ORIGINAL RESEARCH

Coupled HMM-based multimodal fusion for mood disorder detection through elicited audio–visual signals

Tsung-Hsien Yang[1] · Chung-Hsien Wu[1] · Kun-Yi Huang[1] · Ming-Hsiang Su[1]

Received: 31 March 2016 / Accepted: 11 July 2016 / Published online: 19 July 2016

Abstract Mood disorders encompass a wide array of mood issues, including unipolar depression (UD) and bipolar disorder (BD). In diagnostic evaluation on the outpatients with mood disorder, a high percentage of BD patients are initially misdiagnosed as having UD. It is crucial to establish an accurate distinction between BD and UD to make a correct and early diagnosis, leading to improvements in treatment and course of illness. In this study, eliciting emotional videos are firstly used to elicit the patients' emotions. After watching each video clips, their facial expressions and speech responses are collected when they are interviewing with a clinician. In mood disorder detection, the facial action unit (AU) profiles and speech emotion profiles (EPs) are obtained, respectively, by using the support vector machines (SVMs) which are built via facial features and speech features adapted from two selected databases using a denoising autoencoder-based method. Finally, a Coupled Hidden Markov Model (CHMM)-based fusion method is proposed to characterize the temporal information. The CHMM is modified to fuse the AUs and the EPs with respect to six emotional videos. Experimental results show the promising advantage and efficacy of the CHMM-based fusion approach for mood disorder detection.

Keywords Mood disorder detection · Coupled Hidden Markov Model · Multimodal fusion · Autoencoder adaptation

1 Introduction

Nowadays, unipolar depression (UD) and bipolar disorders (BD) are common mental illness. According to the Diagnostic and Statistical Manual of Mental Disorders-Fifth Edition (DSM-5) (American Psychiatric Association 2013), the symptoms of BD could change between the period of mania and depression intensely. UD, also known as Major Depressive Disorder, is a kind of mental illness characterized by a persistent depression. Generally, BD and UD have similar depressive appearances but specific pathophysiology. Current diagnoses are determined mainly according to structured clinical assessment based on DSM-5 and patient's self-report of family medical history. Several recent psychopathological studies have found that a high percentage of bipolar disorder patients are initially misdiagnosed as having unipolar depression. This is because the extremes in intensity and duration of mood disorders are just below the DSM-5 diagnostic threshold for a diagnosis of hypomania. This means that a large portion of BD patients are misdiagnosed as UD. According to (Perlis 2005), around 40–67 % of BD patients were likely to be misdiagnosed as UD because the patients with BD seek treatment more often when they are in depression state on initial presentation. In addition, this misdiagnosis carries significant negative consequences for the treatment of the BD patients. Therefore, it is crucial to distinguish

✉ Chung-Hsien Wu
chunghsienwu@gmail.com

Tsung-Hsien Yang
thyang@mail.ncku.edu.tw

Kun-Yi Huang
iamkyh77@gmail.com

Ming-Hsiang Su
huntfox.su@gmail.com

[1] Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan

between BD and UD in order to make an accurate and early diagnosis, leading to improvements in treatment and course of illness.

