_k)$;
  end for
  Sort L in ascending order, let $maxIG = 0$;
  for each distance value L_m in L **do**
    Assume threshold distance $d = L_m$;
    Calculate IG $currIG$ with $maxIGS$ or $sumIGS$ scheme;
    if $currIG > maxIG$ **then**
      $maxIG = currIG$,$S.d = L_m$;
    end if
  end for
  $S.IG = maxIG$, $S.L = L$;
  return S;

We develop two schemes to evaluate the quality of EmoShapelets based on multi-class information gain (IG). IG describes the randomness in the emotion-class label given that a specific EmoShapelet has been observed and has been shown useful in Shapelet-based algorithms [17], [19], [20].

The IG for EmoShapelet $newS$ in emotion e_k is:

$$IG(e_k|newS) = \Big(-\sum_{e\in[e_k,\neg e_k]} P(e)log\big(P(e)\big)\Big) - \Big(-\sum_{e\in[e_k,\neg e_k]}\sum_{s\in[newS,\neg newS]} P(e|s)log\big(P(e|s)\big)\Big) \quad (3)$$

$P(e_k)$ is the prior probability of emotion e_k. $P(\neg e_k)$ is the prior probability of other emotions except e_k. $P(e_k|newS)$ is the conditional probability of an utterance having emotion label e_k given it contains EmoShapelet $newS$. $P(e|\neg newS)$ refers to the given condition that the utterance does not contain $newS$. The determination of whether an EmoShapelet is in a given utterance is made based on a threshold, described in the next section.

The two EmoShapelet quality schemes are:

- The largest IG of the four emotion classes: $maxIGS = max_{e_k}(IG(e_k|newS))$.
- The sum of the IGs over the four emotion classes: $sumIGS = \sum_{e_k} IG(e_k|newS)$.

4) Identify Threshold: An EmoShapelet is present in an utterance if the DTW distance between the EmoShapelet and that utterance is less than a threshold distance, d. We choose d as the distance that maximizes the multi-class IG of an EmoShapelet as defined by either $maxIGS$ or $sumIGS$ (Algorithm 2, based on [19]).

First, we segment the utterance using sliding windows, ranging in length from $l - 0.5l$ to $l + 0.5l$ ($newS$ is of length l). We calculate the DTW distance between $newS$ and every segment in the set. We define the distance between EmoShapelet $newS$ to an utterance as the shortest DTW distance between a segment of the utterance and $newS$. We calculate the distance between $newS$ and every utterance in the training set using $findDistandNeighborsUCRSuite$ (Algorithm 2). We store these values in a distance vector, D.

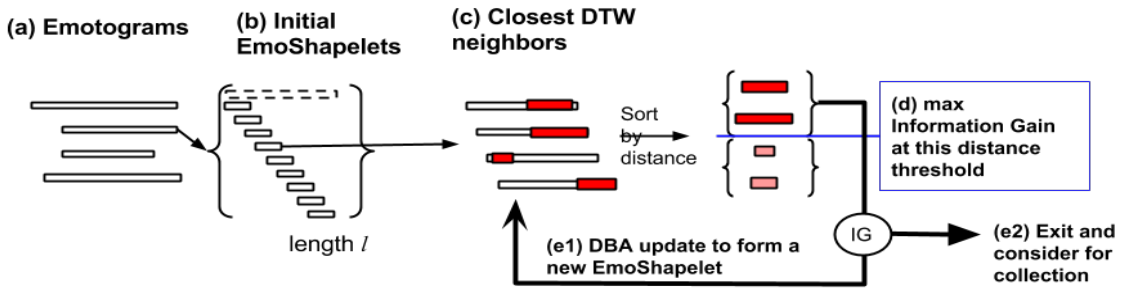

Fig. 2. Proposed EmoShapelet update/generation process. (a) The EGs time series are broken into segments by a sliding window of length l. (b) These segments are the candidate EmoShapelets. (c) The neighbors (shown in red) are identified by finding the closest training segments using DTW and sorting these segments by distance. (d) All neighbors within a threshold distance of the candidate EmoShapelet are clustered (the threshold is chosen to maximize information gain). (e1) A representative EmoShapelet is identified for that cluster and is the new candidate. (e2) This process is iterated until convergence.

