Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features

Md Nasir
mdnasir@usc.edu

Arindam Jati
jati@usc.edu

Prashanth Gurunath Shivakumar
pgurunat@usc.edu

Sandeep Nallan Chakravarthula
nallanch@usc.edu

Panayiotis Georgiou
georgiou@sipi.usc.edu

SCUBA, University of Southern California, Los Angeles, CA, USA
scuba.usc.edu

ABSTRACT

Automatic classification of depression using audiovisual cues can help towards its objective diagnosis. In this paper, we present a multimodal depression classification system as a part of the 2016 Audio/Visual Emotion Challenge and Workshop (AVEC2016). We investigate a number of audio and video features for classification with different fusion techniques and temporal contexts. In the audio modality, Teager energy cepstral coefficients (TECC) outperform standard baseline features; while the best accuracy is achieved with i-vector modelling based on MFCC features. On the other hand, polynomial parameterization of facial landmark features achieves the best performance among all systems and outperforms the best baseline system as well.

Keywords

Multimodal signal processing, behavioral signal processing (BSP), depression, Teager energy operator, i-vector, facial landmark, fusion

1. INTRODUCTION

Depression is an affective disorder which has long been recognized as a major concern for individual well-being, often leading to mental disability, morbidity and mortality. It is characterized by the impairment of the patient's ability to cope with stressful life events and also often associated with persistent feelings of negativity, sadness, loss of interest or pleasure and low self-esteem. Depression is also associated with a range of physiological symptoms such as weight loss, insomnia and fatigue. Severe depression is considered one of the leading causes of suicide [19] and substance abuse. It is linked with other psychological disorders such as bipolar affective disorder (also known as manic disorder), dementia [22], and even cardio-vascular conditions [4]. According to a recent report published by World Health Organization (WHO), it is estimated that around 350 million people worldwide are affected by moderate and severe depression [27]. Depression has been identified as a burden to the economy and affects justice and social systems [17]. These reasons have made detection and treatment of depression a high priority towards improvement of the life and health of millions of people.

The standard of depression diagnosis is based on the criteria of the Diagnostic Statistical Manual (DSM) of mental disorders. The outcome of the diagnosis is based on the scores obtained from answering the DSM questionnaire. There are little to no physical tests that directly provide a diagnosis. Studies have shown an alarming rate of false detection in depression diagnosis by general practitioners [30]. The consequences of false detection could be severe. In fact, under-detection could lead to insufficient treatment whereas over-detection could lead to over-treatment, either leading to decreased quality of life. Thus, any valuable contribution to aid in accurate prediction of depression is critical. Recent advances in machine learning, artificial intelligence, and behavioral signal processing allow researchers to view this as a joint human-machine problem and employ a range of modalities to better quantify depression.

The Audio-Visual Emotion Challenge and Workshop (AVEC 2016) provides a research opportunity to investigate a range of signal processing and machine learning methodologies for depression recognition through the Depression Classification sub-Challenge (DCC). In this work, we use data- and knowledge-driven feature extraction methods followed by machine learning techniques including multimodal fusion to predict depression labels from audio-visual information. The primary contribution lies in proposing multimodal features that capture depression behavior cues in a multi-resolution modeling and fusion framework. Teager energy-based and i-vector features along with phoneme rate and duration are proposed to predict depression from audio. On the other hand, polynomial parameterization of temporal variation along with perceptually motivated distance and area features obtained from facial landmarks in video modality is used for the same task. We also investigate the impact of different temporal resolutions on various modalities. We find that a window-based representation of the features with a large temporal context results in better prediction of depression than frame-level analysis. This is in agreement with the slow-varying nature of depressive behavior, as reflected in human audio-visual signals during communication.

The rest of the paper is organized as follows. Section 2 gives a brief review of related research, while Section 3 introduces the dataset. Section 4 and 5 describe the features for the audio and video modalities respectively. Section 6 describes different aspects of the classification systems. Section 7 focuses on the multimodal fusion methods and Section 8 on the regression technique used

AVEC'16, October 16 2016, Amsterdam, Netherlands

ISBN 978-1-4503-4516-3/16/10...$15.00

DOI: http://dx.doi.org/10.1145/2988257.2988261

for depression severity estimation. The experimental results are reported in Section 9. Finally, conclusions are drawn and future directions are given in Section 10 and 11, respectively.

2. RELATED LITERATURE

In recent years, there has been significant interest in research on identification of depression from behavioral signals, namely, speech and visual modality of human communication. Early work on speech-based markers in relation to depression relied on subjective observation and manual or empirical decision making by clinicians [37]. Recently, more data-driven approaches towards depression detection found several speech features to be useful—such as vocal jitter [32], shimmer [38], harmonics-to-noise ratio (HNR) [1] *etc.* The usefulness of these voice quality features was consistent with the hypothesis that depressed individuals tend to have unnatural and monotonous speech patterns [8]. Further research [31, 35] showed that characterization of glottal flow waveform improves discrimination of depressed speech. Furthermore, pause duration [44] and global speech rate [45] are known to be vocal biomarkers of a depressed individual.