In previous studies, UD has attracted many research efforts in different fields. Cohn et al. (2009) conducted a clinical interview of depressed patients and found that automatic audio signal processing of vocal prosody and facial image analysis effectively detected depression and non-depression by using SVM classifier. Sanchez et al. (2011) used prosodic and spectral features in SVM to detect depression in elderly males and found that both prosodic and spectral speech features are good cues for characterizing speaker's emotional states and could be applied to detect depressed speakers. Low et al. (2010, 2011) investigated the properties of acoustic speech for depression in adolescents based on Gaussian mixture models (GMM) and SVM. They found the glottal flow formation is useful cues for clinical depression. However, these research results did not consider the effect of multimodal fusion. Ooi et al. (2013) used multichannel weighted decision procedure to predict depression in adolescents. They also indicated that the glottal and prosodic features were effective in predicting depression (Sanchez et al. 2011). The results show that these approaches can be used to distinguish depression from control group in short-term detection. However, few researchers investigated detecting the distinction between BD and UD due to the difficulty for detection of BD in depression period and the lack of database. Recently, in short-term BD detection, Erguzel et al. (2015) combined particle swarm optimization and neural network methods to discriminate UD and BD via quantitative electroencephalography (QEEG) data. International Affective Picture System (IAPS)-pictures-slideshow intends to provide emotional stimuli of patients, and analyzes their electrodermal response to identify the state of the bipolar patients (Greco et al. 2014; Lanata et al. 2014). However, those methods need some special equipment to collect data. Other research tried to identify the state of bipolar patients (Bersani et al. 2013) using video to elicit the emotional state of the patients and collected their facial expressions. Besides facial expression-based approaches, in (Howard 2013), speech was transformed into lexical information in real-time and used a pre-constructed model by Alzheimer's database to determine axiological values and time orientation of lexical features. Both facial expressions and speech are the natural and easy ways to get the states of BD patients. In addition, mood disorder is highly relevant to emotions, and emotion recognition from speech and facial expression has been a hot topic in recent years (Wu et al. 2014). For the extension of emotion recognition, recognition of mood disorder based on speech and facial expression plays an important role in helping clinical diagnosis. Many studies indicated that patients with BD have lower accuracy of emotion perception than the others (David et al. 2014; Goghari and Sponheim 2013; Vederman et al. 2012). Such misjudgment of emotion perception could be found even in the remitted patients with BD (Bozikas et al. 2007). The study in (Chen et al. 2006) showed that both the patients in manic mood and depressive mood would exhibit abnormal responses when they are recognizing the facial expressions with surprise, fear and happiness. In the research on UD, some studies have shown that the patients with UD have a bias in the recognition of positive and negative facial expressions compared to normal people (Surguladze et al. 2004; Langenecker et al. 2005). Another study showed that when the patients with BD and the patients with UD are being elicited by emotional stimuli, the responses could be different from that by the healthy people. No matter the emotional stimuli is positive or negative, the intensity of the responses of the patients with UD are slighter than that from the healthy people (Summers et al. 2006). In addition, when the patients with BD are being elicited by the stimuli of fear, surprise, anger and sadness, they may present responses with unpredictable emotions (Bersani et al. 2013).

As shown in these past studies, both the patients with BD and UD have some defects on emotion processing, such as low accuracy of emotion perception, more sensitivity of emotion, and less response capability of emotional stimuli. Accordingly, it is desirable to have a diagnosis aid system which could perform one-time evaluation in order to eliminate misdiagnosis.

In general, an immediate facial expression indicates a patient's feelings and emotions. On the other hand, speech can help patients to express their emotion and mood directly. For this reason, our research focuses on the patient's facial expressions and speech responses which are elicited by watching the videos with emotional stimuli for mood disorder detection.

However, the performance of mood disorder detection based on only facial or vocal modality still has its limitation. To further improve the recognition performance, a promising research area is to explore the data fusion strategy for effectively integrating facial and vocal cues (Wu et al. 2014). In the past, Audio/Visual Emotion Challenges (AVEC 2011–2013) (Schuller et al. 2011, 2012; Valstar et al. 2013) aimed at comparison of audiovisual signal processing and machine learning methods to advance emotion recognition systems. Data fusion strategy effectively integrating the facial and vocal cues has become the most important issue. Generally, a multimodal system that fuses different channels and cues to make a decision is expected to provide more accurate recognition compared to that obtained using only a unimodal system. In many classification tasks, the fusion of decisions or representations from models which used different input modalities for model training yields a significant improvement. There are three approaches to combining multiple modalities: feature-level, decision-level and model-level fusions. The feature-level fusion usually concatenates different modalities features into a super-vector as an input to build a classification model. This type of fusion is appropriate for synchronized modalities. Gunes and Pantic (2010) showed that feature-level fusion obtained better results than the decision-level fusion approach. Although feature-level fusion is expected to contain more information than decision-level fusion (Rattani et al. 2007), several shortcomings of this method exist. For example, feature vectors from different modalities with different temporal structures and metric levels are generally not correlated, and the increase of the dimension of feature vectors might lead to the problem of curse of dimensionality. In decision-level fusion, each modality builds its own unimodal model first, and then these unimodal decision results are combined by several rules (max, average, logical AND, logical OR, etc.) in the end. Compared to feature-level fusion, it is easier to incorporate knowledge of every unimodal models which have the same scale of decision outputs. However, a system that uses hybrid strategies of feature-level and decision-level fusion is conceivably more precise and robust. Model-level fusion aims to combine both feature-level and decision-level fusion methods to overcome their drawbacks by using the interaction and integration between modalities. Zeng et al. (2009) stressed that temporal structures of the facial and vocal modalities as well as their temporal correlations play an important role in the interpretation of human naturalistic audiovisual affective behavior. In this study, two features (AU profiles and Emotion Profiles) transformed from two modalities (visual and audio) are used. The Hidden Markov Model (HMM)-based multimodal fusion method is applied to integrate the audio–visual features as well as considering the temporal evolution of these two media to precisely characterize the signal characteristics for improving the mood detection performance.

