Emotion Analysis Using Audio/Video, EMG and EEG: A Dataset and Comparison Study

Farnaz Abtahi
CUNY Graduate Center
fabtahi@gradcenter.cuny.edu

Tony Ro
CUNY Graduate Center
tro@gc.cuny.edu

Wei Li
City College of New York
wli3@ccny.cuny.edu

Zhigang Zhu
CUNY Graduate Center & City College of New York
zhu@cs.ccny.cuny.edu

Abstract

This paper describes a study on automated emotion recognition using four different modalities – audio, video, electromyography (EMG), and electroencephalography (EEG). We collected a dataset using the 4 modalities as 12 human subjects expressed six different emotions or maintained a neutral expression. Three different aspects of emotion recognition were investigated: model selection, feature selection, and data selection. Both generative models (DBNs) and discriminative models (LSTMs) were applied to the four modalities, and from these analyses we conclude that LSTM is better for audio and video together with their corresponding sophisticated feature extractors (MFCC and CNN), whereas DBN is better for both EMG and EEG. By examining these signals at different stages (pre-speech, during-speech, and post-speech) of the current and following trials, we found that the most effective stages for emotion recognition from EEG occur after the emotion has been expressed, suggesting that the neural signals conveying an emotion are long-lasting.

1. Introduction

Emotional state can be observed or measured in many different ways, including through facial expressions, speech, and physiological signals. The idea of emotion recognition while speaking has been investigated by several researchers in applications such as human-computer interaction (HCI) and call center monitoring. These applications have also produced multiple datasets that are being used by researchers. The goal of the majority of emotion detection work has been to optimize the accuracy of emotion recognition, more recently by utilizing the state-of-the-art statistical or machine learning models and the most relevant modalities such as visual information, vocal features, body movements and posture, or physiological signals. Several attempts have been made to combine multiple modalities to further improve the accuracy of the emotion recognition models.

The goal of the current research was twofold. First, we examined the efficacy of using different modalities and machine learning models for emotion recognition. We collected a rich new dataset by recording video, audio, facial muscle movements (with EMG signals), and brain activity (with EEG signals) while subjects (actors) spoke a generic sentence expressing one of the seven different emotions. We then applied a set of state-of-the-art feature extractors, each suitable for a specific modality, before applying the most effective deep learning and statistical models. Various different machine learning models were compared to determine the best models for the different modalities.

The second goal, which is the main focus of this work, was to compare the characteristics that are specific to each modality and the conditions in which each modality performs optimally. We analytically show that even though each modality might seem ineffective in some settings, if used correctly, their unique contributions to emotion recognition can be effectively applied to increase classification. Scientifically, we investigated, using the same dataset, how much spatial-temporal visual facial expression, auditory speech information, facial muscle movements, and neural activity can classify a person's emotion and in what stages of speech and expression that each modality best captures the emotion. To assess how neural activity may be used to detect emotions, we used EEG to record brain activity while subjects expressed different emotions. Thus, the overall goal of this work was not to improve existing emotion recognition methods, but to thoroughly study different emotional state detecting modalities and to determine the optimal stages of information in the signals, the best categories of feature extractors, and the most appropriate machine learning models that would best fit each modality.

As a summary, the contributions of the work include: (1) A new multimodal dataset was collected with four different modalities: audio, video, EMG, and EEG. (2) A thorough comparison was performed across these four modalities to determine how to optimize the data, which features and models are most informative, and to offer insights into the effectiveness of each modality to classify emotions. (3) Most notably, we study at what stage of each modality, especially for the neural activity measured with EEG signals, emotion information prevails and for how long they remain reliable.

The remainder of the paper is organized as follows.

Section 2 is a review of related literature. We will introduce and explain the models used for emotion recognition in this work in Section 3. The process of data collection and preprocessing is explained in Section 4. Section 5 covers the steps of data preparation and our analyses. We conclude our work and discuss our future direction in Section 6.

2. Related work

There is a large body of work on emotion recognition using different modalities separately and in combination with one another. Three categories of datasets are usually used in analyzing emotions: acted emotions, natural spontaneous emotions, and elicited emotions [1]. Although different actors may understand and interpret instructions differently and may actually experience the emotions to different degrees, data obtained from acted emotions are less ambiguous because actors express the exact emotions they were instructed to act.

In contrast, spontaneous speech and emotions can, for example, be collected from call center data [2] or through human-computer interaction [3]. These emotions are more diversified and are often difficult to classify given that the data must be mapped onto a limited number of classes. Even if it is evident that emotion research should ideally target natural databases, acted databases are more systematically controlled and useful, especially neural activity will be measured. Furthermore, there is a direct correspondence between the collected data and their labels, generally resulting in higher accuracy in emotion recognition [4, 5]. We therefore use acted emotions for data collection in this work.

Generally, facial expressions and speech have been the two most used modalities for emotion recognition, although other modalities have also been investigated. In the area of unimodal emotion recognition, there have been many studies using a variety of different, but single modalities. Facial expressions [6, 7], vocal features [4, 8], body movements and postures [9, 10], physiological signals such as skin temperature, skin conductance, blood volume pulse and heart rate [11, 12], and EMG (facial muscle activity) [13, 14] have been used as inputs during these attempts. Several approaches have also examined the integration of information from facial expressions, speech, and body gestures [1, 15].

Another approach that has more recently been explored for emotion recognition is through EEG, which measures electrical activity in the brain [16, 17, 18]. EEG is especially interesting due to its capability to detect internal emotional states, as opposed to the other modalities mentioned above. Some previous studies ([16, 17, 18]) have incorporated the use of EEG in attempts to determine the inner emotional (affective) state. Here, we recorded EEG signals during different expressed emotional states and compared them with other modalities.