Along with speech and vocal cues, depression is also reflected in facial expression, head movement, eye gaze and gesture. A significant amount of literature is available on detecting depression from facial expressions. Alghowinem *et al.* [2] used yaw, roll, pitch and their statictics to recognize depression from head pose and movement. Ooi *et al.* [34] applied eigenface and Fisherface methods on the facial images of adolescents to detect the risk of being depressed. Hamm *et al.* [18] developed automated facial expression and action units (AU) recognition system to find and analyze neuropsychiatric disorders, which are sometimes actively or passively related to depression.

A good amount of research has been done in multimodal fusion of audio and video features to detect and analyze depression. Kächele *et al.* [20] have used a hierarchical classifier model to find the state of depression using audio-visual fusion. Vocal prosody and facial action units have also been used in [5] to detect depression. Meng *et al.* [28] used Motion History Histogram (MHH) dynamical features [29] to recognize depression from both audio and video.

3. DATASET

The current work uses the publicly available multimodal depression data set, Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [16], which is a collection of interviews of individuals conducted by Ellie, a virtual human designed to help diagnosis of psychological distress conditions. The dataset is annotated with self-reported PHQ-8 [23] scores of depression as well as binarized depression ratings of the subject. The challenge primarily addresses the automatic classification problem of depression (*depressed vs. not depressed*) from audio and video modalities.

4. AUDIO FEATURES

4.1 Baseline features

The baseline audio features for the DCC challenge include a set of features extracted using COVAREP toolbox. It consists of several prosodic features (fundamental frequency and voicing) and some knowledge-driven voice quality features (such as glottal features and formants), along with standard spectral features like MFCCs and harmonic model features. The complete list of these features can be found in [46].

4.2 Extended spectral and prosodic features

Since the baseline audio feature set does not include a few features deemed useful for depression prediction (such as jitter, shimmer *etc.*), we choose to extract some additional features. We use the baseline feature set from INTERSPEECH 2013 computational paralinguistics challenge (ComParE) [39]. This set consists of pitch, energy, spectral, cepstral coefficients (MFCCs) and voicing related frame-level features referred to as low-level descriptors (LLDs). Some other LLDs include logarithmic harmonic-to-noise ratio (HNR), spectral harmonicity, and psychoacoustic spectral sharpness. We use the OpenSMILE toolkit [12] to extract all these features.

4.3 Teager energy cepstral coefficients

The *Teager Energy Operator* (TEO) is motivated by the energy in an oscillating system and has found use in many speech applications [15,50]. This nonlinear energy operator (Ψ), proposed by Teager and Kaiser [21], is defined below for a given signal or feature stream $x[n]$.

$$\Psi(x[n]) = x[n]^2 - x[n-1]x[n+1] \quad (1)$$

Because of its nonlinearity property, TEO can robustly track rapid changes in a local context, which is specifically useful in presence of noise. Teager-energy cepstrum coefficients (TECCs) were introduced in [15] as a robust alternative to MFCCs in an emotion recognition application in noisy conditions. We adopt these features for our depression prediction system. To compute TECCs, first the spectrum of the speech signal is computed and then TEO is applied before the remaining stages of cepstral coeffcient computation in Mel scale. The process also involves a pre-processing stage including frame blocking, windowing with a hamming window and pre-emphasis.

4.4 Feature representation with a large window

All the features discussed so far are extracted at a high temporal resolution due to the non-stationarity of speech signal. We use a 25 ms sliding frame with 10 ms shift to perform the feature extraction. However, it is difficult to characterize depression at such a fine resolution. Unlike emotion, which can change rapidly within a short speech segment (within 2 seconds, for example), expression of depression in speech is a much more slowly varying phenomenon. For this reason, we choose to integrate the vocal information within a large window containing multiple frames, consistent with prior work on behavioral signal processing [25]. In this work, we choose a window of 10s duration and 5s shift and combine features from all frames within that window by taking their arithmetic mean. We will refer to this feature representation as *window-level* features throughout this paper, as opposed to the very short-term *frame-level* representation.

We mask the virtual human speech and the non-speech regions to consider only the speech segments of the subject. We use the start and end time-points of continuous utterances from the transcripts to obtain voice activity detection (VAD) labels.

All the windows from a certain session can be used for training using the label of that session (*depressed* vs. *not depressed*) since depression can be considered a user state that does not change rapidly.