Next, we identify a threshold that will be used to decide whether $newS$ is present in an utterance. The threshold is determined using the distances between $newS$ and every training utterance. Intuitively, we would like utterances with distances below the threshold to be from the same emotion class as $newS$ and utterances above the threshold to be from a different emotion class. We identify a candidate threshold, d and calculate the corresponding multi-class IG (either maxIGS or sumIGS) of the EmoShapelet. The final threshold is the value that maximizes the multi-class IG.

5) Remove Redundant EmoShapelets: It is undesirable to collect a large number of redundant EmoShapelets. We avoid redundancy by comparing new EmoShapelets to EmoShapelets that have already been identified. We discard a candidate EmoShapelet if it is similar to an existing EmoShapelet.

We define dissimilarity using the DTW distance between a candidate EmoShapelet and existing EmoShapelets. We discard newly generated EmoShapelets that fall within a fraction α of the distance threshold of any EmoShapelet collected, i.e. discard $newS$ if $dist(newS, emoS_i) \leq \alpha \times emoS_i.t$ for all i. α is determined empirically to be 0.1 in our experiment. A bigger α causes a portion of emotion salient EmoShapelets to be rejected while a smaller α increases the likelihood of collecting redundant EmoShapelets.

C. EmoShapelet Classification Features

Decision Tree classification has been proposed to classify time series data using Shapelets [19]. Distances between Shapelets and time series data have also been used as input features for classifiers such as k-NN or SVMs [20] [29]. In this paper, we implement a variety of classifiers to examine the utility of the EmoShapelets. We train each classifier using EmoShapelets generated only from the training data. We experiment with three types of EmoShapelet properties as input features to these classifiers:

1) d_D: natural exponential of the minimum DTW distances between the EmoShapelets and each utterance $exp(-dist(emoS, utt))$. See section IV.B.4 for the definition of $dist(emoS, utt)$.
2) $d_{D,IG}$: natural exponential of both distances and the IGs of the EmoShapelet $exp(emoS.IG - dist(emoS, utt))$
3) $d_{D,IG,M}$: natural exponential of distances, IGs, and the maximum magnitudes of the EmoShapelets $exp(max(emoS.magnitude) + emoS.IG - dist(emoS, utt))$

We use natural exponential to avoid invalid numbers during the course of calculations. For example, the distance between an EmoShapelet and an utterance can be zero. This causes the function of $emoS.IG/dist(emoS, utt)$ to be undefined. However, the exponential function of $exp(emoS.IG)/exp(dist(emoS, utt)) = exp\big(emoS.IG - dist(emoS, utt)\big)$ can avoid this undefined value. The inclusion of information gains d_{IG} and/or magnitudes d_M of EmoShapelets is designed to explore the impact of these properties of EmoShapelets on classification accuracies.

We perform feature selection, using multi-class IG scheme value to select the top n EmoShapelets. We sweep through the size of n from $[50, 100, 200, 1000, 2000, ALL]$.

D. EmoShapelet Classification Methods

We classify the test utterances by summing the IGs of EmoShapelets observed in each test utterance. IGs associated with each emotion label are summed separately, and the emotion label associated with the maximum sum of the IGs across the 4 emotion classes is assigned as the label for this test utterance.

We also apply k-Nearest Neighbor (k-NN) classifier to the EmoShapelets as in [30] [31]. We use both cosine similarity and Euclidean distance as means to measure the similarity of the feature vectors. The cosine similarity measure outperforms the Euclidean distance measure significantly ($p < 0.01$, paired t-test).

All models are trained with leave-one-speaker-out cross-validation within the training sets. All results are reported as the unweighted accuracy, the average of the recall associated with each emotion class.