The remainder of this paper is organized as follows: Sect. 2 introduces the collection process of our mood database; Sect. 3 describes the HSC-based autoencoder database adaptation; Sect. 4 gives an overview of system framework, and Sect. 5 presents the experimental setup, experimental results, and discussion. Finally, Sect. 6 is the conclusion and future work.

2 Mood database collection

To distinguish BD from UD, a mood database should be collected first for system training and evaluation. We cooperated with Chi-Mei Medical Center in Taiwan to collect a database containing the elicited facial expressions and speech responses of the patients with UD or BD. The serial number of the project approved by the Institutional Review Board (IRB) of Chi-Mei Medical Center is 10,403-002. We used six emotional videos, including happiness, fear, surprise, anger, sadness and disgust to elicit expressions of the subjects. The subjects' facial expressions and the speech responses of the subjects in the following interviews with a clinician after watching each eliciting video were collected. For elicitation-based database collection, the first thing is the selection of eliciting videos. We selected 11 high-rating emotional videos from (Gross and Levenson 1995; Cheng et al. 2013) and our subjective opinions. Table 1 shows the 11 candidate eliciting emotional videos, including the videos with the emotions of happiness, fear, surprise, anger, sadness and disgust (Bersani et al. 2013; Ekman 1999) for selection. The selection process was conducted by 55 students, based on manual selection and Chi squared test. According to the degree of freedom (dof = 1) and the α value (α = 0.05), we selected the videos with Chi squared value smaller than 3.841 as the eliciting emotional videos. In Table 1, the rows we mark as 'Y' with bold face are the six emotional videos finally selected. The speech responses of the participants in the interviews with a clinician were collected to construct the CHI-MEI mood speech database.

Table 1 The Chi squared test results for video selection

Video	Emotion of the eliciting video	χ^2	Choice
1	Anger	24.01	N
2	**Disgust** (Gross and Levenson 1995)	**0.27**	**Y**
3	Fear	5.98	N
4	**Happiness** (Cheng et al. 2013)	**0.93**	**Y**
5	**Sadness** (Gross and Levenson 1995)	**0.04**	**Y**
6	Surprise	14.27	N
7	**Surprise** (Cheng et al. 2013)	**3**	**Y**
8	**Anger** (Gross and Levenson 1995)	**0.01**	**Y**
9	Sadness	0.42	N
10	Happiness	0.98	N
11	**Fear** (Cheng et al. 2013)	**2.53**	**Y**

The rows marked as 'Y' with bold face are the six emotional videos finally selected

Figure 1 shows the mood database collection procedure. Before data collection, each participant with BD or UD will be assessed by the doctor to check if his/her physical and mental state is stable before participating in the evaluation. This step is to make sure that all the participants are in the stable state and could complete the process for data collection. The set of questionnaires including Depression and Somatic Symptoms Scale (DSSS), Mood Disorder Questionnaire (MDQ), Young Mania Rating Scale (YMRS), Simpson-Angus Extrapyramidal Side Effects

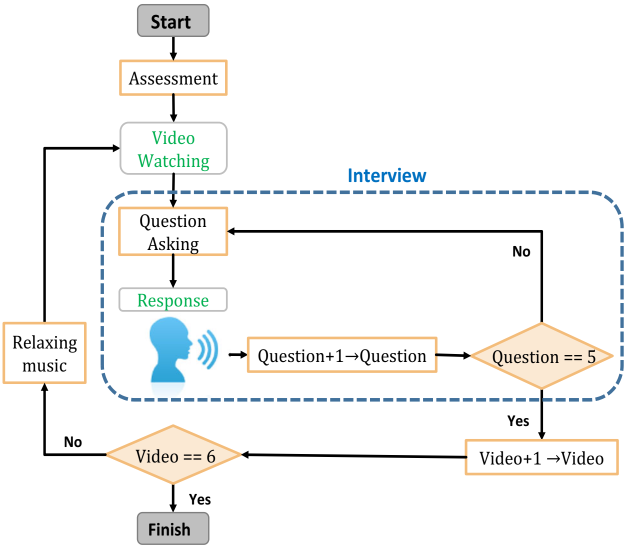

Fig. 1 The data collection procedure

Scale (SAS), Barnes Akathisia Rating Scale (BARS) and Clinical Global Impression (CGI) are used to examine the physical and mental states of each patients. Then, the doctor will determine whether the participant is suitable to participate in the experiment according to the results of the questionnaires. The assessment is based on the criterion presented in (Hamilton 1960; Hirschfeld et al. 2000; Young et al. 1978; Leucht et al. 1999; Barnes 1989; Guy 1976).