4.5 i-vector modeling

Recently i-vector modeling using total variability framework was proposed for speaker verification task [9] with state-of-the-art performance. The total variability framework provides an effective way to capture speaker variability and channel

Table 1: **Different audio feature sets**

Feature	Scope of analysis	Dimension
COVAREP	10*s*-window-level	74
Formant features	10*s*-window-level	5
Baseline	10*s*-window-level	79
IS ComParE 2013	10*s*-window-level	130
i-vector	10*s*-window-level	50 - 600
eGeMAPS	session-level	88
Phone features	session-level	17
session-level features	-	105

variability in a low dimensional sub-space. We hypothesize that the vocal characteristics of depression are reflected in the spectral variability over time. Previous analysis of acoustic properties of depressed subjects has shown spectral properties like amplitude modulation, formants and spectral power distribution to be potential discriminators from neutral speech [7,13]. The i-vector framework has been shown to be extremely effective in capturing spectral variability cues in application to speaker recognition [41], language recognition [10], native language identification [42] and speaker age classification tasks [43]. In [49], the i-vectors were used to detect emotions in speech. In this work, we employ total variability modeling to capture the variability between *depressed* and *non-depressed* speakers. The i-vector extraction process could be represented as follows:

$$M = m + Tv, \tag{2}$$

where m is the mean super-vector of the Gaussian mixture model-universal background model (GMM-UBM). M is the mean centered super-vector of the speech utterance derived using the 0^{th} and 1^{st} order Baum-Welch statistics. T is the low rank total variability matrix representing the speaker and channel variabilities trained using Expectation-Maximization (EM). v is the i-vector representation of the speech utterance.

In our work, the UBM and the total variability matrix are trained on the in-domain data. To compensate for the data sparsity, we over-sample our training set by perturbing the raw audio files by spectral warping and addition of white Gaussian noise at different signal-to-noise ratios. The *depressed* class is oversampled more than the *not depressed* class to compensate for unbalanced class distribution. 13 dimensional MFCC features are extracted using a frame size of 25ms and an overlap of 10ms. Additionally, the first and second order derivatives (Δ and Δ-Δ) of MFCCs are computed to give 39 dimensional feature vector. Voice activity detection (VAD) labels are inferred from the transcripts to mask the non-speech regions, similar to window-level and session-level analysis. The GMM-UBM component dimensions are experimented upon, ranging from 512 to 2048. The total variability subspace rank is experimented for dimensions in the range 50 - 600.

4.6 Session-level acoustic features

Since depression reference ratings are provided per subject at a global level, we also investigate classification based on features extracted from the entire session. The major challenge for such an approach is the small sample size (number of sessions) of the dataset, which has led us to use a small minimalistic feature set. We use the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [11] for this purpose, which has showed promising results in paralinguistic applications. It consists of various functionals over standard spectral, cepstral, prosodic and voice quality features and is extracted from OpenSMILE toolkit [12] for all speech frames in a session.

4.7 Phoneme-based features

4.7.1 Text-audio alignment

We first utilize the transcript to extract the continuous utterances spoken by the subject. Then, we use the start and end time-points of each utterance to obtain the corresponding audio segment. Finally, we run the Gentle forced aligner [33] on these transcript-audio pairs to obtain the start and end time-points of each phoneme in each utterance.

4.7.2 Phoneme rate and duration

We extract phoneme rate and phoneme duration information at the utterance-level as well as the session-level. We obtain the utterance-level phoneme rate by dividing the number of phonemes present in each utterance by the corresponding utterance duration. Statistical functionals (mean, variance, median, min, max, range, skewness and kurtosis) of these utterance-level features (phone rate and duration) are then computed across the entire session. Finally, we obtain the session-level phoneme-rate by dividing the total number of phonemes by the total duration over all turns. This results in a 17-dimensional (2×8+1) feature vector per session.

In order to account for the class size imbalance in the dataset, we create smaller-length sessions for the *depressed* class by splitting each original session into multiple contiguous sub-sessions. This step results in nearly equal number of training samples in both classes.

5. VIDEO FEATURES

Since no raw video is made publicly available for the AVEC 2016 DCC, we use the baseline video features to derive meaningful meta features as described below.

5.1 Baseline features

Two sets of baseline video features are provided in the challenge. The first one contains facial landmarks (2D and 3D), Histogram Oriented Gradient (HOG) features, estimated gaze direction, and 3D position and orientation of the head. These are computed from the raw video using the OpenFace [3] toolkit. The second set consists of continuous measures of 20 facial action units (AUs) and emotions estimated with the FACET [26] toolkit.

5.2 Marker velocity and acceleration features

We use the same 10*s*-window level feature extraction technique as discussed in Section 4.4. Table 2 shows the dimensions of different feature sets we have extracted along with their assigned names. *FacialMarker1* contains the baseline OpenFace features averaged over the frames in a window. These are concatenated with their velocity (first order derivative) and acceleration (second order derivative) to create *FacialMarker2* features. For convenience, we will refer to velocity and acceleration as Δ and Δ-Δ respectively throughout the paper. The *AU* feature set is constructed in similar manner from the raw FACET features. To boost the classification performance, we extract additional features as described below.