V. Results and Discussion

A. Results on EmoShapelets

Table I shows the classification results by the IG summation method (Section IV.D). We make several observations. Firstly,

TABLE I

UNWEIGHTED ACCURACIES (%) OF SUMMING EMOSHAPELETS SCORES WITHIN EACH UTTERANCE WITH AND WITHOUT DBA. THE THREE TYPES OF SCORES d_{IG},$d_{IG,D}$, AND $d_{IG,D,M}$ ARE DEFINED IN THE SECTION IV.C. ‡ AND ‡‡ MEANS THAT NO DBA RESULTS FROM ALL LENGTHS OF EMOSHAPELETS ARE STATISTICALLY HIGHER THAN DBA RESULTS BY $p = 0.05$ AND $p = 0.01$ IN PAIRED T-TEST. ** MEANS THAT RESULTS FROM ALL EMOSHAPELETS LENGTHS USING MAXIGS VS SUMIGS ARE STATISTICALLY SIGNIFICANTLY DIFFERENT BY $p = 0.01$ IN PARIED T-TEST. BOLDED NUMBERS ARE THE BIGGEST CLASSIFICATION ACCURACY GIVEN DBA OR NO DBA.

scheme	DBA present	features	1.25 seconds	2.50 seconds	3.75 seconds	5.00 seconds	0.625-5 seconds	1.25-3.75 seconds
maxIGS	yes	d_{IG}**	51.4	55.1	52.6	52.4	59.3	58.8
		$d_{IG,D}$**	51.4	**56.6**	53.1	54.2	**60.1**	**58.9**
		$d_{IG,D,M}$**	**51.9**	56.1	53.6	**54.8**	59.7	58.7
	no	d_{IG}	46.8	56.2	51.0	51.9	60.1	59.8
		$d_{IG,D}$	48.9	56.6	54.2	53.4	60.7	58.9
		$d_{IG,D,M}$	49.2	**57.2**	53.9	**54.6**	**61.5**	**59.6**
sumIGS	yes	d_{IG}	47.2	54.0	53.1	51.0	57.8	56.4
		$d_{IG,D}$	48.4	55.1	54.1	51.8	58.4	57.4
		$d_{IG,D,M}$	48.6	55.4	**54.4**	52.6	57.25	57.6
	no	d_{IG}	48.8	55.1	51.2	51.9	58.4	57.6
		$d_{IG,D}$‡‡	50.1	55.4	54.5	53.3	60.1	58.5
		$d_{IG,D,M}$‡‡	**50.9**	54.7	**54.8**	52.6	60.1	58.3

EmoShapelets of different lengths have discriminative power in affective speech. Secondly, the classification results improve when EmoShapelets of different lengths are combined (seen in last two columns of Table I). This demonstrates that EmoShapelets of these lengths are complementary. Thirdly, the most discriminative length of the EmoShapelets is 2.5s.

Two experiments, one with DBA updating and one without DBA show different results in classification rates. In Table I, when the IGs score summing is used to classify emotions, classification accuracies when using no DBA are comparable or slightly higher than accuracies derived with DBA. However, the algorithm with no DBA collects double the number of EmoShapelets (Figure 3). The benefit to this is that a larger number of EmoShapelets is collected, resulting in a slightly wider coverage of EmoShapelets over an utterance. This results in a redundancy in the EmoShapelets' IGs scores and a slightly higher classification rate.

When other classification methods are used, the classification accuracies of systems with DBA are generally higher than those without DBA. This is shown by the † in Table III. DBA not only reduces the number of EmoShapelets collected, it also improves inter-class balance between the number of EmoShapelets found in different EGs dimensions. Class imbalance and an increase in the number of EmoShapelets adversely affect classification results, indicating that DBA update is an important step for EmoShapelet collection.