3. Baseline and deep learning approaches

Different machine learning techniques have been used in emotion recognition. One approach, which has been successful, is to use deep learning approaches because they have the ability to learn the most relevant features with respect to the task. Two deep learning models were used in this work to classify the emotions: Long-Short Term Memory (LSTM) [19] and Deep Belief Network (DBN) [20]. Although both models are characterized as being 'deep', as they use layers of latent or hidden variables, they have very different characteristics.

The architecture of deep learning techniques can be categorized into two different categories: generative and discriminative. The deep models that fall into each category often share the properties of the other category, making it difficult to draw a clear boundary between the two groups of models. Generative models are very useful for both classification and regression tasks, especially when data preparation and pre-training of the parameters of the model are necessary. These models have the ability to initialize the search through the parameter space in an area that potentially contains the solution. On the other hand, the architecture of the discriminative models has direct ability to classify the data [21]. In other words, the former models describe the distribution of data, whereas the latter models describe the distribution of targets conditioned on data [22]. In the current work, we investigated how the two different models handle the data from the four different modalities and draw some useful conclusions about the effectiveness of these models for different types of datasets.

Example of discriminative architectures include Convolutional Neural Network (CNN) [23], Recurrent Neural Network (RNN) [24], and LSTM. RNN is an artificial neural network model that has feedback connections in the hidden units, therefore RNN can store historical information like memory and can solve context-dependent tasks with the architecture. However, the vanishing and explosion gradient problem makes learning of RNN difficult. LSTM is a special type of RNN architecture to overcome the vanishing gradient problem of RNN. Thus, we choose LSTM for both its memory and its performance.

The DBN on the other hand, is an example of generative models. DBNs are probabilistic graphical models that are built by stacking up Restricted Boltzmann Machines (RBMs). An RBM is an undirected graphical model that consists of one layer of visible and one layer of hidden Bernoulli units. There are no connections between units of the same layer, but the two layers are fully connected to each other. Connections between layers are bidirectional and symmetric, so the weights are also shared between both layers. The effect of pre-training is studied in detail by Erhan et al. [25]. They explain that the reason why pre-trained DBNs work much better than traditional neural networks is that pre-training initializes the parameters of the

DBN in a more desirable area of parameter space where a better local optimum can be found. Therefore, pre-training introduces a bias towards configurations of the parameters that the supervised learning phase can explore, which is adopted in this paper.

For applications that involve speech, i-vector features [26] have been state of the art. For the sake of completeness of comparisons, in addition to DBN and LSTM, we will use this method for classification of voice signals. i-vectors convey the speaker characteristic among other information such as transmission channel, acoustic environment or phonetic content of the speech segment. The i-vector extraction could be seen as a probabilistic compression process that reduces the dimensionality of speech-session super-vectors according to a linear-Gaussian model.

4. Data collection and feature extraction

The data we used to train and test the models was gathered from 12 human subjects (5 female and 7 male individuals), who participated after informed consent. The study was approved by the Institutional Review Board of the City University of New York. In general, having a relatively small number of subjects is typical in neuroscience studies due to the difficulty in data collection. For this study, each testing session lasted approximately two hours and it took approximately four months to recruit the actor subjects and collect all of the data. We included a large number of repetitions/instances in the dataset to minimize variability. For each of the 7 emotions, 50 trials were expressed for a total of 350 emotion instances per subject. *We will release the dataset following publication of the paper for research purposes*.

The subjects either had acting experience or were acting students because the emotions needed to be expressed as naturally and believably as possible. Every five seconds, one of the seven standard emotion labels were presented on a monitor placed 57 cm in front of the subject. The emotions were *happiness*, *sadness*, *anger*, *surprise*, *fear*, *disgust*, and *neutral*. Each time an emotion label appeared on the screen, the subject uttered the sentence "*The sky is green*" while trying to mimic the facial expression and experience the emotion associated with that label. This sentence was chosen because of its neutral content, thereby minimizing interference with any emotion that the subject was trying to experience and express. During the utterance, the subject's face and voice were video recorded and EMG and EEG signals were acquired from their facial muscles and scalp using gold plated surface electrodes that were connected to Grass amplifiers. The camera and microphone were placed in front of the subject to ensure an adequate quality of the acquired video and voice.

Each emotion label was displayed for 4 seconds, and a one-second break was given between every emotion. Overall, the longest it took the subjects to speak the sentence was approximately 2.5 seconds. The entire interval, therefore, was not completely filled with the utterance of the sentence and started and/or ended with periods of silence.

The entire session was divided into five sub-sessions. Each sub-session contained 10 repetitions of each emotion in random order. That is, we used a 7x10 total number of emotions per sub-session, or 5x7x10=350 emotions overall, for each of the 12 subjects. Between every two consecutive sub-sessions, the subject took an optional break that was arbitrarily long.

The video was used for extracting two types of information: 1) a clip of audio signals with a 44.1kHz sampling rate, and 2) an image sequence of 24 screenshots per sentence. The 24 images were evenly sampled from the 2.5 seconds (on average) during speech, such that this window included most of the emotional expression. Only 24 frames were used for computational efficiency, following the work presented in [6]. A few samples of the screenshots are shown in Figure 1.

