Diagnosis of Depression by Behavioural Signals: A Multimodal Approach

Nicholas Cummins
School of Electrical Eng. and Tele.
The University of New South Wales
and National ICT Australia
n.p.cummins@unsw.edu.au

Jyoti Joshi
Human-Centred Comp. Lab
University of Canberra
Bruce ACT 2617 Australia
jyoti.joshi@canberra.edu.au

Abhinav Dhall
Res. School of Computer Science
Australian National University
Canberra ACT 2601 Australia
abhinav.dhall@anu.edu.au

Vidhyasaharan Sethu
School of Electrical Eng. and Tele.
The University of New South Wales
Sydney NSW 2052 Australia
v.sethu@unsw.edu.au

Roland Goecke
Human-Centred Comp. Lab
University of Canberra
Res. School of Computer Science
Australian National University
roland.goecke@ieee.org

Julien Epps
School of Electrical Eng. and Tele.
The University of New South Wales
and National ICT Australia
j.epps@unsw.edu.au

ABSTRACT
Quantifying behavioural changes in depression using affective computing techniques is the first step in developing an objective diagnostic aid, with clinical utility, for clinical depression. As part of the AVEC 2013 Challenge, we present a multimodal approach for the Depression Sub-Challenge using a GMM-UBM system with three different kernels for the audio subsystem and Space Time Interest Points in a Bag-of-Words approach for the vision subsystem. These are then fused at the feature level to form the combined AV system. Key results include the strong performance of acoustic audio features and the bag-of-words visual features in predicting an individual's level of depression using regression. Interestingly, in the context of the small amount of literature on the subject, is that our feature level multimodal fusion technique is able to outperform both the audio and visual challenge baselines.

Categories and Subject Descriptors
G.3 [**Mathematics of Computing**]: Probability and Statistics—*Correlation and regression analysis*; *Robust regression*
I.4.7 [**Computing Methodologies**]: Image Processing and Computer Vision—*Feature measurement: Feature representation*
I.5.4 [**Computing Methodologies**]: Pattern Recognition – *Applications: Signal processing; Computer vision; Waveform analysis*
J.3 [**Computer Applications**]: Life and Medical Sciences—*Medical information systems*;

General Terms
Algorithms, Measurement, Performance, Design, Reliability, Experimentation, Verification.

Keywords
Depression, Behavioural Signals, Multimodal Technologies, Acoustic Speech Features, Space-Time Interest Points, Pyramid of Histogram of Gradients, Bag-of-Words, Multimodal Fusion, Support Vector Regression.

ACM-Multimedia 2013 Barcelona, Spain

1. INTRODUCTION
Clinical depression has long been recognized as one of the leading causes of disability and burden worldwide; it has been estimated that depression will be one of the three leading causes of burden by disease along with HIV/AIDS and heart disease by 2030 [1]. Despite these growing socio-economic costs, diagnosis is achieved, almost exclusively, on the basis of an interview style assessment between a clinician and a patient, and patient self-reporting. These assessments attempt to assign an objective score to a patient based on a weighted sum of key symptoms observed in depression [2], including but not limited to; negative conceptualizations, fatigue, cognitive impairments and observable psychomotor retardation [3]. In practice, these tests are subjective in nature, requiring a large degree of clinical training to produce acceptable results, and can be conducted only infrequently [4].

It is believed that finding a set of objective markers of depression will increase diagnostic accuracy, aiding optimal patient care. Whilst research into biological markers for depression has revealed several promising results, such as low serotonin levels, no biomarker specific to depression has been found [5]. Whilst biomarkers remain elusive, significant advances have been made in using affective computing and social signal processing as a diagnostic tool. These systems rely either on speech processing techniques [6]–[8], facial and bodily expression analysis [9]–[11] or eye movement [12] to capture characteristic behavioural changes relating to depression.

In this paper, we present our system design for the Depression Recognition Sub-Challenge (DSC) for the 2013 Audio/Visual Emotion Challenge and Workshop (AVEC 2013) [13]. This challenge requires participants to predict, using multimodal signal processing techniques, an individual's self-reported level of depression from a given multimedia file. Herein, we compare the accuracy of systems designed to capture relevant audio and facial information, which represent some of the key behavioural changes associated with depression [14]. We will also present fused systems designed to capture the complementary information capture in the individual modes. The aim of all testing described is twofold, firstly to outperform the challenge benchmark [13] and secondly to help us gain further insights into the design of affective sensing systems, with clinical utility, to aid objective diagnosis of depression.

2. BEHAVIOURAL MARKERS ASSOCIATED WITH DERPESSION

Although over 60% of human communication is believed to be non-verbal, large parts of this channel of communication are beyond conscious control, and current diagnostic methods for depression do not utilize this extra information [14].

2.1 Non Linguistic Audio Cues

Speech has long been recognized as a key component for any behavioural based depression recognition system. Speech in patients with depression is often described having diminished prosody, forcing it to sound dull, monotonous and "lifeless" [15]. This leads to depressed speech having longer pauses, decreased utterance length and a reduced speech rate [14], [15].

It has been hypothesized that it takes more articulatory effort for a depressed individual to produce and sustain speech; this is evident in decreased formant frequencies often reported with increasing levels of depression [4], [16]. This sustained effort causes a wide range of prosodic, articulatory and phonetic errors in depressed speech [17]–[20]. These effects, combined with changes to vocal tract properties potentially caused by increased vocal tract tension [16], [21], a lack of motor coordination [17], [20] or possibly the result of anti-depressant medications drying out the vocal tract [21], altering the spectral properties of the speech produced by a depressed individual [20].

Mel Frequency Cepstral Features (MFCC) are one of the strongest performing spectral features, when combined with Gaussian Mixture Models (GMM), for classifying either low/high levels of depression [22], [23], or the presence/absence of depression [7], [8]. Classification using MFCCs in combination with GMM-UBM (universal background model) supervectors has recently gained popularity for many paralinguistic tasks: many entrants to similar Interspeech Challenges on speaker affect, intoxication and sleepiness have used this style of system to obtain competitive results [24], [25].

Motivated by recent results showing a decrease in energy variability with increasing levels of depression, due in part to a decrease in the motor action associated with speech production [17], [20], and by results suggesting that this decrease in variability can be captured in a GMM [22], we explore a range of GMM-UBM supervector systems in combination with Support Vector Regression (SVR) for the task of predicting depression. Supervectors have been employed just once previously in a depression detection system [24], and have not been investigated for a regression problem.

2.2 Non Verbal Visual Cues

Facial expression cues are among the most popular visual cues that are utilized for behavioural analyses by both machines and humans. Whilst facial expression recognition in affect analysis and behaviour understanding is well researched, research in depression analysis via facial cues has been a more recent undertaking. Based on work done by Ellgring [26], in which depressed subjects showed significant decreases in facial activity, McIntyre *et al.* [27], proposed the use of person-dependent Active Appearance Models (AAM) [28], to compare the facial activity of depressed and healthy control subjects. At the same time, Cohn *et al.* [9] reported a 79% accuracy when combining AAM features with a Gaussian kernel Support Vector Machine for classifying the presence / absence of depression.