Before playing eliciting videos to the participant, a basic recording is conducted to ensure that the baseline characteristics of the speech responses of the participant can be properly collected. In this step, the clinician explained to the participant the whole recording procedure and asked them the following two questions to collect the baseline speech data of each participant before he/she was elicited by the emotional videos.

1. *What kind of videos did you watch on YouTube?*
2. *What is your favorite movie? Please describe it.*

After baseline recording, each participant watches six eliciting emotional video clips one by one. After watching each video clip, five pre-recorded questions are played to the participant for response collection sequentially. The five questions are:

1. *What do you think about the above video?* (*happy, sad, angry, disgusting, fearful and surprised*).
2. *How intense is it?* (*ranging from 1 to 5*).
3. *Which scene in the movie is impressive? Why?*
4. *Do you have any similar experience like that scene?*
5. *Are you feeling sick after watching above film?*

In order to ensure the data recording process will not be interrupted and the psychological state of the participant is stable, data collection was performed in a closed room. There were one clinician leading the process of data collection and one participant in the room at the same time. We designed a simple program installed on a laptop to guide the process of data collection. The clinician used this program to play six video clips sequentially. The video clips were played on a portable monitor, and the elicited audio and video information of the participant were recorded by the webcam (including camera and microphone) placed on the top of the monitor. Figure 2 shows the environment of data collection, arrangement of devices and the activity of the participants. The clinician side controlled the recording process. The participant side had a recoding screen on the desktop computer.

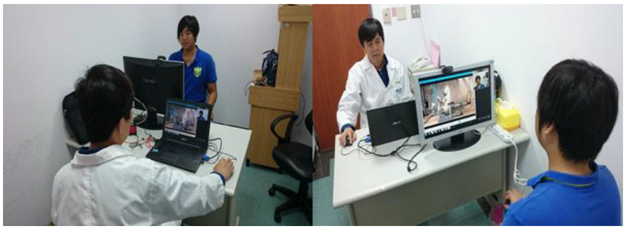

Fig. 2 Experimental environment from the clinician/participant side

This work collected the facial and speech responses from 26 participants, including 13 UDs and 13 BDs to construct the CHI-MEI mood database. The audio and video data were recorded simultaneously by the webcam placed on the top of the computer monitor. The recorded videos are AVI format with 640 × 480 dots per inch and 30 frames per second (30FPS), and the recorded speech data are monophonic with the format of 16 bits and 44.1 kHz sampling rate. Figure 3 shows the data structures of the speech responses during interview after watching the eliciting videos. Each participant answered five questions in six video phases and totally provided 30 facial image sequences and 30 speech response segments.

3 HSC-based autoencoder database adaptation

As this work focuses on speech emotional expression and the changes of facial expressions responding to emotional stimuli, a database big enough with manual labels, consisting of labeled Action Units (AUs) and emotions, is important for training the speech emotion model and the AU model for emotion and AU profile generation. Because the collected CHI-MEI mood database is small and difficult for emotion labeling, in order to deal with the small data problem, we apply a domain adaptation method called Hierarchical Spectral Clustering based denoising Autoencoder (HSC–DAE) to improve system performance. Inspired by the stacked denoising autoencoder domain adaptation (Glorot et al. 2011), in this study, we first use a

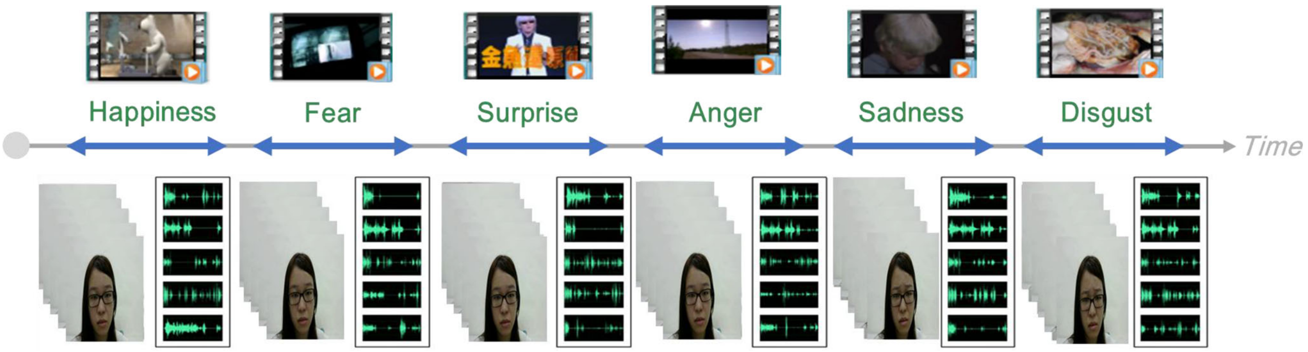

Fig. 3 The mood database structure for each facial and speech response after watching the corresponding eliciting video