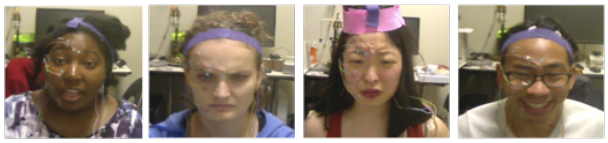

Figure 1: Screenshots from videos of four subjects

The audio was recorded using a laptop's microphone with the 44.1kHz sampling rate. We divided the audio into 20ms intervals with 10ms offsets and then extracted MFCC features from each interval separately. The features extracted from the intervals formed a sequence that embed both frequency and time information.

The EMG data consisted of six channels captured through six surface electrodes with a 5kHz sampling rate that was then downsampled to 1 kHz. Six muscles were chosen: the depressor anguli oris, zygomaticus major, levator labii superior alaeque nasi, levator labii superioris, procerus, and occipitofrontalis (Figure 2, top). These are the major muscles that are involved during facial expressions and their equivalent facial Action Units (AUs) have often been used in the literature for facial expression recognition [6, 27]. A band-pass Butterworth filter (20 to 450 Hz) was applied to the EMG data to eliminate noise and meaningless parts of the signals.

EEG data were acquired using the same sampling frequency as the EMG data but through 8 surface electrodes placed onto the scalp: F3, Fz, F4, Cz, P3, Pz, P4, and O2 (Figure 2, bottom). The preprocessing steps applied to the EEG data were similar to the EMG data but with different bandpass filter settings (0.1 to 30 Hz). Figure 3 shows samples of EMG and EEG signals collected from one of the subjects while acting "fear".

All electrode impedances were below 10 kΩ at the start of the experiment. After filtering the EEG and EMG

signals, wavelet transforms (WT) [28] were applied for feature extraction.

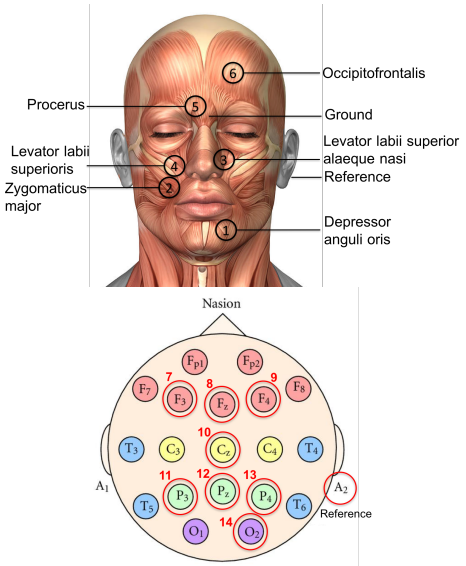

Figure 2: the position of the sensors on the face (top) and scalp (bottom).

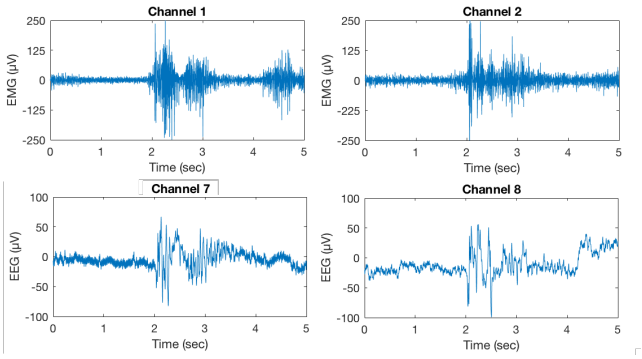

Figure 3: The EMG/EEG Data (Subject: 03, Emotion: Fear, EMG channels 1 and 2, and EEG channels 7 and 8)

To extract emotions from the images, we first cropped the face area on each frame using the open source DLib C++ Library [29]. A few cropped frames from a "happy" sequence are shown in Figure 4. Each cropped image was 186x186 pixels, which was enlarged to 224x224 pixels to fit into the feature extraction method. Note that transparent tape was used on the faces to minimize occlusion of the facial features.

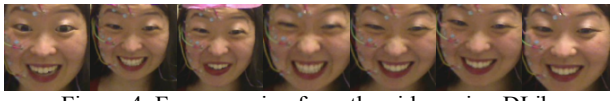

Figure 4: Face cropping from the video using DLib

Using pre-trained models for visual feature extraction has become a very common and promising approach among researchers [6, 30]. As explained earlier, we have a sequence of 24 images per utterance of the sentence. The features were extracted by applying an integrated deep learning model with the pre-trained VGG-16 network [31], followed by the ROI networks, as proposed by Li et al., 2017 [6]. The reason we chose the VGG+ROI model rather than more sophisticated models, such as ResNet [32], is that the VGG model is sophisticated enough for our data and VGG+ROI has been a further trained model using more than 10K facial expressions [6].

More specifically, the ROI nets were designed to ensure that regions of interest on the faces were learned independently; each sub-region (out of 20 in this case) had a local CNN - an ROI net, whose convolutional filters were only trained for the corresponding region for facial expression recognition. The structure of the VGG+ROI model is illustrated in Figure 5. The VGG net's output from fully connected layer 7 (fc7) provided the input to the ROI net. Each feature vector was obtained from the output of the last layer and had 2048 elements.