In addition to facial activity, the relative movements of body parts have also been found to be indicators of both affect [29], [30] and depression [31]. Spatio-temporal features [32] extracted from upper body expression and head movement cues and combined in a bag-of-words framework, have shown strong performance when identifying the presence/absence of depression [11]. It can be argued that these spatio-temporal descriptors capture very subtle yet discriminative movements exhibited by individuals suffering from depression [11], [31].

Based on these recent studies, this paper explores the use of facial cues for predicting levels of depression, specifically the suitability of both Space-Time Interest Points (STIP) [32] and Pyramid of Histogram of Gradients (PHOG) [33] in meeting our specified aims. The STIP concept has found much attention in computer vision and video analysis research. Our motivation for using this approach is its robustness to temporal misalignments within the spatio-temporal feature space [32]. Similarly, the motivation behind using PHOG is its superior performance in facial analysis tasks when using data collected in non-lab conditions [34].

2.3 Multimodal Prediction of Depression

Whilst multimodal affect recognition is a well-established field [35], to the best of the authors' knowledge, papers in this current challenge will represent some of the first attempts at depression recognition via fusion of audio and visual features. Both affect and depression recognition share common traits; a continuous negative affect is a key symptom of depression [36]. However, it should be stated that whilst affect is a more continually changing condition, depression is more steady-state in its nature, with individuals inflicted for weeks or months [24], [37]. A meta-analysis, conducted by D'Mello and Kory, into multimodal affect detection has shown that across 30 different published studies, all of which report both unimodal and multimodal affect detection results, on average multimodal systems offer an 8% relative improvement over unimodal systems [35]. Whilst visual and audio modalities share considerable redundancy in terms of affect detection [35], the improvement when fusing both is often attributed to results seen in the valence-arousal space. Recent studies have shown that audio cues are often better at recognizing arousal, whilst visual cues correlate better with valence [38]. Motivated by these results, we perform feature-level fusion to find a complementary set of audio and visual features for the task of predicting an individual's level of depression.

3. DEPRESSION CORPUS

The AVEC corpus is part of the Audio-Visual Depressive Language Corpus (AVDLC). Tasks completed by speakers include vocal exercises, free and read speech tasks. It is important to note that there is a large degree of phonetic variability captured within each file: not all files include all tasks. For an in-depth description of the corpus, the reader is referred to [13].

The clinical annotation of the corpus employs the Beck Depression Inventory (BDI), comprising 21 items with each item scored between 0-3, to give a total score range of 0-63. The BDI is the most widely used self-reported measure, originally developed in 1978 and updated in 1996 [2]. BDI has been shown to have clinical validity when differentiating between depressed and non-depressed individuals and when tracking changes in a patient's level of depression as they undergo treatment. Its major criticism is that it reflects the BDI authors' interest in the negative self-evaluative symptoms experienced in depression [2].

Given the steady-state nature of depression, the extent of depression in AVDLC files is likely not the strongest source of variability in all the behavioural signals available in this corpus [8], [11]. Therefore, the make-up of the corpus provides unique challenges in the DSC: the varying length of the files (Table 2),

the variety of speech tasks contained within each file, the skewed distribution of BDI scores in both the training and development data (Figure 1) – both favoring lower BDI scores, and the continually changing speaker affect [13]. These all introduce unwanted variability into the files making it harder to produce an accurate, meaningful regression score over larger file lengths.

Table 2. Summary of AVEC 2013 Depression Corpus

Partition	Number of Files	Max Length	Min Length	Average Length	Score Range
Train	50	27min 20s	8min 5s	14min 20s	0-44
Devel.	50	23min 55s	14min 20s	14min 20s	0-45
Test	50	23min 57s	5min 15s	15min 57s	N/A

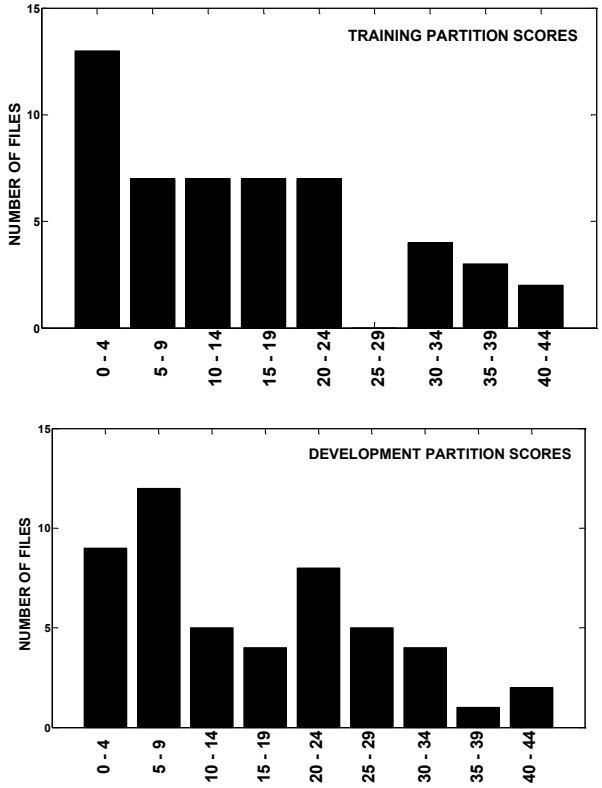

Figure 1. Distribution of BDI scores in both the training (top) and development (bottom) partitions

3.1 System Performance Metric

All system accuracies reported, unless otherwise stated, are in terms of Root Mean Square Error (RMSE). A baseline RMSE has been set by the challenge organisers: using the audio baseline feature set and an epsilon-SVR (ε-SVR), with a linear kernel. The audio baseline RMSE, for the development set, is 10.75. The visual baseline set using the baseline feature set and an ε-SVR with an intersection kernel is 10.72. For details of the baseline system and feature sets the reader is referred to [13].

3.2 Chance-Level System Accuracy

To establish the chance-level RMSE on the development partition, we ran four different tests (Table 3). The first was a repeated trial of a uniformly distributed random guesses between 0-45 where 45 represents the highest BDI score appearing in the development partition. The second was also repeated trials of uniformly distributed random guesses, this time generating numbers between 0-63, where 63 is the highest possible BDI score. The third and fourth tests used the mean (15.02) and median (12.5) scores of the training set respectively as an estimate for all development set scores. These scores are higher than the development set chance-level RMSE of 11.90 stated in the challenge baseline paper [13].

Table 3. Chance-level RMSE's for AVEC 2013 development partition

System	1	2	3	4
RMSE (st.dev.)	19.1 (0.01)	27.1 (0.015)	12.3	12.5

4. MULTIMODAL DEPRESSION RECOGNITION SYSTEM

4.1 Audio Subsystem

All speech systems were based on supervectors extracted using the GMM-UBM paradigm, using MFCCs appended with the first and second order time derivatives as the extracted features. A Gaussian Mixture Model (GMM) trained using the Expectation-Maximisation (EM) algorithm was employed as the UBM, which serves as a rough acoustic model.