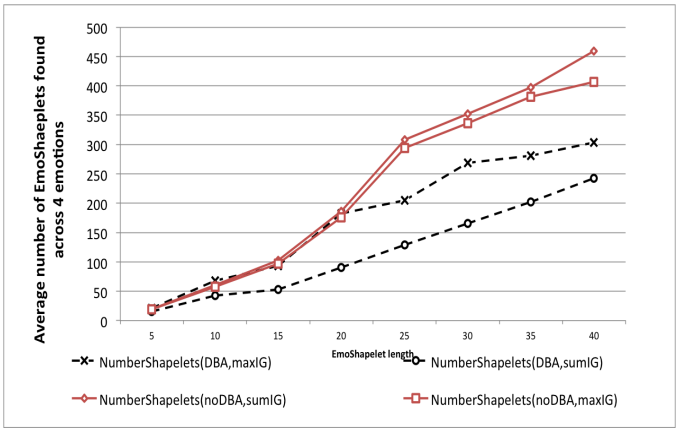

Fig. 3. Average number of EmoShapelets collected. x-axis: lengths of EmoShapelets from 0.625s to 5.0s. y-axis: number of EmoShapelets. Black: with recursive DBA EmoShapelet Update. Red: without DBA update.

In Figure 4, we show examples of EmoShapelets selected using the maxIGS scheme. We group EmoShapelets generated from the same dimension of the EGs. For example, the top-left figure groups EmoShapelets extracted from the happy EG dimension. These four EmoShapelets are the most relevant (have the highest IGs) to the classes of happiness (blue), anger (red), sadness (yellow), and neutrality (purple). The flat blue EmoShapelet suggests that the consistent presence of happiness over all frames best distinguishes happiness. The yellow EmoShapelet suggests that lack of happiness, followed by an increase at the end of the segment is indicative of sadness.

The bulk of the system's computational runtime lies in the iterations of DTW algorithm used in DBA update and EmoShapelets selection. Exact brute force DTW algorithms take $O(n^2)$ time and are slow in large scale applciations [32]. To reduce computational complexity, we used the DTW measure with UCR suite heuristics [32]. These heuristics speed up DTW distance calculations and are able to search trillions of data points within reasonable amount of time [32]. In addition, EmoShapelets algorithm in the four EGs dimensions were implemented in parallel. The DBA update and the EmoShapelet comparisons for selection methods are both embarrassingly parallel problems – both processes can be done by distributing chunks of workload to many parallel processors without incurring overhead costs in run time. With 8 processors on Intel(R) Xeon(R) Quad-Core E5620 2.40GHz with 48GB RAM machine, our system processes 120,000 segments of EGs (30,000 in each emotion category), and takes

TABLE II

UNWEIGHTED CLASSIFICATION ACCURACIES (%) FROM APPLYING FOUR TYPES OF CLASSIFIERS TO THE EMOSHAPELETS GENERATED WITH DBA. TWO TYPES OF SCHEMES FOR EMOSHAPELET SELECTIONS ARE COMPARED: MAXIGS AND SUMIGS. † AND †† MEANS THESE RESULTS ARE STATISTICALLY SIGNIFICANTLY HIGHER FROM RESULTS GENERATED WITHOUT DBA AT $p = 0.05$ AND $p = 0.01$. * MEANS THE RESULTS FROM MAXIGS AND SUMIGS ARE STATISTICALLY DIFFERENT AT $p = 0.05$.

Classifier	EmoShapelet propoerty	maxIGS	sumIGS
SVM linear kernel	d_D	61.1 ± 5.2	63.1 ± 5.7
	$d_{IG,D}$†	63.1 ± 3.8	63.4 ± 5.8
	$d_{IG,D,M}$††	64.2 ± 4.4	64.6 ± 5.1
k-NN cosine distance	d_D††	62.3 ± 5.1	62.4 ± 5.7
	$d_{IG,D}$††	63.2 ± 5.2	62.9 ± 5.4
	$d_{IG,D,M}$††	**64.9 ± 4.3**	63.3 ± 5.7
Decision Tree	d_D	23.8 ± 6.0	29.1 ± 6.1
	$d_{IG,D}$	25.4 ± 5.0	25.6 ± 4.8
	$d_{IG,D,M}$††	41.2 ± 9.2	37.8 ± 8.1
Naive Bayes	d_D††	62.3 ± 5.2	62.4 ± 5.7
	$d_{IG,D}$††	63.3 ± 5.3	62.9 ± 5.4
	$d_{IG,D,M}$††	64.4 ± 4.9*	63.3 ± 5.7*

4 hours to produce a total of 2000 EmoShapelets[1].