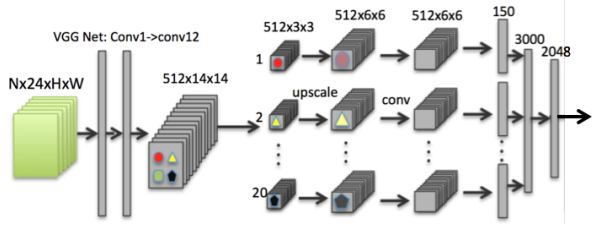

Figure 5: Framework of the CNN model we used: VGG Net followed by the ROI Nets [6]

5. Hypotheses and analytical experiments

The goal was both to classify the emotion using each of the four modalities and to explore the unique characteristics of the EEG modality. Although the former is a very important topic, it has been explored by many researchers, therefore the latter will be the focus of most of the analyses in this section.

Before describing the details of each analysis, we first describe the experimental setup for the models. The following configurations were shared among all the analyses unless otherwise is explicitly stated.

Even though LSTM and DBN are both deep models and are able to extract hidden information from the data and represent it as learned features, each has different strengths. Since each modality captured in our data is a *sequence* of information over time, we used the LSTM, which is very powerful in dealing with temporal information. On the other hand, DBN is a very powerful feature extractor when pre-trained properly. For this reason, either LSTM or DBN is more suitable depending on the task. These models have also been combined and used as a hybrid to provide strengths from both models [21]. In this paper, we will focus on the comparison of LSTM and DBN for all modalities instead of integration.

5.1. Data preparation

In this section, we provide details and parameter configurations for the different feature extraction methods.

Images: For each of the 24 images, we extract a feature vector size of 2048. When using LSTM, these feature vectors were provided to the model as a sequence and an output was reported after the entire sequence. Given a sequence of n frames $\mathbf{X}_i = \{X_1, ..., X_n\}$, the target prediction is the class of the last frame X_n (Figure 6). When the length of the sequence of images was shorter than what the model expects, we padded the sequence with a blank black frame. In such cases, the sequence of images was padded at the beginning with black images and the features extracted from those frames were also appended to the rest of the feature vectors. On the other hand, when the data was provided to the DBN, all feature vectors from the sequence were concatenated and used as a single input vector to the model. Note that each image frame has been turned into a feature vector of a length of 2048 using the ROI Net, so the input dimension to DBN from 24 frames is 24*2048.

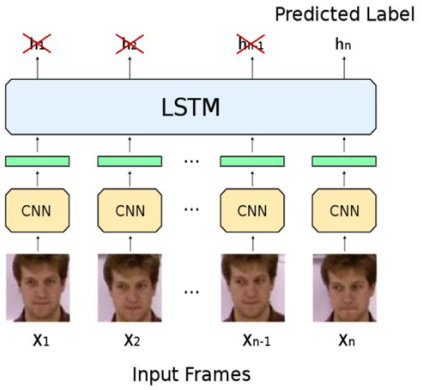

Figure 6: Feature extraction via CNN (in our case the VGG+ROI model) and prediction using LSTM (Bellantonio, 2016)

Voice: For extraction of MFCC features, we used a Hamming window of size 20ms with a 10ms offset, following the work presented by Wang, 2014 [33]. From each 20ms interval, we then extracted MFCC features and chose the 20 most significant coefficients. The input to the LSTM model was a sequence of such feature vectors. For the DBN, we concatenated these feature vectors into a single vector. Again, if the sequence was shorter than expected, we padded it with a frame of silence. The MFCC features extracted from these silent frames were used normally as part of the sequence. We also used the i-vector features along with PLDA, for which we set up the input sequence similar to LSTM. For the extraction of the i-vector features and classification of the feature vectors using PLDA, we utilized the MSR Identity Toolbox [34].

EMG and EEG: Since both EMG and EEG are non-stationary signals that share similar characteristics with the voice signal, we used the same settings for processing them. The same 20ms window with 10ms offset was used to cut the signal of each of the six channels into a sequence of intervals. WT was applied to each interval channel by channel and the 20 most significant coefficients were kept as a feature vector for each channel. To apply LSTM to EMG, the WT coefficients of all 6 channels from a single interval were concatenated and used as the feature vector associated with that time step. Similarly, the WT coefficients of all 8 channels of EEG were concatenated to form the feature vector of each time step. The input feature vector to DBN on the other hand, was formed by concatenating the WT coefficients of the entire sequence. In the case of EEG, shorter sequences were padded with WT coefficients corresponding to Electrocerebral inactivity (ECI) or electrocerebral silence (ECS), which is defined as no EEG activity over 2μV [35]. In this case we padded the sequence with zeros to represent the inactivity of the area, both for EMG and EEG, and extracted WT features from the padded sequence.

5.2. Artifact removal

EEG recordings are usually corrupted by spurious extracerebral artifacts, which should be rejected or corrected. Since manual screening of human EEGs is inherently error prone and might induce experimenter bias, we used an automatic artifact detection method. The impact of muscular activity on the EEG signal can be evaluated using artifact removal approaches [36, 37, 38]. To remove the effect of EMG from EEG signals, we used the AAR plug-in for EEGLAB [38]. We compared the results on unfiltered EEG data with filtered EEG to assess the effectiveness of this removal procedure, but once the artifact removal procedure was established, we preformed the remaining analyses with filtered EEG signals only.

5.3. Initial experiments

We began our analyses by comparing different modalities for emotion recognition using DBN and LSTM as the classification methods. First, we randomly chose 60% of the samples for training and the remaining 40% for test. For each modality, the entire 5-second sequence was provided to the model and the model classified the sequence into one of the seven emotions. For image sequences, a more accurate term to use would be facial expressions instead of emotions, but since our goal was to label the data based on the underlying emotion, we simply refer to the task as emotion recognition/classification.