In order to form a statistically rich audio description, a single supervector was formed per audio segment (file or subfile as shown in Fig. 2) by first adapting the UBM to fit the distribution of the features extracted from that segment via MAP-adaptation. The Gaussian mixture components were then stacked (as outlined in the following sections) to form the supervector. An SVR system was trained to operate on this supervector space. An advantage in using supervectors is the range of transforms that exist to help minimize variability in the acoustic supervector space, such as Nuisance Attribute Projection [39]; a similar methodology was used in depressed speech in [23] with promising results. Finally, the use of appropriate kernels in place of the standard inner product used in linear-SVR allows much more powerful non-linear support vector regression.

4.1.1 Kullback-Leibler Divergence Kernel

The use of the Kullback-Leibler (KL-means) divergence kernel allows a method of estimating the similarity between two utterances. It is possible to estimate this kernel by the means from a speaker-specific GMM, then transform these means by normalizing them by the corresponding non-adapted (UBM) covariance and weights. Using the transformed supervectors in combination with a linear SVR kernel is equivalent to using a KL-means kernel in the SVR model [39]. The KL transformation of the set of speaker-adapted means corresponding to the *i*-th mixture is given by:

$$\widetilde{\Phi}_i = \sqrt{\omega_i}(\sum_i^u)^{-\frac{1}{2}}\boldsymbol{m}_i^\lambda \quad (1)$$

where ω_i represents the weight component and $\sum_i^u$ the covariance matrix from the *i*-th UBM mixture and $\boldsymbol{m}_i^\lambda$ the *i*-th mean vector from λ-th speaker adapted GMM. The overall supervector for a given speaker (MAP adapted GMM) is formed by stacking all the speaker specific $\widetilde{\Phi}_i$'s:

$$\Phi_{KL}^\lambda = \left[\widetilde{\Phi}_1^T, \widetilde{\Phi}_2^T, \cdots \widetilde{\Phi}_M^T\right]^T \quad (2)$$

where *M* represents the total number of GMM mixtures.

4.1.2 GMM-UBM Mean Interval Kernel

A potential drawback of the KL divergence kernel is the lack of speaker-specific covariance information included in the supervector. Work in [22] shows the potential importance of including covariance information when forming an acoustic

model of depressed speech. The Bhattacharyya distance based GMM-UBM mean interval (GUMI) kernel allows us to include both covariance and weighting information in the kernel in a SVR kernel. It is possible to implement the GUMI SVR kernel by combining stacked and transformed speaker-specific GMM parameters with a linear SVR kernel [40]. The GUMI transformation of the i-th set of speaker adapted GMM parameters corresponding to the i-th mixture is given by:

$$\breve{\Phi}_i = \begin{bmatrix} \left(\frac{\sum_i^{\lambda}+\sum_i^{u}}{2}\right)^{-\frac{1}{2}}\left(\boldsymbol{m}_i^{\lambda}-\boldsymbol{m}_i^{u}\right) \\ diag\left(\left(\frac{\sum_i^{\lambda}+\sum_i^{u}}{2}\right)^{\frac{1}{2}}\left(\sum_i^{\lambda}\right)^{-\frac{1}{2}}\right) \\ \frac{\omega_i^{u}}{\omega_i^{\lambda}} \end{bmatrix} \quad (3)$$

where $\omega_i^u, \boldsymbol{m}_i^u$ and $\sum_i^u$ are the UBM's i-th weight, mean and covariance components, respectively, and $\omega_i^\lambda, \boldsymbol{m}_i^\lambda$ and $\sum_i^\lambda$ are the i-th weight, mean and covariance components, respectively, from λ-th speaker adapted GMM. As with the KL kernel, the overall supervector for a given speaker is formed by stacking all the speaker-specific $\breve{\Phi}_i$'s:

$$\Phi_{GUMI}^{\lambda} = \left[\breve{\Phi}_1{}^T \breve{\Phi}_2{}^T \cdots \breve{\Phi}_M{}^T\right]^T \quad (4)$$

4.1.3 UBM Weight Posterior Probability

The UBM weight posterior probability (UWPP) supervector is defined as the averaged posterior probability associated with each GMM-component [41]:

$$\Phi_{UWPP}^{\lambda} = \left[\sqrt{n_1}, \sqrt{n_2}, \cdots, \sqrt{n_i}, \cdots, \sqrt{n_M}\right]^T \quad (5)$$

where n_i is the posterior probability of the i-th mixture component, χ_i, given a set of training vectors $X = \{\boldsymbol{x}_1 \cdots \boldsymbol{x}_T\}$ for a given GMM mixture component:

$$n_i = \frac{1}{T}\sum_{t=1}^{T} P(\chi_i|\boldsymbol{x}_t) \quad (6)$$

$$P(\chi_i|\boldsymbol{x}_t) = \frac{\omega_i \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{m}_i, \sum_i)}{\sum_{j=1}^{M} \omega_j \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{m}_j, \sum_j)} \quad (7)$$

where $\mathcal{N}$ denotes the normal distribution. The UWPP uses the Bhattacharyya probability (BPP) kernel [41].

As with the KL and GUMI transformations, it is possible to apply a transformation outside the SVR and use a linear kernel, this operation is achieved through the square root operator in (5). As the posterior probability reflects the amount data assigned to a Gaussian component during training, the UWPP represents spectral variability on a global level as opposed to the localized variance captured in the covariance matrices.

4.1.4 Nuisance Attribute Projection

Nuisance Attribute Projection (NAP) removes subspaces that cause variability in the SVR kernel space [39]. NAP constructs a new kernel space through a projection matrix ***P,*** which projects the training and testing data into a more resilient subspace. NAP was originally designed as a way of mitigating effects due to different recording channels for speaker recognition.

After first extracting overlapping windows of voiced data from each file in the development partition, we then trained a NAP projection matrix. Viewing each of the different files as belonging to the same class and each of the extracted windows of data (subfiles) as representing a different 'channel' (Figure 2), we attempted to mitigate the effects of phonetic and affect variability captured in a file using NAP. First, we formed the matrix $\boldsymbol{A}$ from the supervectors formed from each extracted subfile:

$$\boldsymbol{A} = \left[\Phi_1^1, \cdots, \Phi_L^1, \Phi_{1,}^2 \cdots, \Phi_M^2, \cdots \cdots, \Phi_1^{50}, \cdots, \Phi_N^{50}\right] \quad (8)$$

where Φ_n^m is the supervector extracted from the n-th subfile of the m-th file in the development set. The projection matrix is formed by solving the eigenvalue problem:

$$\boldsymbol{A}(diag(\boldsymbol{W1}) - \boldsymbol{W})\boldsymbol{A}^t\boldsymbol{v} = \gamma\boldsymbol{v} \quad (9)$$

where $\mathbf{1}$ is a column matrix of all ones and $\boldsymbol{W}$ is the weighting matrix formed using:

$$W_{i,j} = \begin{cases} 1 & \text{if } \Phi_i \text{ and } \Phi_j \text{ are from same file} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

the effect of $(diag(\boldsymbol{W1}) - \boldsymbol{W})$ in Eq. (9) is to replace every matching vector with the extent of its deviation from the average vector in its file. The projection matrix $\boldsymbol{P}$ is formed by concatenating the eigenvectors corresponding to the k-th highest eigenvalues calculated in Eq. (9). The projection matrix will be used to project out unwanted variability across all supervectors used in the formation of **A**:

$$\hat{\Phi}_n^m = (\boldsymbol{I} - \boldsymbol{P}\boldsymbol{P}^t)\Phi_n^m \quad (11)$$

where $\boldsymbol{I}$ is the identity matrix.