B. Results on EmoShapelets Classification

Table II shows the classification results with three types of EmoShapelet properties as input features to the classifiers. There is no statistically significant difference between maxIGS and sumIGS, indicating that the two IG selection schemes are comparable. Decision tree classifiers perform significantly poorer than other three classifiers listed ($p < 0.01$, paired t test between the accuracies of Decision Tree with respect to each of the other classifier in Table III)

We find that the highest unweighted per-class accuracy is produced by the k-Nearest Neighbor classifier. We achieve a result of unweighted per-class accuracy of $64.91 \pm 4.32\%$ of classification rate, comparable to state-of-the-art results with the same sliding window size (0.25s) on IEMOCAP data: 63.80% is achieved by approximating local dynamics in EGs as straight lines in [14]; 63.95% is achieved using a DTW-k-NN classifier to capture global dynamics in EGs [9], and 64.20%, which is achieved through HMM models [33].

[1]This code is available at http://umich.edu/ juneysg/EmoShapelets.html. We implement EmoShapelet algorithms with C++ on OpenMP parallel programming platform.

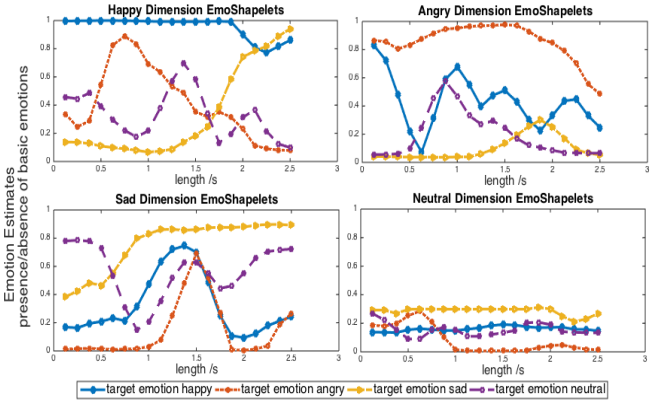

Fig. 4. Examples of EmoShapelets of length 2.5s gathered using $maxIGS$ scheme. Each quadrant corresponds EmoShapelets generated from one of the four emotion dimensions. Top-left: happy, top-right: angry, bottom-left: sad, bottom-right: neutral. Different colors indicate the emotion that these EmoShapelets best distinguish: happiness (blue), anger (red), sadness (yellow) and neutrality (purple).

TABLE III

AVERAGE CONFUSION MATRIX (%) OVER 10 TEST SPEAKERS GENERATED FROM K-NN CLASSIFIER WITH COSINE DISTANCES. ROWS: ACTUAL EMOTION LABELS OF TEST UTTERANCES, COLS: CLASSIFIER PREDICTED EMOTION LABELS. H,A,S,N STAND FOR HAPPY, ANGRY, SAD AND NEUTRAL EMOTIONS.

	predict H	predict A	predict S	predict N
actual H	**78.34**	5.01	10.97	5.68
actual A	17.48	**63.25**	6.05	13.21
actual S	8.34	1.82	**81.00**	8.83
actual N	25.88	7.01	36.84	**30.27**

In Table III, we show the mean confusion matrix, derived from k-NN classifier with cosine distance, averaged over 10 test speakers. Neutral emotion is the most difficult emotion to classify with EmoShapelets, while happiness and sadness are relatively more accurately classified. This observation can partly be attributed to the observation that almost three times as many EmoShapelets were discovered from each of the happiness and sadness dimensions of EGs than from the neutral dimension.

VI. CONCLUSION

In this paper, we propose the novel concept of EmoShapelets which describe the sub-utterance dynamics in emotional speech. We propose methods to 1) generate EmoShapelet candidates, 2) update candidates into representative EmoShapelets, 3) evaluate the relevance of EmoShapelets with respect to the emotion labels of training utterances, 4) find the threshold distances of EmoShapelets, and 5) select a final collection of non-redundant EmoShapelets. We explore the use of EmoShapelets with a variety of classifiers and attain classification accuracies comparable to the state-of-the-art results with the IEMOCAP dataset. EmoShapelets provide a interpretable method for classifying emotion content using local Emotogram dynamics in speech.