The process of randomly dividing the data into 60% training and 40% test samples was repeated 100 times and the results averaged over all repetitions. The goal was to classify the emotions, not the subjects. The training and test datasets do not overlap from an emotion classification perspective, but they did contain samples from the same subject expressing the same emotion on different trials. Emotions were quite different across sub-sessions, and the order of the emotions that were being expressed was randomized and different across the sub-sessions to minimize information leak. Table 1 summarizes the means

and standard deviations of the results, showing that LSTM (the discriminative model) classifies better on voice and images, whereas DBN (the generative model) performs better for EMG and EEG data. The results also show that facial images discriminate between emotions the best, followed by voice, EMG, and then EEG activity.

Table 1: Comparison of emotion recognition accuracies on all modalities using DBN and LSTM. Results are reported as "mean (standard deviation)".

	DBN	LSTM
EEG before artifact removal	40.4% (8.5)	35% (10.9)
EEG after artifact removal	**51.7% (7.3)**	37.6% (8.2)
EMG	**62.8% (8.7)**	58.4% (9.1)
Images	71% (7.7)	**81.1% (5)**
Voice	64.5% (7.9)	**70.3% (5.7)**

As also seen in Table 1, the EEG signals result in higher accuracies after artifact rejection. However, the classifier, especially in the case of EEG signals, are not as accurate and accuracy variations across the trials was high. Nonetheless, classification performance was still much greater than would be expected by chance (14.3%). The confusion matrix in Table 2 shows the class assignments for classifying *EEG after artifact removal* with DBN (top) vs. *images* with LSTM (bottom). As can be seen in this table, the model gets easily confused between different classes, especially for Sad (as Disgust, Fear and Anger for around 10%) and Fear (as Disgust and Anger for around 10%). We also note that approximately 10% of each of the non-Neutral emotions (except Surprise) was classified as Neutral, which indicates that the overall emotion information provided through the EEG signals was not very strong. The images still do a much better job in classifying the emotions; only Sad was classified as Neutral for more than 10% of the classifications. Note that the goal of Table 2 is to clarify the cause of the low accuracy achieved by DBN on EEG rather than showing the best cases from Table 1. In other words, the table shows which emotions are confused by the model and cause the low accuracy in those classes using the EEG.

Table 2: Confusion matrix of DBN classifier on filtered EEG (top) and the LSTM classifier on image sequences (bottom)

	Neutral	Sad	Happy	Disgust	Fear	Surprise	Anger
Neutral	**49.8%**	10.3%	0.7%	4.7%	5.1%	19.7%	10.0%
	69.2%	15.1%	0.1%	0.2%	9.9%	0.1%	5.4%
Sad	14.9%	**45.6%**	0.1%	9.6%	9.7%	5.5%	14.3%
	15.1%	**79%**	0.2%	0.3%	5.4%	0.4%	0.1%
Happy	9.6%	5.6%	**59.8%**	0.1%	5.1%	10.3%	9.9%
	0.1%	0.2%	**88.1%**	0.3%	0.2%	10.3%	0.2%
Disgust	10.3%	5.2%	0.0%	**70.6%**	4.9%	0.3%	9.7%
	5.1%	0.2%	0.4%	**84.1%**	0.3%	0.4%	10%
Fear	10.2%	5.1%	5.0%	9.9%	**48.6%**	5.4%	14.9%
	0.2%	5%	0.3%	0.4%	**78.9%**	15%	0.1%
Surprise	0.2%	0.5%	10.3%	4.5%	10.1%	**64.6%**	9.8%
	0.3%	0.1%	5.3%	0.1%	10.3%	**83.9%**	0%
Anger	9.7%	4.9%	0.5%	5.3%	4.9%	4.7%	**69.3%**
	5.3%	5.1%	0.2%	10.3%	0%	0.2%	**78.9%**

The fact that several of the examples were classified incorrectly led us to manually investigate the data and to check if classes with different labels have similarities. We suspected that the emotions do not exactly start or end in the dedicated time slot and might leak into the previous or following slots. For EEG signals, this can particularly be more explainable due to the possible delay in emotional state taking effect in the brain [18], and in more general visual classification tasks [17]. In the next analysis, we verify this hypothesis by analyzing different parts of each sequence separately.

5.4. Dividing data into more meaningful segments

The easiest way to divide each sequence into sensible intervals is by using the beginning and end of the voice signal to mark the sequence. Since we previously found that the maximum length of the utterance was approximately 2.5 seconds, but the data was sampled for 5 seconds after the emotion word onset, we segmented each sample into three parts: *pre-speech*, *during-speech*, and *post-speech*, using the beginning and end of the voice signal as the timestamps to divide the sample. "During-speech" starts as soon as the subject begins uttering the sentence and ends once the utterance ends. This is done by automated speech segmentation. We standardized the length of all "during-speech" segments (by resampling) across the entire dataset so that they were all 2.5 secs. We considered the 1.25 sec segment before the beginning of speech as "pre-speech". The 1.25 sec segment beginning at the end of voice was considered as "post-speech".

Using this segmentation method, we performed a series of analyses where we compared every segment against all the three segments from the same emotion across trials. We repeated this analysis for all modalities, using both DBN and LSTM, for a total of 9x2 results for each modality (except speech). Since i-vector is often used in speech signal classification, we compared the classifiers with the i-vector approach as well. In addition, since we used voice to split the segments, pre-speech and post-speech segments are not meaningful for voice-based emotion recognition; thus, we did not include those combinations in our analyses.