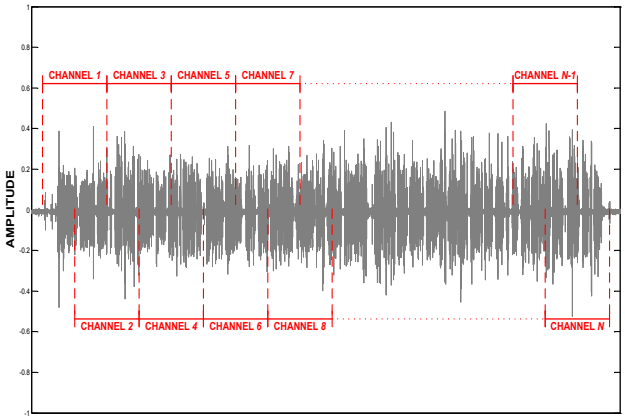

Figure 2. Example of creating different 'channels' from a single audio file for use with NAP compensation

4.1.5 Experimental Settings

The experimental settings (unless otherwise stated) of the classification system were as follows: a 39-dimensional feature vector was formed from thirteen MFCCs, including C_0, extracted using the openSMILE toolkit [42], every 10ms using a 25ms window. The 13 coefficients were concatenated with delta (Δ) and delta-delta (ΔΔ) coefficients extracted using the conventional regression equation, 39 features in total. Only voiced frames were used in the modeling, determined using openSMILE's voicing probability function. UBMs were formed with 10 iterations of the EM algorithm using MFCCs, Δ and ΔΔ features extracted from the entire training partition.

Where multiple scores were generated per file, using overlapping subfiles to extract multiple supervectors per file, the median operators were used to generate one regression score per file from all scores predicted per file. All features were normalized to a range of [0, 1] before SVR with the testing features being normalized by the range of the training features. All system testing and training was done using LIBSVM's built-in ε-SVR, in combination with a linear kernel, and default settings were used to avoid overfitting [43].

4.2 Visual System

All visual systems reported on here are based on the following four steps: (1) Extract STIPs on aligned faces for each entire video file. (2) Perform *K*-means clustering on STIP over time, allowing for variable durations. (3) Form time slices from the clusters. (4) Compute Bag-of-Words (BoW) where the time slices form the words. Similar to the audio system, an SVR is trained on these visual features to determine a subject's BDI score.

A limitation when using spatio-temporal visual descriptors is that they require robust alignment [13]. To preserve local temporal information, a reasonable approach is to obtain time slices (subfiles) prior to visual feature extraction by dividing the video file into time slices [44], [45]. A similar approach is followed in the systems explored here. However, determining the appropriate time slice duration is a non-trivial problem ([44], [45] used fixed duration) and, further, the complexity of the approach increases with an increasing duration of the overall video files. To over-come these alignment challenges, STIPs [32] are computed for creating variable duration time slices, which are created on the basis of the presence of facial activity.

4.2.1 Space-Time Interest Points

STIPs detect salient points in a video by extending the idea of the Harris interest point detector to local structures in the spatio-temporal domain. Salient points are detected where image values have sufficient local variation in both the space and time domain. The STIPs reflect the spatio-temporal changes, which account for movements inside the facial area and elsewhere; such as hands, shoulders and head movements. Finally, *K*-means clustering is performed on the STIPs, which lie within the face blob area using temporal separation as the distance metric. The clusters then determine the time slice, i.e. each time slice extends from the earliest STIP in a cluster to the final STIP in that cluster.

4.2.2 Time Slice based Bag-of-Words

For each word (time slice), a Pyramid of Histogram of Gradients (PHOG) [33] descriptor is computed. PHOG is an extension of the popular Histogram of Gradient descriptor, where an image is divided into blocks on various pyramid levels and orientation histograms are computed based on orientations. Orientations are fused at block and pyramid level into histograms. This approach represents an image by both its local shape, the individual histograms calculated per block, and its spatial layout, the result of multiple resolution tiling [33]. The motivation behind using PHOG is based on its superior performance for face analysis as compared to local binary patterns and its variants, especially when the data is recorded in non-lab conditions, which introduces errors in alignment [34]. PHOG features have also been used successfully in automatic emotion recognition [34]. One method to compute PHOG for a time slice is to evaluate all its frames and then to perform an averaging or max operation. Averaging can dampen the signal and max can be biased towards inaccuracy in alignment. Therefore, we use a simpler approach where the central frame of each time slice is selected for computation of the PHOG descriptors, which the results suggest performs better than the organisers' baseline (see Section 6.1.2).

BoW represents a sample as an unordered frequency of words and has been very popular in computer vision for its performance advantage of handling the change in appearance of an object under consideration easily. A BoW dictionary is computed using the PHOG descriptors. The dictionary is learnt by considering each time slice a word and each video file a document. This approach allows the vision system to cater for varying file lengths.

From the STIPs, one can also compute histograms of gradient and flow around the interest point. Following [45], we compute histograms and we refer to this method as STIP_BoW, which we use as a visual baseline with which to compare the performance of the PHOG based systems in the experiment section below. Similar features have been used for the classification of the presence / absence of depression using visual features [11], [31], [45].

4.2.3 Experimental Settings

To perform facial dynamics analysis, first a face tracker [46] was employed to each video to register the face. This gave 66 face landmark points, which were further used for face alignment. STIP and PHOG feature extraction was then performed on these aligned faces. The STIP features were extracted using the methodology presented in [45], with the number of clusters empirically set to 1000. This results in 1000 time slices extracted per video clip, the range of times for these slice, across the entire corpus, is 0.04s to 131.2s with an average time slice time of 0.49s.

The default PHOG implementations [34] were used here, in which the pyramid size was set to 3, the orientation range was [0-360] and the number of bins of the histogram was set to 8. The aligned faces were then rescaled to 96×96 pixels and the BoW computed. Deciding the dictionary size is non-trivial and has an effect on the performance of the system. Similar to the method used in [11], a BoW was computed, the dictionary was learnt using approximate nearest neighbour and the dictionary size was decided empirically (a range of 100-300 was tested). For our STIP_BoW baseline, the dictionary size was empirically set 2000.

System testing and training was done using LIBSVM's built-in ε-SVR in combination with either the histogram intersection (HIK) or Chi-Squared (Chi^2) kernel due to their proven suitability with the chosen visual features [47], with both kernels implemented using the Maji *et al.* add-on package [48]. Again, to avoid overfitting, default hyper parameters were used with both kernels.

4.3 Fused Systems

Common fusion approaches in affect detection include both naïve feature based fusion and modal level fusion [35]. Given the different extraction methods used when calculating the audio and visual features it is highly unlikely that the important cues in each feature group match up in temporally, therefore we will be fusing together file length feature representations. As the challenge requires prediction of an individual depression score, which lends itself to regression analysis, we performed feature level fusion on different combinations of features described in Sections 4.1 and 4.2. We also compared accuracy of the non-linear Chi^2 and HIK SVR kernels described in Section 4.2.