5.3 Polynomial parameterization on the facial marker features

To capture the temporal variation of the facial expression, polynomial fitting is performed on the features of different frames in a particular window and the coefficients of the fitted polynomial are used as new features [47]. The central idea is to capture the temporal variation of facial expression rather than the position of the landmarks at any particular time. The 3D positions of the 68 facial markers, along with eye gaze and head pose are taken from the baseline features provided. For every feature, all the values in

Table 2: **Different video feature sets**

Feature set name	Description	Dimension
FacialMarker1	3D landmarks	204
	Eye gaze	12
	Head pose	6
	Total	**222**
FacialMarker2	FacialMarker1 + Δ + Δ-Δ	**666**
Polyfit	FacialMarker1	222
	Polynomial parameterization	**666**
AU	AUs	30
	Δ AUs	30
	Δ-Δ AUs	30
	Total	**90**
Geometrical	Distance	11
	Δ distance	11
	Area	26
	Δ area	26
	Total	**74**
VideoFeatSet1	FacialMarker2+Geometrical	**740**
VideoFeatSet2	VideoFeatSet1+Polyfit	**1406**

a particular window (10s) are fitted into a 2nd order polynomial and the three coefficients of the polynomial are taken into account. So, the three coefficients carry the information about the temporal variation of the corresponding feature in a compact way. In total, this generates a 666 dimensional feature vector (termed *Polyfit*) for every window.

5.4 Geometrical features

We also extract geometrical features, namely distance and area features, from the face. This set contains distances between particular points as shown in Fig. 1a and areas formed by joining specific points as in Fig. 1b [48]. Clearly, the distance features carry information about mouth opening and closing, mouth stretching, eye opening and closing, eyebrow lifting *etc.* Similarly, the area features reflect changes in face topology. Our hypothesis is that these features capture information about facial expressions relevant to depression detection. 11 distance features and 26 area features are extracted from the x and y co-ordinates of the 68 landmark points (please see Table 2).

The distances are normalized by the width of the face since the face width does not change in different expressions or emotions [48]. The areas are normalized by the mean area of the face of a particular person over the whole session. So, these features are less dependent on different face structures and face areas of different persons as well as invariant to the individual differences in face topologies for the same expression.

The Δ features are also taken into consideration to capture the nature of change in the face topology. As shown in Table 2, the dimension of the whole geometrical feature vector is 74.

6. CLASSIFICATION

6.1 Feature selection

Because of the high dimensionality of the video feature sets and consequent risk of overfitting, we perform feature selection to choose a reduced subset of features before classification. This strategy is specifically useful for facial marker feature sets. We use Mutual Information Maximization (MIM) based feature selection method [24], where every feature X_k is evaluated based on a mutual information score $J_{MIM} = I(X_k;Y)$ based on the class label Y. This feature selection procedure is also used while performing session-level classification, as the number of sessions becomes comparable to the number of features used.

6.2 Classifier: Support Vector Machine

6.2.1 Window-level: SGD-SVM

A linear support vector machine (SVM) classifier with stochastic gradient descent (SGD) learning is used for all feature sets except for the session level feature sets and the i-vector system. The hyperparameters of the classifier (α, loss function and regularizer) are tuned by a grid search on 10-fold cross-validation of train and development set. Then training is performed only on the training data with optimized parameters. This process is also similar to the baseline system [46]. The implementation of SGD-SVM in Scikit-learn [36] toolbox is used for this purpose. The use of such a fast and simple classifier enables us to experiment on various feature sets more efficiently.

6.2.2 Session-level: Kernel-SVM

Since we have small number of data samples in session-level system, we use support vector machine (without SGD) for classification. Different types of kernels (linear, polynomial, RBF) are used with SVM. The choice of the kernel and other hyperparameters including C, γ *etc.* are optimized by grid search with cross-validation.

6.3 Gaussian Probabilistic LDA

Gaussian Probabilistic Linear Discriminant Analysis (G-PLDA) models have shown to be effective classifiers for i-vector based systems [40] which constitute the current state-of-the-art speaker recognition systems [14]. The i-vectors extracted were length-normalized as in [14] and used in a G-PLDA framework for classification. To compensate for the limited sample size for training the G-PLDA model, each speaker session is split using 10 second window with 5 second overlap. The output of the G-PLDA system consists of the log-likelihoods of both classes for each session. For final classification, the log-likelihoods of the constituent splices representing a session are fused as described in section 7.2