Note that, for instance, when we test post-speech against pre-speech, we randomly chose 60% of the pre-speech segments as the training set and the post-speech segment part of the remaining 40% as the test set. This process was repeated 100 times and the accuracies were averaged. On the other hand, for post-speech against post-speech (or any other matching pair), we randomly chose 60% of the segments for training and the remaining 40% for test, as usual.

Furthermore, since the length of the pre- and post-speech sequences are shorter than during-speech, we padded the shorter sequences to the length of the longer sequence in order to test and train on a non-matching pair (e.g. pre-

against during-speech).

Tables 3 through 6 demonstrate the accuracies of DBN (left) and LSTM (right) classifiers on all modalities, per segment. We compared every segment of an emotion against other segments (including the segment itself) of the same emotion to verify whether the emotion consistently continued over the entire 5 second interval. The following observations should be noted from these results:

(1) The overall observation based on Tables 3 through 6 is that even though all segments that are compared belong to the same emotion, they do not exactly match if the pair is from non-matching segments, i.e. any combination other than pre-vs-pre, during-vs-during and post-vs-post-speech. This observation holds for all modalities. In particular, the large difference between the pre- and post-speech convinced us that the data is contaminated before and after the speech, either by random emotions/facial expressions, or hypothetically by the leakage of each emotion into the following, which result in mismatch between the pre and post- speech segments.

(2) Another important observation is that post-vs-post-speech comparison is always more accurate than pre-vs-pre-speech for all modalities. By manually examining the data, we realized that the subjects tend to keep the same facial expression after the 5-sec duration of the trial, until they fully read and process the label displayed on the next trial and then switch to the next emotion. This makes the post-speech segment more stable compared to the pre-speech segment. We suspect that the same phenomenon happens with EEG signals, similar to the observation made in [17, 18]. To further test this hypothesis, we performed a more thorough analysis. The results will be reported in Sections 5.5 and 5.6.

(3) Interestingly, the post-speech segment is more accurate in classifying the emotions compared to during-speech for EEG signals. This suggests that the EEG response begins taking place slightly later than other modalities and stays active longer or that the movement contaminated EEG signals are not as reliable. We will investigate this more thoroughly later in this section.

(4) For EMG and EEG signals, DBN often does a better job in classifying the emotions correctly. In contrast, LSTM performs better on image sequences. This observation holds for results on both the whole segment and sub-segments. This can be due to generative vs. discriminative capabilities of the models. EEG and EMG require a model with a strong ability to extract hidden information within the data. However, the image sequences and voice signals can readily be classified using LSTM, especially since these images and sounds have already been processed by another deep model, the CNN+ROI platform and MFCC feature extractor, respectively, and valuable information has already been extracted from the data before the LSTM was applied.

Since Tables 3 through 5 have several entries, full segment results from Table 1 are added to those tables for easy comparison.

Table 3: Classification of emotions based on EEG signals with DBN and LSTM. Results are reported as "mean (standard deviation)" in all following tables.

	DBN	LSTM
Pre-vs-pre	**61.3% (6.4)**	55.1% (6.1)
Pre-vs-during	34.7% (5.3)	36.5% (5.3)
Pre-vs-post	36.2% (4.2)	32.3% (4.4)
During-vs-pre	31.3% (2.9)	34.1% (4.2)
During-vs-during	**66.4% (2.7)**	62.7% (3.5)
During-vs-post	56.1% (3.1)	49.1% (3.6)
Post-vs-pre	42% (4.8)	39.2% (7.1)
Post-vs-during	38.9% (4.1)	36% (7.2)
Post-vs-post	**70.9% (6)**	66.5% (3.1)
Full segment	**51.7% (7.3)**	37.6% (8.2)

Table 4: Classification of emotions based on EMG signals

	DBN	LSTM
Pre-vs-pre	**67.9% (4.5)**	64.3% (3.4)
Pre-vs-during	54.3% (5)	49.2% (5.9)
Pre-vs-post	41.2% (6.2)	35.7% (4.8)
During-vs-pre	33.8% (5.1)	37.9% (6.1)
During-vs-during	**77.6% (3.6)**	71.5% (2.5)
During-vs-post	66.2% (4.1)	61.5% (3.3)
Post-vs-pre	47.8% (2.8)	43.4% (4.8)
Post-vs-during	53.8% (5.3)	56.9% (6.1)
Post-vs-post	**72.1% (4.9)**	68.1% (5.7)
Full segment	**62.8% (8.7)**	58.4% (9.1)

Table 5: Classification of emotions based on image sequences

	DBN	LSTM
Pre-vs-pre	59.1% (6.1)	**66.5% (3.3)**
Pre-vs-during	52.2% (8.2)	60.8% (7.1)
Pre-vs-post	44.6% (4.5)	55.2% (5.7)
During-vs-pre	35.7% (4.2)	37.9% (5.6)
During-vs-during	80.1% (5.7)	**89.9% (2.7)**
During-vs-post	51.7% (6.1)	60.1% (5.2)
Post-vs-pre	44.9% (3.3)	57.1% (4.4)
Post-vs-during	41.1% (3.9)	56.5% (3.6)
Post-vs-post	63.2% (7.3)	**73.6% (3.6)**
Full segment	71% (7.7)	**81.1% (5)**

Table 6: Classification of emotions based on voice signals (during speech)

	DBN	LSTM	i-vector
During-vs-during	67.7% (7.2)	**88.7% (3.3)**	76.6% (5.1)

5.5. Unclear boundaries between consecutive emotions

Based on the previous observations (inaccuracy of the comparisons between pairs of different segments), we trained the models on post-speech from the current expression and tested if this emotion can be detected in the pre-speech signal from the next expression. We applied LSTM on image sequences, and DBN on the EEG and EMG signals, since they have shown the best performance on those modalities respectively (Table 7). As can be seen in Table 7, the post-speech of the current emotion signals and the pre-speech of the following emotion signals match

with a surprisingly high accuracy for the EEG data (62.8%). We repeated this same analysis except we switched the training and test sets, i.e., we trained the models on pre-speech from the next emotion (using the current emotion as the label) and tested them on the post-speech from current emotion. The accuracy in this case was 66.1%, which is very close to (actually higher than) the accuracy in our previous analysis (62.8%). This is not the case for EMG or images, with fairly low classification performance.