4.3.1 Experimental Settings

All features used in fusion were extracted in the methods described in Sections 4.1.5 and 4.2.3. All features were normalized to a range of [0, 1] before SVR, with the testing features being normalized by the range of the training features. Again LIBSVM's built-in ε-SVR was employed with both kernels implemented using the Maji *et al.* add-on package. Default kernel settings were used to avoid overfitting of hyper parameters.

4.4 System Development Settings

To help minimize our chances of overfitting, two separate cross-fold validation systems were used for system development. The first was employed for all audio development set tests. The UBM was formed from the training partition; supervectors were then extracted from the development partition. 20 trials of 5-fold cross-validation were then employed, where the development set was randomly split into 5 folds, which were used in all different

training and test permutations in each trial (Figure 3). This experimental scheme was used to generate the development results for the audio systems.

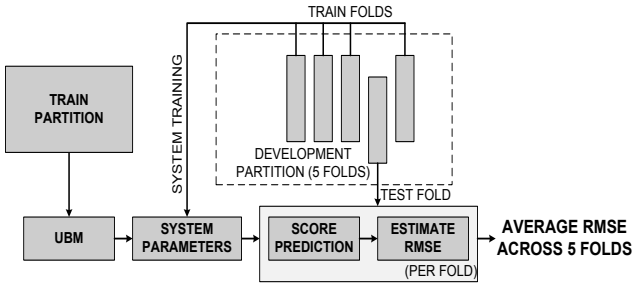

Figure 3. Regression scheme used to generate all system development audio scores

The experimental scheme for visual and fused scores uses both training and development data to train the SVR. Again 20 trials of 5-fold cross-validation were used. Within each trial, a different set of development files was randomly split into different training and test permutations for each trial, only this time the development vectors held out for training were combined with the training partition vectors to form the SVR model (Figure 4). This system was used for all visual and fusion system development results.

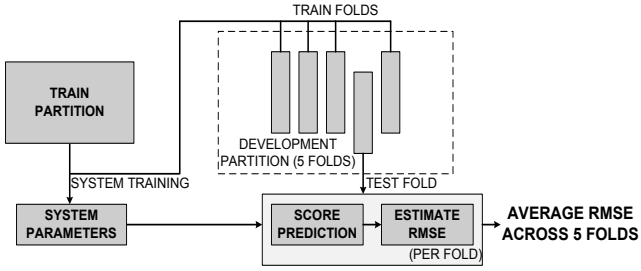

Figure 4. Regression scheme used to generate all system development visual and fusion scores

For both validation systems, mean RMSE was recorded for each trial. Overall scores for system development were then reported in terms of the mean, minimum and maximum RMSE, and standard deviation of all trials. To meet the challenge requirement of predicting a single depression score per file, all testing was done using LIBSVM's built-in ε-SVR. Default settings were used to avoid overfitting of hyper parameters.

5. SYSTEM DEVELOPMENT

5.1 Audio Sub System

5.1.1 Initial System Design

An initial series of comprehensive tests were run on 3 supervector systems in which the number of Gaussian mixtures (powers of 2 between 8 and 1024) and number of MAP adaptations (2, 5, 10, 20) were varied. All tests were done both with and without C_0, and corresponding deltas appended to the feature set. The results confirmed that on average, across the three supervectors, a feature space of 39 dimensions, twelve MFCC coefficients appended with C_0, and corresponding deltas and a supervector space of 128 mixtures adapted from the UBM using 5 iterations of MAP gave the best results (Table 4).

For the multi subfile systems, we then ran an exhaustive series of tests to determine the optimal parameters in terms of subfile length (measured in seconds) and subfile overlap. Results from this analysis indicated that changing the window size between 30 and 60s and varying the amount of overlap between 5 to 30s does not have a large effect on system RMSE. However, increasing the VP (voicing probability) setting beyond 0.55 resulted in a significant increase in RMSE. On average, the best results across the 3 supervector systems, KL-means, GUMI and UWPP were found using a window size of forty seconds overlapped by 20s (Table 4). Herein, this is the default audio multi subfile setting.

Table 4. A selection of RMSEs generated using AVEC 2013 Development Partition from tests run to determine suitable parameters for the audio systems.

System	mean	min	max	st.dev.
KL-mean	10.07	9.40	11.17	0.40
GUMI	10.13	9.34	11.35	0.46
UWPP	12.21	11.06	13.39	0.62
KL-mean (mult.)	10.35	9.47	11.66	0.52
GUMI (mult.)	10.60	9.93	11.63	0.48
UWPP (mult.)	12.22	11.34	13.03	0.50

Whilst both the KL-mean and GUMI systems give below chance-level RMSE (according to [13]), results of the UWPP system were disappointing and could be a reflection of the acoustic variability seen within the corpus. We speculate that if the corpus had matched utterances, UWPP performance might improve.

Whilst better results, not shown, could be found for each supervector, to fine-tune the number of mixes, MAP iterations, subfile length and overlap could result in overfitting to the development partition.

To test the effect of variability captured within a file on score prediction, we extracted a number of supervectors per development set file, using the overlapping methodology described in Section 3.1, and generated a range of prediction scores per file (Figure 3). Results shown in Figure 4 were found using KL-means with 30s windows overlapping every 5s and leave-one-out (file) cross-validation.

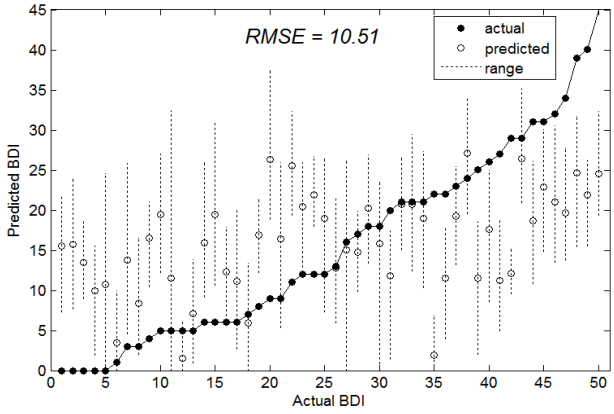

Figure 4. Range of scores predicted per development set file using KL-means supervector, 30s windows overlapping every 5s and leave-one-out (file) cross-validation

5.1.2 NAP Results

The dimensionality of the projection matrix is the critical parameter in ensuring how much variability is removed from each supervector. If too much variability is removed then useful depression information contained in the speech signal may also be removed. To ascertain how much variability is captured per dimension, the eigenvalues extracted in Eq. (9) were sorted in descending order (these can be considered estimates of variability along the directions defined by the corresponding eigenvectors). The percentage of variability captured by a given number of dimensions can then be estimated from a plot of the ratio of

eigenvalues to the sum of all eigenvalues. For the KL means and GUMI systems, almost 100% of the variability is captured by the first 500 eigenvectors whilst for the UWPP supervector the full dimensionality is needed, 128, to capture all the variability (Figure 5).