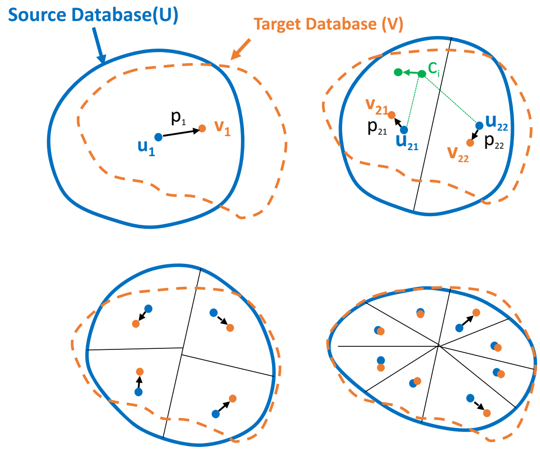

Fig. 4 Hierarchical spectral clustering (HSC)

cluster-based linear transform method, Hierarchical Spectral Clustering (HSC) (Furui 1989), to adapt the source domain data to the target domain and generate the transformed data which are the source domain data with some "relevant noises". Then, we use the transformed data as the input to train a denoising autoencoder (DAE) to reconstruct the source domain data as the domain-adapted data for further process. The function of the HSC algorithm is described as follows. The concept of database adaptation is shown in Fig. 4. In the upper left of the figure, the source data (U) and target data (V) are the two input databases. In the beginning, the data centroids of the source data and the target data are calculated as $u_1$and v_1 so that the deviation vector p_1 between these two centroids can be obtained. All of the source data are shifted by this deviation vector so that the corresponding centroids can coincide with each other. In the upper right of the figure, the number of clusters is doubled, and each datum in the target database is assigned to its nearest cluster. The same clustering process is applied to the source data. Then the centroid for each database and deviation vector are calculated the same as the preceding procedure.

3.1 Target and source databases

In order to apply domain adaptation, we collect the facial expressions and speech responses which are the reactions of the subjects when they are answering the questions after watching the eliciting videos. These collected audio–visual data are then used to construct the CHI-MEI mood database as our target domain.

We also select eNTERFACE and Extended Cohn-Kanade Database (CK+) for speech and action unit (AU) adaptation, respectively, as the databases of the source domain. The eNTERFACE database (Martin et al. 2006), containing six emotional expressions, including happiness, fear, surprise, anger, sadness and disgust, which are the same as the eliciting emotional videos, is thus adopted for emotion model training. In the eNTERFACE database, there were 42 subjects from 14 different nationalities including 24 males and 18 females. Each subject was asked to record 6 emotions, and there were 5 different sentences for each emotion. So each subject totally recorded 30 sentences. To express the characteristics of emotion better when recording, each subject was asked to listen to six consecutive stories. Each clip of the stories was expressed with a specific emotion. After the subject was familiar with six emotions, they expressed each of the six emotions to two judges for their determination if the data could be included in the database.

On the other hand, the Extended Cohn-Kanade Database (CK+ database) (Lucey et al. 2010) is adopted as the source domain for facial expression. Since we need labeled data for training, only expressions with explicit labels were selected in this database. This part of CK+ database contains 593 frame sequences of facial expressions from 123 subjects. Each of the frame sequences shows the variation of the facial expressions from the neutral frame to the peak frame in duration, and the length of these frame sequences are 10–60 frames. Because the selected data from the CK+ database are posed expression while CHI-MEI database is non-posed natural expression, the hierarchical spectral clustering method followed by a denoising autoencoder is employed for domain adaptation to alleviate the difference between them. In addition, we use the learned joint source and target feature representations to construct an SVM classifier to predict the AU profile and the other SVM to predict speech emotion profile for further process.