7. FUSION SCHEMES

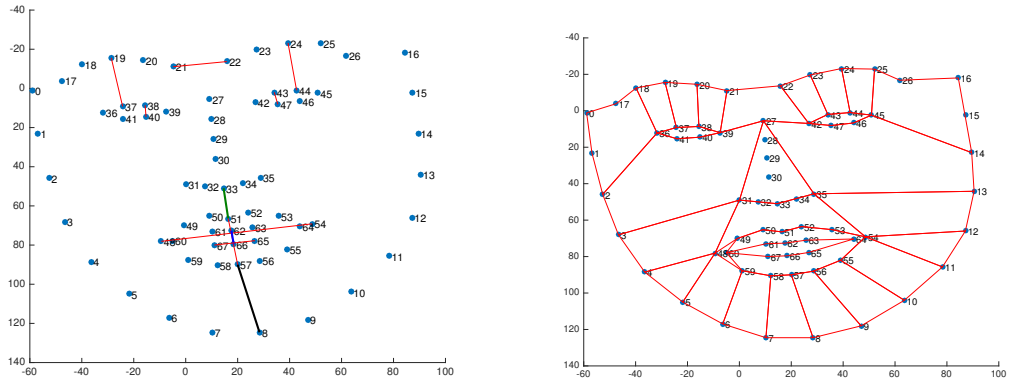
7.1 Feature-level fusion

Feature-level fusion is used for the different feature sets of the video modality. Table 2 shows the description of the fused feature sets for video. *FacialMarker2* and geometrical features are concatenated to create *VideoFeatSet1* feature set. Finally, the *Polyfit* features are appended with the *VideoFeatSet1* features to create *VideoFeatSet2* feature set. The reason for not using the *AU* features in early fusion is their poor performance (discussed in Section 9) on the development set. Moreover, feature level fusion is performed while using session-level audio features and phoneme features.

7.2 Temporal fusion

The majority of the classification experiments are performed at a finer temporal resolution than the originally provided depression labels. In other words, the class labels are given per session, while window-level classification predicts a label for each window. Therefore, we can employ predictions of all windows throughout a given session to derive a single session-level *depressed* or *not depressed* label. The signed confidence score $C(\mathbf{x})$ for a feature vector $\mathbf{x}$ is defined as a function of the class posterior probabilities against the class label y as shown below:

$$C(\mathbf{x}) = logP(\mathbf{x}|y=1) - logP(\mathbf{x}|y=0), \quad (3)$$



(a) 11 distance features

(b) 26 area features

Figure 1: Geometrical features computed from the facial landmarks.

where 1 and 0 indicate *depressed* or *not depressed* classes respectively. For all the window-based classification experiments, class-posterior log-probabilities are added for all windows in a given session and the class that maximizes the cumulative posterior log-probability is chosen as the hypothesized class. Similarly, for the ivector classification, the log-likelihood scores from the G-PLDA system are summed for all the splices constituting a session and the class that maximizes the total log-likelihood score is selected. This confidence-based technique proved to be more robust and effective compared to the simple majority voting for both SVM and i-vector systems.

7.3 Multimodal late fusion

We perform late-fusion of session-level decisions from multiple modalities to investigate its effectiveness at predicting session-level depression ratings. We first build depression-prediction SVM models for each modality separately. Then, we obtain classification-confidence scores at window-level and session-level for video and audio modalities respectively.

In the video stream, we uniformly resample the window-level confidence scores to obtain 10 scores per session. From the audio stream, we obtain 2 confidence scores per session, one each for i-vector and phoneme features. Finally, we use these scores in a classification scheme to predict whether a session belongs to *depressed* or *not depressed* class. Thus, we also examine if the phoneme stream provides complementary information to the audio and video by testing the system with a fusion of only the latter two modalities.

To account for class-size imbalance, we randomly select a subset of the *not depressed* samples that is equal in size to that of the *depressed* class. We then train an RBF-kernel SVM model on these samples and predict the labels for the development set samples. The final prediction labels are obtained by taking a majority vote over 51 independent runs.

In addition, we also perform simple logical AND operation between predictions made by audio and video modalities, similar to the baseline system.

8. RANDOM FOREST REGRESSION

Random Forest estimator is an ensemble learning method which fits multiple decision tree classifiers/regressors by random selection of features and optimizes by bagging and aggregating the results. The number of estimators are tuned using grid search on 10-fold cross-validation of train and development set, and the best parameters are retained and used for final regression. The system setup is identical to the baseline system [46]. The trimmed mean of regression scores of constituent windowed samples is used to get the regression score representing the corresponding session. The regression performance is evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

9. RESULTS AND DISCUSSION

9.1 Classification results

The results of each modality are presented in this section. Since the dataset has a very skewed distribution of classes, we provide unweighted and weighted F1 scores, precision, and recall.

9.1.1 Audio system

Table 3 illustrates the performance of the proposed audio systems. All audio feature sets except IS2013 *ComParE* features outperform the baseline. TECC features alone perform better than the baseline, while the session level systems perform better than any of window-level SVM classifier systems. We also achieve improvement in performance by the use of phoneme features in addition to acoustic features for session-level classification.