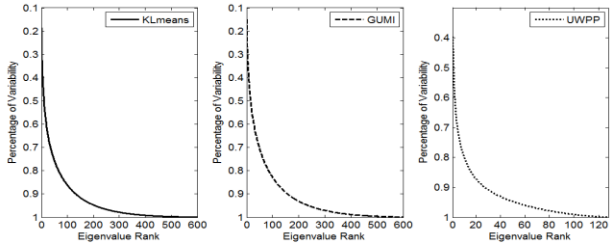

Figure 5. Variability captured by increasing dimensionality of NAP projection matrix for the KL-mean (left), GUMI (middle) and UWPP (right).

The results of applying Eq. (11) are shown in Figure 6. Again, to minimize the effects of overfitting, no attempts were made to fine tune the dimensionality of the projection matrix. Both the GUMI and UWPP systems were able to lower their RMSEs when compared with the results seen in Table 5, whilst the KL-means system was unable to. None of the systems beat the RMSE set in Table 4. We speculate that this is due to the uncontrolled nature calculating the NAP variability - we have no guarantees that the variability being removed is useful or detrimental - as well as the (relatively) small amount of data being used to compute ***P***. A less naïve partition of the data set, such as turn based, might also be able to help improve results in both sections 5.1.1 and here.

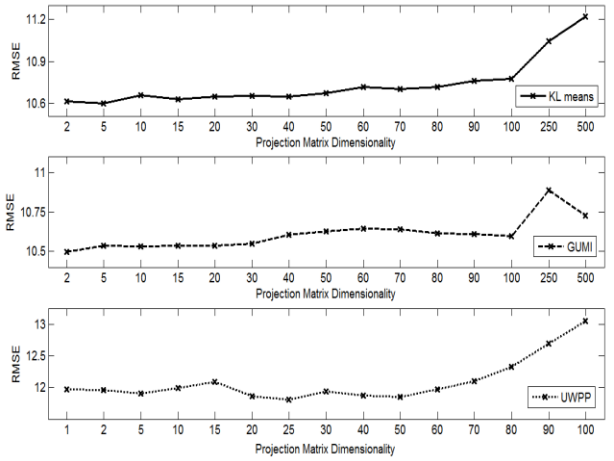

Figure 6. Effect on RMSE, for overlapping window technique, of increasing dimensionality of NAP projection matrix for the KL-mean (top), GUMI (middle) and UWPP (bottom)

5.2 Visual Sub System

A series of initial tests was run to help determine the dictionary length for the PHOG feature, (Table 6). All results were generated, using the second testing methodology from Section 4.4. There are small performance gains to be found when increasing the PHOG dictionary length from 100 to 200, but there is a saturation effect when increasing this length above 200. We speculate this effect can be attributed to the curse of dimensionality. The PHOG appears better suited to the challenge's regression task than the STIP_BoW feature. The reason behind this is that the STIP_BoW descriptors are computed around interest points only; therefore may miss potentially important global information. It is noticeable that the Chi^2 kernel performs best for histogram based visual features. This re-affirms earlier results seen for similar features used in human action recognition [32].

Table 6. RMSE for Initial Visual System Design

System	kernel	mean	min	max	st.dev.
STIP_BoW	Chi^2	13.21	12.70	13.67	0.27
	HIK	12.18	11.73	12.56	0.23
PHOG_100	Chi^2	8.18	7.72	8.73	0.28
	HIK	8.32	7.76	8.82	0.28
PHOG_200	Chi^2	7.27	6.97	7.69	0.18
	HIK	8.05	7.35	8.39	0.25
PHOG_ 300	Chi^2	9.16	8.76	9.70	0.21
	HIK	9.16	8.71	9.85	0.25

5.3 Fused Systems

To accurately allow for comparison between our unimodal and multimodal results, we firstly generated a set of results on a subset of individual systems using the experimental settings described in Section 4.3.1 and the second experimental scheme from Section 4.4 (Table 7). All results in this section are reported in terms of mean RMSE only. Surprisingly, the HIK kernel worked well with the KL-mean audio feature; this result could be due in part to the increased training data in this regression system (90 vectors vs. 40). We speculate that as the KL-means works well with the HIK kernel, it could potentially have properties, which resemble Bag-of-visual words (BOV) based features, as in [47]. For the visual features, we see an improved system performance when using the non-linear SVR kernels; this matches well with results presented in [47] for the comparison of different kernel methods for evaluating histogram based visual features. The differences in visual results between those presented in Table 6 and those in this section can be attributed to the normalisation of the feature space (Section 4.3.1); as the visual features are histograms they typically do not need further normalisation.

Table 7. RMSE found for features from different modes using the same regression system.

System	Mode	Kernel		
		Lin	Chi^2	HIK
STIP_BoW	V	12.23	11.67	11.81
PHOG_100	V	12.01	11.72	11.81
PHOG_200	V	12.15	11.84	11.84
PHOG_300	V	12.12	11.85	11.89
KL-means	A	11.36	11.40	9.88

When fusing the KL-means with visual features, small improvements can be found (Table 8). The results for the fusion with HIK kernel are well below chance level RMSE, whereas only the KL-means feature, with HIK kernel, is significantly lower than chance level in Table 7. This result matches well with experimental work presented in [45], where increases in system performance for a presence/absences of depression classifier are found using feature level fusion.

Table 8. RMSE found fusing KL-means with a range of different visual features

System	Kernel		
	Lin	Chi^2	HIK
KL-Means + STIP_BoW	11.74	11.67	10.30
KL-Means + PHOG_100	11.39	11.38	9.75
KL-Means + PHOG_200	11.29	11.32	9.67
KL-Means +PHOG_300	11.36	11.44	9.84

6. AVEC 2013 CHALLENGE RESULTS

6.1 Development Set Results

To make results comparable with baseline system accuracy, a selection of the systems presented in Section 5 was trained using the training partition and tested using the development partition (Figure 7).

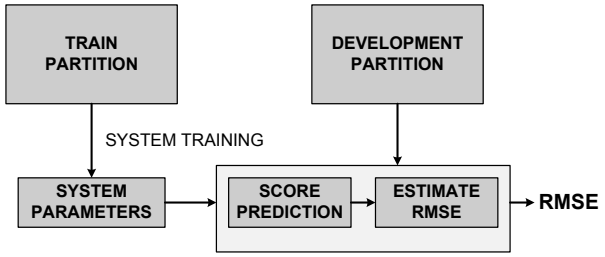

Figure 7. Regression system experimental scheme to generate RMSE scores comparable with AVEC 2013 baseline

To generate the audio results default experimental settings were used, whilst for the visual and fusion results results Chi^2 and HIK kernels were used, respectively in the SVR. All features were normalized to lie in the range [0, 1] before SVR training and testing, except for a subset of the non-fused visual features.

6.1.1 Audio Systems

All KL-means and GUMI audio systems outperformed the AVEC development baseline of 10.75 (Table 9). Whilst the single UWPP result is not surprising given results in Section 5, the performance of the two UWPP (multiple subfile) systems is encouraging. Both these systems almost match the challenge baseline on a feature whose dimensionality is 128, compared with 95256 features supplied in the challenge feature set [13].

Table 9. RMSE generated using Audio (A) feature systems trained using the AVEC 2013 Training Partition and tested on Development Partition.