Table 7: The aftereffect of EEG compared to EMG and Images

	EEG	EMG	Images
Accuracy	**62.8 (5.1)**	34.1 (3.3)	41.2 (5.9)

5.6. Continuation of emotions through time

The EEG aftereffects can be controlled for by giving the subjects enough time to recover from the emotions, as in Palazzo et al., 2017 [16] and Spampinato et al., 2016 [17]. In those studies, the subjects were shown a sequence of images for 25 secs while EEG activity was recorded, followed by a 10 sec pause where a black image was shown. The black image was used to "flush" any high-level class information present from the previous one. We, on the other hand, analyzed the data in order to check the length of this aftereffect by comparing each emotion trial with the next five trials. We performed this analysis for EEG, as well as EMG and images, to show, unlike other modalities, this effect is unique to EEG signals (Table 8). Similar to the previous analysis, we applied LSTM on image sequences and DBN on EEG and EMG.

Table 8: The possible aftereffect of EEG propagated through the next five trials

	Next	2nd next	3rd next	4th next	5th next
Accuracy with EEG	**62.8 (5.1)**	**43.4 (7.3)**	31.2 (9.1)	33.9 (5.9)	21.2 (6.7)
Accuracy with EMG	34.1 (3.3)	22.8 (7.2)	19.5 (5.1)	26.9 (10.8)	21.8 (8.1)
Accuracy with Images	41.2 (5.9)	25.3 (6.2)	26 (4.8)	23.9 (7.7)	18.9 (5.1)

Table 8 shows that the emotion aftereffect is the strongest into the next trial (within 10 seconds), and still has some effect in the n+2 trial (within 15 seconds), but gradually decreases after the n+3 trial. Table 8 also shows that unlike EEG, the EMG and image modalities reflecting a given emotion do not significantly propagate through the next trials and their effect only lasts through the pre-speech segment of the emotion immediately following the current one. Again, we do not track the effect of audio in this case, since the audio signal does not appear throughout the pre- and post-speech segments.

We should note that the pure random chance of each emotion is around 14.3% (1 in 7 emotions) and the results we obtained for EMG and images after the immediate next trial are close to chance and only slightly higher. Note that the probability of the same emotion appearing in the sequence in each of the next 2nd, 3rd, … trials was also 14.3%.

6. Conclusions

This paper presents a thorough study on emotion recognition using four different modalities – audio, video, EMG and EEG. To this end, we collected a dataset with 7 emotion categories, the 4 modalities, and 12 human actor subjects. Both generative models (DBNs) and discriminative models (LSTMs) were applied to the four modalities. Our analyses indicate that LSTM is better for classifying information from audio and video, each with their own sophisticated feature extractors (MFCC and CNN), whereas DBN is better for classifying information from both EMG and EEG. Importantly, we examined how different stages of a trial (pre-speech, during-speech and post-speech) and the following trials affect EEG signals and found long-lasting neural signatures that represent different emotional states.

We believe that the dataset collected in this work can be valuable for affective computing and facial analysis, *thus it will be made publically available following publication*. This paper has focused on the comparison of the four modalities, especially the two bio-sensing datasets (EEG and EMG) versus the commonly used visual and audio data. In particular, one of the most interesting aspects of the analyses is the observation that neural signals conveying an emotion are long-lasting and can be detected by the use of machine learning. This kind of temporal effect has been noted in the psychology and neuroscience literature, but this seems to be the first time it has been exploited by the computer vision community in such a significant capacity.

In the future, we would like to further this research in three directions. The first is to integrate the modalities to optimize performance by using the results of this comparison study. This can be done either at the feature level (early fusion) or the classification level (late fusion). In particular, since LSTM works better on audio and video, and DBN works better on EEG and EMG, it would be interesting to develop models combining generative and discriminative neural networks as in [21], but for emotion recognition. We would like to implement and compare both approaches. The second is to compare machine algorithms and humans in reading the emotion from audio and video, drawing more insights into emotion recognition. The third is to investigate how sensing processing can be improved (especially on EEG and EMG) to obtain more robust signals for reading emotions.

Acknowledgement. This research was supported by NSF EFRI Award #1137172, NSF BCS Awards #1358893 and #1561518, and an Enhanced Chancellors Fellowship from the CUNY Graduate Center.

References

[1] L. Kessous, G. Castellano and G. Caridakis. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces, 3(1), pp.33-48, 2010.

[2] P. Gupta and N. Rajput. Two-stream emotion recognition for call center monitoring. In 8th Annual Conference of the International Speech Communication Association, 2007.

[3] N. Fragopanagos and J. G. Taylor. Emotion recognition in human–computer interaction. Neural Networks, 18(4), pp.389-405, 2005.

[4] T. Vogt and E. André. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In Multimedia and Expo, ICME 2005, (pp. 474-477), 2005.

[5] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier and B. Weiss. A database of german emotional speech. In Interspeech (Vol. 5, pp. 1517-1520), 2005.