3.2 HSC-based denoising autoencoder

As mentioned before, we have labeled the training data drawn from the CK+ and eNTERFACE databases (source domain), while it is unlikely to label the training data drawn from the CHI-MEI database (target domain). The goal of this work is to find two functions for transferring the CK+ database to fit the facial expression part and the eNTERFACE to fit the speech response part of CHI-MEI database, respectively. Using the stacked denoising autoencoder framework, a data adaptation method based on Hierarchical Spectral Clustering followed by a denoising autoencoder (HSC–DAE) is proposed. The training process of the proposed HSC–DAE consists of two major steps. First, as shown in Fig. 4, the HSC algorithm is used for linear feature shift from the source database (CK+ and eNTERFACE, respectively) to the corresponding target data cluster (facial and speech data in the CHI-MEI database, respectively). After performing HSC algorithm, we get an adjusted vector (called target-adapted source data) which represents each sample of the source database adjusted to the target database. As we can see in Fig. 5, the target-adapted source data is mapped to a hidden representation $\boldsymbol{h}$ by:

$$\boldsymbol{h} = f_{\theta}(\tilde{}) = S(\boldsymbol{W}\tilde{\boldsymbol{x}} + \boldsymbol{b}) \tag{1}$$

and then the source data are reconstructed by:

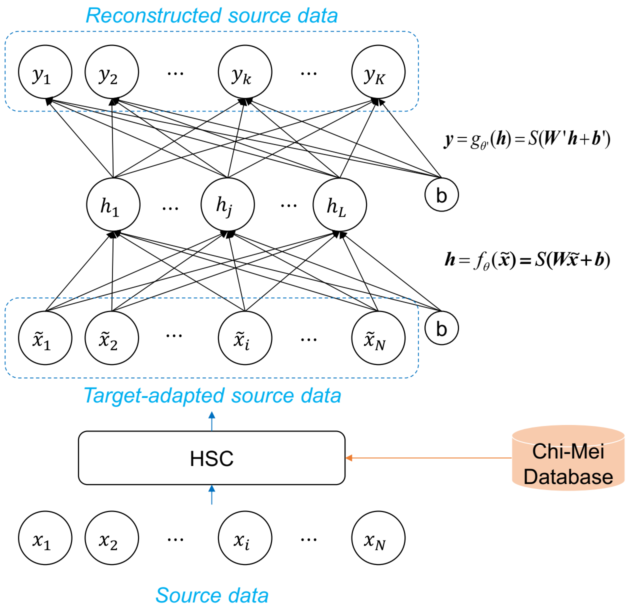
Reconstructed source data
y_1
y_2
…
y_k
…
y_K
$y = g_{\theta'}(\boldsymbol{h}) = S(\boldsymbol{W}'\boldsymbol{h} + \boldsymbol{b}')$
h_1
…
h_j
…
h_L
b
$\boldsymbol{h} = f_{\theta}(\tilde{x}) = S(\boldsymbol{W}\tilde{x} + \boldsymbol{b})$
$\tilde{x}_1$
$\tilde{x}_2$
…
$\tilde{x}_i$
…
$\tilde{x}_N$
b
Target-adapted source data
HSC
Chi-Mei Database
x_1
x_2
…
x_i
…
x_N
Source data

Fig. 5 Hierarchical spectral clustering-based denoising autoencoder

$$\boldsymbol{y} = g_{\theta'}(\boldsymbol{h}) = S(\boldsymbol{W}'\boldsymbol{h} + \boldsymbol{b}'), \tag{2}$$

where $\boldsymbol{W}$ is the weight matrix, $\boldsymbol{b}$ is the bias vector and S is the non-linear activation function (typically a logistic sigmoid function). Because HSC–DAE is trained to have reconstructed source data $\boldsymbol{y}$ as close as possible to the original source data $\boldsymbol{x}$ rather than the target-adapted source data $\tilde{\boldsymbol{x}}$(the same as the denoising concept), the backpropagation algorithm is used to update the weights of DAE and minimize the average reconstruction error by the following cost function:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{2}\|x - y\|^2 + \frac{\lambda}{2}\sum_{l=1}^{L}\sum_{j=1}^{J}(W_j^l)^2 \tag{3}$$

where the first term is an average sum-of-squared error and the second term is the weight decay (with a hyper-parameter λ) which can prevent overfitting. After training the HSC–DAE model, the reconstructed data of facial pointers are used as the input to build the AU detectors for predicting AU profiles of the CHI-MEI database. Similarly, the reconstructed speech features are used as the input to build the emotion detectors for predicting emotion profiles of the CHI-MEI database.