The i-vector - PLDA system outperforms the baseline by a significant margin. We observe that the i-vector system provides improvement to both the *depressed* and *not depressed* classes, thereby asserting the robustness of the system. An absolute increase of 17% is achieved with F1 score metric for the *depressed* class whereas for the *not depressed* class the performance increases by an absolute 32%.

On test dataset, the best performing audio system achieves F1 scores of 0.48 and 0.85 for *depressed* and *not depressed* classes respectively, which are much above the baseline performance (0.46 and 0.68).

9.1.2 Video system

A comparison of the performances of different video feature sets along with the baseline is presented in Table 4. The results are obtained after fusing the results of 100 independent runs of SGD-SVM (Section 6.2) on the development set. We observe poor performance of the *AU* feature set. We further notice that *FacialMarker2* achieves better performance than *FacialMarker1*, indicating that Δ and $\Delta - \Delta$ of facial markers are informative. The geometrical features perform better than facial marker feature sets, despite a comparatively smaller dimensionality.

Table 3: **Results on development set for audio (values for *not depressed* class are shown in parentheses)**

Features	Classifier	F1 score	Precision	Recall	Weighted F1 score
Baseline	SVM-SGD	0.41 (0.58)	0.27 (0.94)	0.89 (0.42)	0.55
IS2013 ComPareE	SVM-SGD	0.40 (0.84)	0.38(0.85)	0.43 (0.82)	0.75
TECC features	SVM-SGD	0.43(0.86)	0.43(0.86)	0.43(0.86)	0.77
eGeMAPS	SVM-polynomial	0.46 (0.88)	0.5 (0.86)	0.43 (0.89)	0.80
eGeMAPS+phone	SVM-polynomial	0.47 (0.83)	0.4 (0.88)	0.57 (0.78)	0.76
i-vector (MFCC)	G-PLDA	**0.57 (0.89)**	**0.57 (0.89)**	**0.57 (0.89)**	**0.83**

Table 4: **Results on development set for video (values for *not depressed* class are shown in parentheses)**

Feature set name	F1 score	Precision	Recall	Weighted F1 score
Baseline	0.58 (0.86)	0.47 (0.94)	0.78 (0.79)	0.81
FacialMarker1	0.36 (0.71)	0.27 (0.85)	0.57 (0.61)	0.64
FacialMarker2	0.38 (0.73)	0.29 (0.86)	0.57 (0.64)	0.66
Polyfit	0.38 (0.73)	0.28 (0.86)	0.57 (0.64)	0.66
AU	0.11(0.69)	0.09 (0.75)	0.14 (0.64)	0.57
Geometrical	0.40 (0.76)	0.31 (0.86)	0.57 (0.67)	0.69
VideoFeatSet1	0.48 (0.78)	0.36 (0.91)	0.71 (0.67)	0.72
VideoFeatSet2	0.42 (0.78)	0.33 (0.87)	0.57 (0.71)	0.71
VideoFeatSet2 + Feature selection	**0.63 (0.89)**	**0.56 (0.92)**	**0.71 (0.86)**	**0.84**

Table 5: **Regression results on development set**

Modality	Features	Regressor	MAE	RMSE
Video	Baseline	Random Forest	5.8767	7.1332
	VideoFeatSet2 + Feature Selection	Random Forest	6.4799	7.8644
Audio	Baseline	Random Forest	5.3566	6.7418
	i-vector PLDA log-likelihood scores	Linear Regression	5.8237	6.7334

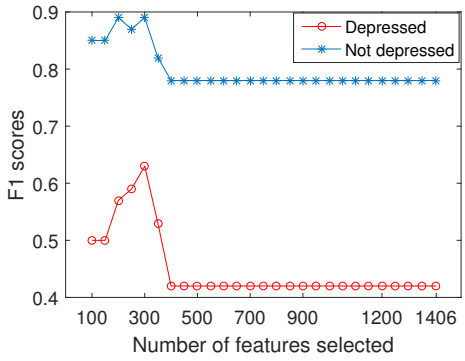


Figure 2: Variation of F1 scores for two classes on the development set with number of features selected.

Table 6: **F1 scores on development set for mulimodal late fusion (values for *not depressed* class are shown in parentheses)**

System	F1 score
Audio only	0.57 (0.89)
Video only	**0.63 (0.89)**
Late fusion 1 (A+V)	0.40 (0.84)
Late fusion 2 (A+V)	0.40 (0.84)
AND(A,V)	**0.63 (0.89)**

Using geometric features along with facial marker feature sets (*VideoFeatSet1*) achieves improved F1 score over using either of them separately. However, further adding polynomial parametric feature sets (*VideoFeatSet2*) leads to decline in the performance, possibly because of overfitting caused by the high dimensionality of the feature set and limited training data. To address this, feature selection (Section 6.1) is applied on this feature set. Fig. 2 shows the variation of F1 scores of the two classes on the development set with number of features selected. We start with 100 features and successively add 50 features in each iteration. At every step we run SGD-SVM 100 times and obtain the final result by fusing different runs.