[6] W. Li, F. Abtahi and Z. Zhu. Action Unit Detection with Region Adaptation, Multi-labeling Learning and Optimal Temporal Fusing. CVPR 2017. Also arXiv preprint arXiv:1704.03067, 2017.

[7] A. Mollahosseini, B. Hasani and M. H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. arXiv preprint arXiv:1708.03985, 2017.

[8] C. Parlak and B. Diri. Emotion recognition from the human voice. In 21st Signal Processing and Communications Applications Conference (SIU), (pp. 1-4). IEEE, 2013.

[9] D. Bernhardt. Emotion inference from human body motion (Doctoral dissertation, University of Cambridge), 2010.

[10] S. Piana, A. Stagliano, F. Odone, A. Verri and A. Camurri. Real-time automatic emotion recognition from body gestures. arXiv preprint arXiv:1402.5047, 2014.

[11] K. Gouizi, F. Bereksi Reguig and C. Maaoui. Emotion recognition from physiological signals. Journal of medical engineering & technology, 35(6-7), pp.300-307, 2011.

[12] I. Uma. physiological signals based human emotion recognition a review. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 6(1), 2014.

[13] S. Yang and G. Yang. Emotion Recognition of EMG Based on Improved LM BP Neural Network and SVM. JSW, 6(8), pp.1529-1536, 2011.

[14] S. Jerritta, M. Murugappan, K. Wan and S. Yaacob. Emotion recognition from facial EMG signals using higher order statistics and principal component analysis. Journal of the Chinese Institute of Engineers, 37(3), pp.385-394, 2014.

[15] W. Liu, W. L. Zheng and B. L. Lu. Multimodal emotion recognition using multimodal deep learning. arXiv preprint arXiv:1602.08225, 2016.

[16] P. Lahane and A. K. Sangaiah. An approach to EEG based emotion recognition and classification using kernel density estimation. Procedia Computer Science, 48, pp.574-58, 2015.

[17] C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, M. Shah and N. Souly. Deep Learning Human Mind for Automated Visual Classification. 2017 IEEE Conference on Computer Vision and Pattern Recognition. arXiv preprint arXiv:1609.00344, 2016.

[18] S. Palazzo, C. Spampinato, I. Kavasidis, D. Giordano and M. Shah. Generative Adversarial Networks Conditioned by Brain Signals. PDF available on ucf.edu, 2017.

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8), pp.1735-1780, 1997.

[20] G. Hinton, S. Osindero, Y. W. Teh. A fast learning algorithm for deep belief nets. Neural Computation, vol. 18, no. 7, pp. 1527-1554, 2006.

[21] E. P. Giri, M. I. Fanany and A. M. Arymurthy. Combining Generative and Discriminative Neural Networks for Sleep Stages Classification. arXiv preprint arXiv:1610.01741, 2016.

[22] L. Deng and N. Jaitly. Deep discriminative and generative models for pattern recognition. USENIX–Advanced Computing Systems Association, 2015.

[23] Y. LeCun and Y. Bengio. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), p.1995, 1995.

[24] D. Britz. Recurrent Neural Networks Tutorial, Part 1– Introduction to RNNs, 2015.

[25] D. Erhan, P. Manzagol, Y. Bengio, S. Bengio and P. Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. Proc. AISTATS, vol. 5, pp. 153-160, 2009.

[26] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel and P. Ouellet. Front-end factor analysis for speaker verification. Proc. IEEE Trans. on Audio, Speech, and Language, vol. 19, issue 4, pp. 788 – 798, 2011.

[27] Z. Wang, S. Wang and Q. Ji, Q. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. Proc. of the IEEE CVPR (pp. 3422-3429), 2013.

[28] Phinyomark, A., Limsakul, C., Phukpattaranont, P., 2011. Application of wavelet analysis in EMG feature extraction for pattern classification. Measurement Science Review, vol. 11, no. 2, pp. 45-52.

[29] D. E. King. Dlib-ml: A Machine learning toolkit. J. of Machine Learning Research 10, pp. 1755-1758, 2009.

[30] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt and C. L. Zitnick. From captions to visual concepts and back. Proc. of the IEEE CVPR (pp. 1473-1482), 2015.

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[32] K. He, X. Zhang, S. Ren and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778), 2016.

[33] J. Wang. A Tutorial on Speaker Verification. Center for Speech and Language Technologies. Tsinghua University, 2014.

[34] S. O. Sadjadi, M. Slaney and L. Heck. MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research. Microsoft Research Technical Report, 2013.

[35] American Clinical Neurophysiology Society. Guideline 3: minimum technical standards for EEG recording in suspected cerebral death. J Clin Neurophysiol, 23, pp.97-104, 2006.

[36] F. B. Vialatte, J. Solé-Casals and A. Cichocki. EEG windowed statistical wavelet scoring for evaluation and discrimination of muscular artifacts. Physiological Measurement, 29(12), p.1435, 2008.

[37] J. Hu, C. S. Wang, M. Wu, Y. X. Du, Y. He and J. She. Removal of EOG and EMG artifacts from EEG using combination of functional link neural network and adaptive neural fuzzy inference system. Neurocomputing, 151, pp.278-287, 2015.

[38] G. Gómez-Herrero, W. De Clercq, H. Anwar, O. Kara, K. Egiazarian, S. Van Huffel and W. Van Paesschen. Automatic removal of ocular artifacts in the EEG without an EOG reference channel. In Signal Processing Symposium. NORSIG, (pp. 130-133), IEEE, 2006.