4 System framework

The system framework of this study is shown in Fig. 6. In the framework, first, the facial features of the CK+ database and the speech features of the eNTERFACE database are extracted and adapted to that of the CHI-MEI database using the HSC–DAE. The adapted facial data and speech are then used for training the support vector machine (SVM)-based AU detector and the emotion detector, respectively. Based on the constructed SVM-based AU detector and the emotion detector, the AU profiles and the emotion profiles are generated for feature representation. As each AU profile only characterizes one facial image, the AU profile sequence corresponding to the entire image sequence for one question response is extracted for facial feature representation. Finally, the Coupled Hidden Markov Model (CHMM)-based multimodal fusion method integrating the AU and the emotion profile sequence which characterize the temporal context of facial expression and speech emotion is adopted for mood disorder detection.

4.1 Speech segmentation and feature extraction

The speech signals in CHI-MEI mood database were collected by asking five questions to the subjects after watching each video clip. In order to make the subjects to

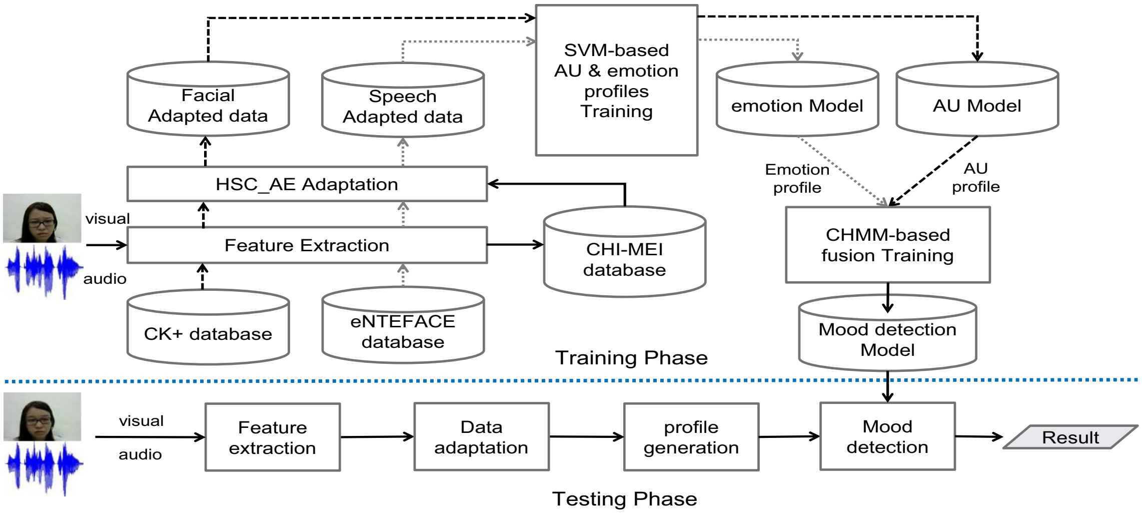

Fig. 6 The system framework

respond to the questions naturally and to encourage the subjects to provide more responses, a clinician conducted an interview with each subject in order to meet the above requirements for data collection. Accordingly, the speech segments were segmented to obtain only the speech responses from the patient by removing the silence as well as the speech segments from the clinician.

The research in (Giannakopoulos 2009) presented a simple method for silence removal and speech signal segmentation. The method is based on energy and spectral centroid as speech features to define a threshold. Through this threshold criterion, the silence area of speech signal is removed precisely instead of traditional segmentation using Endpoint Detection (Bou-Ghazale and Assaleh 2002), Zero Crossing Rate (ZCR) and Short Time Energy (STE) (Saha et al. 2005).

After dividing the speech data into multiple segments, 39 dimensions of Mel-frequency cepstral coefficients (MFCC) are extracted as acoustic features using the tool presented in (Hermansky and Morgan 1994). The K-means clustering is used to generate speaker-based vector quantization (VQ) codebooks for speaker recognition. In this paper, two codebooks are generated from each experimental subject, including codebook of the clinician and the subject. Probability Density Function (PDF) is used to estimate similarity of the segments for speaker recognition. The formula of PDF is shown in Eq. (4),

$$g(x, \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^{d}|\sum|}} \exp\left[-\frac{1}{2}(x-\mu)^{T}\sum{}^{-1}(x-\mu)\right] \tag{4}$$

where x is the input feature vector and μ and are the mean and covariance matrix of the training data, respectively.