The best performance is obtained with 300 features, with F1 scores of 0.63 and 0.89 for *depressed* and *not depressed* classes respectively. We get a 5% increase in F1 score for the *depressed* class and 3% increase for the *not depressed* class over the baseline performance.

We obtained F1 scores of 0.29 and 0.80 on test set using the video modality. The comparatively poor performance might be because of data mismatch between development and test dataset.

9.2 Multimodal late fusion results

Multimodal fusion is performed on the best individually performing audio and video systems.

The methods discussed in Section 7.3 do not seem to improve the performance. The results are shown in Table 6, where first and second fusion schemes refer to majority voting and summation of confidence scores across multiple runs, respectively. Based on our initial analysis, it appears that the decline in performance after fusion might be because of inconsistency of confidence scores across individual modalities. For example, the audio system seems to have higher confidence value with a wrongly classified sample, whereas the prediction of the video system turns out to be correct, yet with a lower confidence score–causing a misclassification of the sample.

However, we obtain F1 scores of 0.63 and 0.89 (for *depressed* and *not depressed* classes respectively) after logical AND-based fusion between two modalities, which is similar to the performance of the video modality itself. This observation is also consistent with baseline fusion.

9.3 Regression results

Table 5 lists the regression performance of our best-performing

audio and video systems. In both the cases, our systems are comparable to the baseline. We did not create a multimodal system for regression and neither did we optimize the individual systems for regression, and hence the gains we see in classification do not translate to regression. We observe that in our systems the audio modality is a better regressor compared to the video, while in the classification task video modality performs better, similar to the baseline results.

10. CONCLUSIONS

In this paper, we propose a multitude of features for depression classification, addressing the AVEC 2016 depression sub-challenge (DCC). The multimodal classification system performs better than the baseline systems on the development dataset. The i-vector system performs the best for the audio modality, while polynomial parameterization of facial landmarks along with geometrical features turns out to be the best video feature set. Our model considers a temporal context of overlapping windows to integrate information relevant to depression. The experiments show that this approach outperforms the baseline approach of frame-level analysis. The contribution of this work is proposing potentially robust and knowledge-driven feature sets in both audio and visual modalities, which may be used in conjunction with a more sophisticated classifier to achieve even better classification.

11. FUTURE DIRECTIONS

In future we plan to use class specific UBM for i-vector extraction. The idea behind this approach is to examine the projection of the *depressed* class statistics on the *not depressed* class and vice-versa. Speaker-level normalization of features to remove inter-speaker variability could potentially improve the performance of the systems [6]. Vocal Tract Length Normalization (VTLN) for MFCC features under the i-vector framework [42] is a potential audio normalization technique to be investigated. Observing the effectiveness of TECCs over MFCCs, we intend to use TECC as front-end features to i-vectors for depression detection task. Also applying i-vector modeling to video features can be a future direction after obtaining promising results on audio data.

12. ACKNOWLEDGMENTS

This research is supported by NSF and DoD.

13. REFERENCES

[1] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker. Detecting depression: a comparison between spontaneous and read speech. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7547–7551. IEEE, 2013.

[2] S. Alghowinem, R. Goecke, M. Wagner, G. Parkerx, and M. Breakspear. Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction (ACII), Humaine Association Conference on*, pages 283–288. IEEE, 2013.

[3] T. Baltru, P. Robinson, L.-P. Morency, et al. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2016.

[4] R. M. Carney, K. E. Freedland, G. E. Miller, and A. S. Jaffe. Depression as a risk factor for cardiac mortality and morbidity: a review of potential mechanisms. *Journal of psychosomatic research*, 53(4):897–902, 2002.

[5] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre. Detecting depression from facial actions and vocal prosody. In *3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009.

[6] N. Cummins, J. Epps, M. Breakspear, and R. Goecke. An investigation of depressed speech detection: Features and normalization. In *Interspeech*, pages 2997–3000, 2011.

[7] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke. Modeling spectral variability for the classification of depressed speech. In *Interspeech*, pages 857–861. Citeseer, 2013.

[8] J. K. Darby, N. Simmons, and P. A. Berger. Speech and voice parameters of depression: A pilot study. *Journal of Communication Disorders*, 17(2):75–85, 1984.

[9] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2011.

[10] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak. Language recognition via i-vectors and dimensionality reduction. In *INTERSPEECH*, pages 857–860, 2011.

[11] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2016.

[12] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462. ACM, 2010.

[13] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and M. Wilkes. Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE transactions on Biomedical Engineering*, 47(7):829–837, 2000.

[14] D. Garcia-Romero and C. Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *Interspeech*, pages 249–252, 2011.