Our future work will focus on simplifying the EmoShapelet classification pipeline. We will also explore the cross-dataset transferability of the EmoShapelets. Experiments can be conducted to discover whether EmoShapelets found in one dataset can be used to classify emotion in a similar dataset. In addition, we will explore the integration of EmoShapelets with global dynamic patterns in affective speech classification.

ACKNOWLEDGMENT

We thank Professor Donald C. Winsor for his support. We also thank members of the University of Michigan Computational Human-Centered Analysis and Integration lab for their advice on this project.

REFERENCES

[1] D. Hutchison and A. Esposito, *Cross-modal analysis of speech, gestures, gaze and facial expressions*. Springer Science & Business Media, 2009, vol. 5641.

[2] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," *Multimedia, IEEE Transactions on*, vol. 2, no. 3, pp. 141–151, 2000.

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

[4] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern Recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[5] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th International Conference on Multimodal Interfaces*. ACM, 2004, pp. 205–211.

[6] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realisation," in *Affect and Emotion in Human-Computer Interaction*. Springer, 2008, pp. 75–91.

[7] L. D. Sanders and D. Poeppel, "Local and global auditory processing: Behavioral and erp evidence," *Neuropsychologia*, vol. 45, no. 6, pp. 1172–1186, 2007.

[8] M. Pantic and I. Patras, "Detecting facial actions and their temporal segments in nearly frontal-view face image sequences," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, vol. 4. IEEE, 2005, pp. 3358–3363.

[9] Y. Kim and E. M. Provost, "Emotion classification via utterance-level dynamics: A pattern-based approach to characterizing affective expressions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3677–3681.

[10] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1057–1070, 2011.

[11] D. Ozkan, S. Scherer, and L.-P. Morency, "Step-wise emotion recognition using concatenated-hmm," in *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. ACM, 2012, pp. 477–484.

[12] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden markov models," in *Affective Computing and Intelligent interaction*. Springer, 2011, pp. 378–387.

[13] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.

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

[15] E. M. Provost and S. Narayanan, "Simplifying emotion classification through emotion distillation," in *Proceedings of APSIPA Annual Summit and Conference*, 2012.

[16] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2009, pp. 947–956.

[17] L. Ye and E. Keogh, "Time series shapelets: a novel technique that allows accurate, interpretable and fast classification," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 149–182, 2011.

[18] J. Lines, L. M. Davis, J. Hills, and A. Bagnall, "A shapelet transform for time series classification," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 289–297.

[19] A. Mueen, E. Keogh, and N. Young, "Logical-shapelets: an expressive primitive for time series classification," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2011, pp. 1154–1162.

[20] J. Hills, J. Lines, E. Baranauskas, J. Mapp, and A. Bagnall, "Classification of time series by shapelet transformation," *Data Mining and Knowledge Discovery*, vol. 28, no. 4, pp. 851–881, 2014.

[21] F. Petitjean, G. Forestier, G. Webb, A. Nicholson, Y. Chen, and E. Keogh, "Dynamic time warping averaging of time series allows faster and more accurate classification," in *IEEE International Conference on*, 2014.

[22] F. Petitjean, A. Ketterlin, and P. Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, no. 3, pp. 678–693, 2011.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[24] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–8.

[25] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[26] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1226–1238, 2005.

[27] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[28] M. Cuturi, "Fast global alignment kernels," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 929–936.

[29] P. Senin and S. Malinchik, "Sax-vsm: Interpretable time series classification using sax and vector space model," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 1175–1180.

[30] S. L. Salzberg, "On comparing classifiers: Pitfalls to avoid and a recommended approach," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 317–328, 1997.

[31] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, "Fast time series classification using numerosity reduction," in *Proceedings of the 23rd International Conference on Machine Learning*. ACM, 2006, pp. 1033–1040.

[32] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2012, pp. 262–270.

[33] E. Mower and S. Narayanan, "A hierarchical static-dynamic framework for emotion classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2372–2375.