System	Mode	RMSE		
		Single	Multi.	Multi. (NAP)
KL-means (single)	A	9.60	9.00	8.94 (***P*** dim. 5)
GUMI (single)	A	9.56	9.39	9.39 (***P*** dim. 5)
UWPP (single)	A	12.01	10.81	10.77 (***P*** dim. 2)

6.1.2 Visual Systems

All variants of the non-normalized PHOG features outperformed the challenge visual baseline of RMSE of 10.72 (Table 10). The disappointing performance of the STIP_BoW helps support our earlier speculation that these features, computed around interest points only, may omit potentially important global cues, in terms of depression recognition.

Table 10. RMSE generated using Visual (V) feature systems trained using the AVEC 2013 Training Partition and tested on Development Partition.

System	Mode	RMSE (No Norm.)	RMSE
STIP_BoW	V	12.58	11.90
PHOG_100	V	8.64	12.07
PHOG_200	V	8.37	12.08
PHOG_300	V	8.84	11.98

6.1.3 Fused Features

Whilst the fusion results were unable to match the non-fused RMSE of KL-means, both systems were able to outperform both the audio and visual challenge baselines, with feature dimensionality well below that used to set the challenge baselines. These results show the potential that combining modalities has in predicting an individual's self-reported level of depression.

Table 11. RMSE generated using both audio and visual features system trained using the AVEC 2013 Training Partition and tested on Development Partition.

System	Mode	RMSE
KL-Means + STIP_BoW	A+V	10.65
KL-Means + PHOG_200	A+V	10.44

6.2 Test Set Results

We submitted two audio systems, two visual systems and a fusion system as our official challenge entry (Table 12). For the audio systems and fusion system, the training and development set were used to train the SVR, again the features were normalized to lie within [0, 1] and the ranges of the training and development set were used to normalize the testing features. No feature level normalisation was used to generate our visual predictions.

The first system tested (Table 12) was KL-means, representing the most consistent performing audio feature in system development. This feature obtained a very competitive RMSE of 10.17, well below the Challenge test audio baseline of 14.12. Given the consistent performance of this feature in system development and in other paralinguistic tasks [24], [25], this result is not surprising. The second audio system chosen was another KL-mean system. This time the multi system including NAP, the training and development set was used to calculate the projection matrix, with a dimensionality of 5, and this was then applied to the test set. The same feature normalisation method was applied to the training/development set and test set. The RMSE for this system was 13.34, still beating the challenge baseline. We speculate that the increase in RMSE could be due to an incorrectly chosen **P** dimensionality due to combining the training and development data, as well as the controlled nature of NAP (see Section 5.1.3).

Two versions of the most consistent performing visual system, PHOG_200 in combination with a Chi^2 SVR kernel, were submitted for test evaluations. To generate our predictions for the first visual system, the SVR was trained using both the challenge's training and development sets, and using this set-up we just beat the Challenge's visual baseline of 13.61. Given the strong performance of this feature in system development, this result was disappointing. Therefore, we re-ran our system but this time only used the training partition for SVR training, and we were able to significantly lower our RMSE. We argue that in certain cases more training data does not always guarantee better system performance [49].

To generate our fusion result, we combined the two strongest performing visual and audio features in development: KL-mean and PHOG_200, in combination with a HIK SVR. As with the development set results for this feature combination we outperformed the baseline RMSE of both modalities.

Table 12. AVEC 2013 Test Set Results

System	Mode	RMSE
KL-means (single)	A	10.17
KL-means (multi + NAP **P** dim. =5)	A	13.34
PHOG_200 (Train + Devel.)	V	13.51
PHOG_200 (Train Only)	V	10.45
KL-mean+PHOG_200	A+V	10.62

7. CONCLUSION

Learning from previous challenges, overfitting to development sets can be a major confounding factor [50]. As a result, various attempts, such as random trials of cross fold validation and no system fine tuning were made during system development. When comparing our scores to those presented in the challenge baseline paper [13], we feel we met our stated aim of outperforming the challenge benchmark. Secondly, given the strong performance of our features in both system development and testing, our attempts minimize overfitting were successful.

In terms of audio features, whilst the NAP approach to minimizing speaker variability did not perform as strongly in the test conditions, the development results show promise and future work will be spent developing this approach further. Interestingly, results seem to show that correctly choosing the correct number of Gaussian Mixtures and MAP iterations to suit the task at hand goes a long way towards minimising unwanted variability. For the visual features, dividing the video clip into variable duration time slices preserves local temporal change information and the PHOG feature is well suited to the task of depression detection and future work will be done developing systems around this feature.

A key result in this paper was the promising result shown for the feature level fusion results in both system development and test conditions. Given the wide range of symptoms associated with depression, and recent affect recognition studies showing the benefits of multimodal approaches [35], [38], we feel that a multimodal approach represents the best prospect for building a successful depression classification system. Therefore, future research will focus on improving our fusion strategies.

8. ACKNOWLEDGMENTS

This research was funded in part by ARC Discovery Project DP130101094. The authors would like to thank Dr. Jia Min Karen Kua for her speech processing advice and code and Sharifa Alghowinem for her insights and advice.

National ICT Australia is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

9. REFERENCES

[1] C. D. Mathers and D. Loncar, “Projections of Global Mortality and Burden of Disease from 2002 to 2030,” *PLoS Med*, vol. 3, no. 11, pp. 2011–2030, 2006.

[2] D. Maust, M. Cristancho, L. Gray, S. Rushing, C. Tjoa, and M. E. Thase, “Chapter 13 - Psychiatric rating scales,” in *Handbook of Clinical Neurology*, vol. Volume 106, F. B. Michael J. Aminoff and F. S. Dick, Eds. Elsevier, 2012, pp. 227–237.

[3] A. T. Beck and B. A. Alford, *Depression: Causes and Treatment*. University of Pennsylvania Press, 2008.

[4] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. S. Geralts, “Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology,” *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[5] H. D. Schmidt, R. C. Shelton, and R. S. Duman, “Functional Biomarkers of Depression: Diagnosis, Treatment, and Pathophysiology,” *Neuropsychopharmacology*, vol. 36, no. 12, pp. 2375–2394, 2011.

[6] E. Moore, 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 Trans. on Biom. Eng.*, vol. 55, no. 1, pp. 96–107, 2008.

[7] L. S. A. Low, N. C. Maddage, M. Lech, L. Sheeber, and N. Allen, “Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents,” in *2010 IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 5154–5157.

[8] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, “An Investigation of Depressed Speech Detection: Features and Normalization,” in *Interspeech2011*, 2011, pp. 2997–3000.

[9] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Ying, N. Minh Hoai, M. T. Padilla, Z. Feng, and F. De la Torre, “Detecting depression from facial actions and vocal prosody,” in *3rd Int. Conf. on Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009.*, 2009, pp. 1–7.

[10] S. Scherer, G. Stratou, M. Mahmoud, and J. Boberg, “Automatic Behavior Descriptors for Psychological Disorder Analysis,” *IEEE Conf. on Automatic Face and Gesture Recognition 2013*, p. NA, 2013.