[15] A. Georgogiannis and V. Digalakis. Speech emotion recognition using non-linear teager energy based features in noisy environments. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2045–2049. IEEE, 2012.

[16] J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, et al. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128, 2014.

[17] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler. The economic burden of adults with major depressive disorder in the united states (2005 and 2010). *The Journal of clinical psychiatry*, 76(2):155–162, 2015.

[18] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of neuroscience methods*, 200(2):237–256, 2011.

[19] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders. Risk factors for suicide in individuals with depression: a

systematic review. *Journal of affective disorders*, 147(1):17–28, 2013.

[20] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker. Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. *depression*, 1(1), 2014.

[21] J. F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. In *Acoustics, Speech, and Signal Processing, ICASSP-1990.*, pages 381–384 vol.1, Apr 1990.

[22] V. Kral. The relationship between senile dementia (alzheimer type) and depression. *The Canadian Journal of Psychiatry/La Revue canadienne de psychiatrie*, 1983.

[23] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1):163–173, 2009.

[24] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of the workshop on Speech and Natural Language*, pages 212–217. Association for Computational Linguistics, 1992.

[25] H. Li, B. Baucom, and P. Georgiou. Sparsely connected and disjointly trained deep neural networks for low resource behavioral annotation: Acoustic classification in couplesâĂŹ therap. In *accepted in Interspeech*, 2016.

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

[27] M. Marcus, M. Taghi Yasamy, M. van Ommeren, et al. Depression: A global public health concern. geneva, switzerland: Who department of mental health and substance abuse; 2012, 2015.

[28] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, pages 21–30. ACM, 2013.

[29] H. Meng and N. Pears. Descriptive temporal template features for visual motion recognition. *Pattern Recognition Letters*, 30(12):1049–1058, 2009.

[30] A. J. Mitchell, A. Vaze, and S. Rao. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*, 374(9690):609–619, 2009.

[31] E. Moore II, M. A. Clements, J. W. Peifer, and L. Weisser. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE transactions on biomedical engineering*, 55(1):96–107, 2008.

[32] Å. Nilsonne, J. Sundberg, S. Ternström, and A. Askenfelt. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. *The Journal of the Acoustical Society of America*, 83(2):716–728, 1988.

[33] R. M. Ochshorn and M. Hawkins. Gentle forced aligner.

[34] K. E. B. Ooi, L.-S. A. Low, M. Lech, and N. Allen. Prediction of clinical depression in adolescents using facial image analaysis. In *WIAMIS 2011, Delft, The Netherlands, April 13-15, 2011*, 2011.

[35] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. *IEEE Transactions on Biomedical Engineering*, 51(9):1530–1540, 2004.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[37] B. Pope, T. Blass, A. W. Siegman, and J. Raher. Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology*, 35(1p1):128, 1970.

[38] T. F. Quatieri and N. Malyska. Vocal-source biomarkers for depression: A link to psychomotor activity. In *Interspeech*, pages 1059–1062, 2012.

[39] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, et al. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. 2013.

[40] M. Senoussaoui, P. Kenny, N. Brümmer, E. De Villiers, and P. Dumouchel. Mixture of plda models in i-vector space for gender-independent speaker recognition. In *INTERSPEECH*, pages 25–28, 2011.

[41] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel. An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. In *Odyssey*, page 6, 2010.

[42] P. G. Shivakumar, S. N. Chakravarthula, and P. Georgiou. Multimodal fusion of multirate acoustic, prosodic, and lexical speaker characteristics for native language identification. In *INTERSPEECH*, 2016.

[43] P. G. Shivakumar, M. Li, V. Dhandhania, and S. S. Narayanan. Simplified and supervised i-vector modeling for speaker age regression. In *ICASSP*, pages 4833–4837. IEEE, 2014.

[44] E. Szabadi, C. Bradshaw, and J. Besson. Elongation of pause-time in speech: a simple, objective measure of motor retardation in depression. *The British Journal of Psychiatry*, 129(6):592–597, 1976.

[45] A. C. Trevino, T. F. Quatieri, and N. Malyska. Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, 2011(1):1–18, 2011.

[46] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. T. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic. Avec 2016-depression, mood, and emotion recognition workshop and challenge. *arXiv preprint arXiv:1605.01600*, 2016.

[47] M. F. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *International Workshop on Human-Computer Interaction*, pages 118–127. Springer, 2007.

[48] P. Wang, F. Barrett, E. Martin, M. Milonova, R. E. Gur, R. C. Gur, C. Kohler, and R. Verma. Automated video-based facial expression analysis of neuropsychiatric disorders. *Journal of neuroscience methods*, 168(1):224–238, 2008.

[49] R. Xia and Y. Liu. Using i-vector space model for emotion recognition. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[50] G. Zhou, J. H. Hansen, and J. F. Kaiser. Classification of speech under stress based on features derived from the nonlinear teager energy operator. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 549–552. IEEE, 1998.