[11] J. Joshi, R. Goecke, M. Breakspear, and G. Parker, “Can body expressions contribute to automatic depression analysis,” *Proceedings of the 10th IEEE Int. Conf. on Automatic Face and Gesture Recognition FG2013. Shanghai, China*, 2013.

[12] S. Alghowinem, R. Goecke, M. Wagner, G. Parker, and M. Breakspear, “Eye Movement Analysis for Depression Detection,” in *2013 IEEE Int. Conf. on Image Processing ICIP2013, Melbourne, Australia, 15-18 Sep 2013*, 2013.

[13] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, “The Continuous Audio/Visual Emotion and Depression Recognition Challenge,” in *The 21st ACM Int. Conf. on Multimedia*, 2013, p. NA.

[14] M. J. H. Balsters, E. J. Krahmer, M. G. J. Swerts, and A. J. J. M. Vingerhoets, “Verbal and Nonverbal Correlates for Depression: A Review,” *Current Psychiatry Reviews*, vol. 8, no. 3, pp. 227–234, 2012.

[15] C. Sobin and H. A. Sackeim, “Psychomotor symptoms of depression,” *Am J Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

[16] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, “Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression,” *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 309–319, 1993.

[17] T. F. Quatieri and N. Malyska, “Vocal-Source Biomarkers for Depression: A Link to Psychomotor Activity,” in *Interspeech2012*, 2012, pp. 1059–1062.

[18] J. C. Mundt, A. P. Vogel, D. E. Feltner, and W. R. Lenderking, “Vocal Acoustic Biomarkers of Depression Severity and Treatment Response,” *Biological Psychiatry*, vol. 72, pp. 580 –587, 2012.

[19] A. Trevino, T. Quatieri, and N. Malyska, “Phonologically-based biomarkers for major depressive disorder,” *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–18, 2011.

[20] N. Cummins, J. Epps, and E. Ambikairajah, “Spectro-Temporal Analysis of Speech Affected by Depression and Psychomotor Retardation,” in *2013 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, p. NA.

[21] 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 Trans. on Bio-Eng,*, vol. 47, no. 7, pp. 829–837, 2000.

[22] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling Spectral Variability for the Classification of Depressed Speech," in *14th Annual Conference of the International Speech Communication Association Interspeech2013, Lyon, France, 25-29 Aug 2013*, 2013.

[23] D. Sturim, P. A. Torres-Carrasquillo, T. F. Quatieri, N. Malyska, and A. McCree, "Automatic Detection of Depression in Speech Using Gaussian Mixture Modeling with Factor Analysis," in *Interspeech2011*, 2011, pp. 2983–2986.

[24] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language; State-of-the-art and the challenge," *Computer Speech & Language*, p. NA, 2012.

[25] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, and F. Eyben, "Medium-term speaker states - A review on intoxication, sleepiness and the first challenge," *Computer Speech & Language*, p. NA, 2012.

[26] H. Ellgring, *Non-verbal communication in depression*. Cambridge University Press, 1989.

[27] G. McIntyre, R. Goecke, M. Hyett, M. Green, and M. Breakspear, "An approach for automatically measuring facial activity in depressed subjects," in *3rd Int. Conf. on Affective Computing and Intelligent Interaction and Workshops, 2009. (ACII '09)*, 2009, pp. 1–8.

[28] J. Saragih and R. Goecke, "Learning AAM fitting through simulation," *Pattern Recognition*, vol. 42, no. 11, pp. 2628–2636, Nov. 2009.

[29] H. Gunes and M. Piccardi, "Bi-modal emotion recognition from expressive face and body gestures," *Journal of Network and Computer Applications*, vol. 30, no. 4, pp. 1334–1345, Nov. 2007.

[30] A. Kleinsmith and N. Bianchi-Berthouze, "Affective Body Expression Perception and Recognition: A Survey," in *IEEE Transactions on Affective Computing,*, 2013, vol. 4, no. 1, pp. 15–33.

[31] J. Joshi, A. Dhall, R. Goecke, and J. F. Cohn, "Relative Body Parts Movement for Automatic Depression Analysis," in *Proceedings of Affective Computing and Intelligent Interaction ACII2013, Geneva, Switzerland, 2-5 Sep 2013*, 2013.

[32] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conf. on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, 2008, pp. 1–8.

[33] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM Int. Conf. on Image and video retrieval (CIVR '07)*, 2007, pp. 401–408.

[34] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using PHOG and LPQ features," in *2011 IEEE Int. Conf. on Automatic Face & Gesture Recognition and Workshops (FG 2011),*, 2011, pp. 878–883.

[35] S. D'Mello and J. Kory, "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *14th ACM Int. Conf. on Multimodal Interaction (ICMI '12)*, 2012, pp. 31–38.

[36] R. J. Davidson, D. Pizzagalli, J. B. Nitschke, and K. Putnam, "Depression: Perspectives from Affective Neuroscience," *Annual Review of Psychology*, vol. 53, no. 1, pp. 545–574, 2002.

[37] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 32–80, 2001.

[38] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space," *IEEE Trans. on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[39] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *2006 IEEE Int. Conf. on Acoustics, Speech and Signal Processing, (ICASSP' 06).*, 2006, vol. 1, pp. 97–100.

[40] C. H. You, K. A. Lee, and H. Li, "GMM-SVM Kernel With a Bhattacharyya-Based Distance for Speaker Recognition," *IEEE Trans.on Audio, Speech, and Lang. Proc.*, vol. 18, no. 6, pp. 1300–1312, Aug. 2010.

[41] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language*, vol. 27, no. 1, pp. 151–167, Jan. 2013.

[42] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the Int.Conf. on Multimedia*, 2010, pp. 1459–1462.

[43] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. on Intelligent Systems and Technology*, vol. 2, no. 3, p. 1:27, 2011.

[44] K. Sikka, A. Dhall, and M. Bartlett, "Weakly Supervised Pain Localization using Multiple Instance Learning," in *Automatic Face and Gesture Recognition, FG 2013 IEEE Conference*, 2013, p. NA.

[45] J. Joshi, R. Goecke, S. Alghowinem, A. Dhall, M. Wagner, J. Epps, G. Parker, and M. Breakspear, "Multimodal Assistive Technologies for Depression Diagnosis and Monitoring," *Journal on Multimodal User Interfaces*, 2013.

[46] X. Xiong and F. De la Torre, "Supervised Descent Method and its Applications to Face Alignment," in *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2013, p. NA.

[47] J. Wu and J. M. Rehg, "Beyond the Euclidean distance: Creating effective visual codebooks using the Histogram Intersection Kernel," in *2009 IEEE 12th Int.Conf. on Computer Vision*, 2009, pp. 630–637.

[48] S. Maji, A. C. Berg, and J. Malik, "Classification using intersection kernel support vector machines is efficient," in *IEEE Conf.on Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, 2008, pp. 1–8.

[49] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes, "Do We Need More Training Data or Better Models for Object Detection?," in *BMVC*, 2012, pp. 1–11.

[50] N. Cummins, J. Epps, J. Min, K. Kua, and J. M. K. Kua, "A Comparison of Classification Paradigms for Speaker Likeability Determination," in *Interspeech2012*, 2012, pp. 282–285.