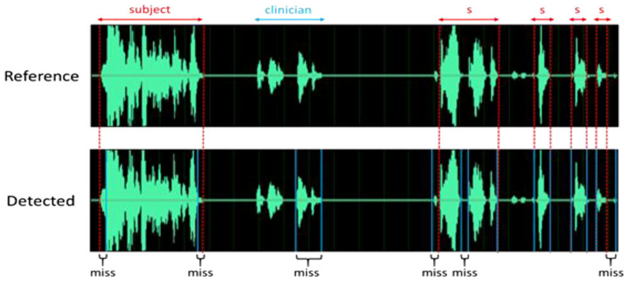

Fig. 7 Illustration of duration accuracy calculation

Figure 7 shows an example for estimating the duration accuracy. The red dotted lines are the real boundaries of subject's voices and the blue lines are the detected boundaries using the proposed system. The average length of each segment is 1,034 ms. After speech/speaker segmentation, openSMILE (Eyben et al. 2010) is employed to extract the acoustic features of 384 dimensions illustrated in Table 2. To represent the emotional expression of the input speech, Emotion Profile (EP) providing a quantitative measure for expressing the degree of the presence or absence of a set of basic emotions within an expression is obtained by constructing an SVM-based EP detector.

4.2 Facial feature extraction

The video signals in CHI-MEI mood database were collected by the subjects during the interview with the clinician after watching the eliciting video clips. We select the video segments which synchronize with the corresponding speech segments. For each frame in the video segments, Gauss–Newton Deformable Part Models (GN-DPM) is adopted for facial feature extraction (Tzimiropoulos and Pantic 2014). Figure 8 shows the facial feature points

Table 2 384-dimension acoustic features extracted using openSMILE

LLDs (16 × 2)	Functionals (12)
(Δ)RMS Energy	mean, standard deviation, kurtosis, skewness, extremes, value, range, relative position, linear regression, offset, slope, MSE
(Δ)ZCR	
(Δ)F0	
(Δ)Harmonic-Noise-Ratio	
(Δ)MFCC 1-12	

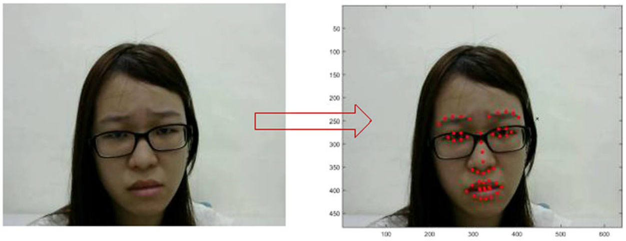

Fig. 8 The facial feature points extracted from a frame

extracted from a frame. Different from traditional feature extraction methods, the GN-DPM performs well for the unconstrained data by the appearance model trained in-the-wild. Therefore, for the data of all subjects of CHI-MEI database, GN-DPM is adopted to extract the facial features by a general appearance model, $A(c) = A_0 + A_c$. GN-DPM could optimize the facial feature parameters. First, at the pixel level, the vector of the appearance parameters and the vector of the shape parameters are optimized by Eqs. (5) and (6).

$$\Delta c = A^T (I - A(c) - J\Delta p) \tag{5}$$

$$\Delta p = H_P^{-1} J_P^T (I - A_0) \tag{6}$$

Based on the result from pixel level and the cost function of GN-DPM in (7),

$$\arg\min_{\Delta p, \Delta c} \sum_{j=1}^{N_P} \left|\left| I^j - A^j(c) - A^j \Delta c - J^j \Delta p \right|\right|^2 \tag{7}$$

the vector of the appearance parameters and the vector of the shape parameters at the part level are optimized by Eqs. (8) and (9),

$$\Delta c = \left(A_w^T A_w\right)^{-1} A_w^T (W(I - A(c)) - J_w \Delta p) \tag{8}$$

$$\Delta p = H_{P_w}^{-1} J_{P_w}^T (W(I - A_0)) \tag{9}$$

Finally, the fitting result of the facial feature parameters could be obtained. In each frame of the frame sequence, 49 facial feature points for describing the shapes and positions of facial components are extracted. For all of the facial feature vectors extracted from the image frames, the AU detectors trained by the HSC–DAE adapted database are used to estimate the occurrence probability of each AU. The AU profiles represent the posterior probabilities providing a quantitative measure for expressing the degree of the presence or absence of the basic AUs. Since the AU profile sequence is composed of the AU profiles generated from all the frames in the video segment, the variation of AU profile in the sequence could be considered as the temporal fluctuation of the subject's facial expression in the video segment, which could be used to distinguish the difference between AU profile sequences for mood disorder detection.

4.3 Coupled Hidden Markov Model (*CHMM*)

Hidden Markov Model (HMM) is a probabilistic graphical model that describes stochastic evolution of a set of random variables over time. An HMM with K states $S = \{s_1, s_2, \ldots s_K\}$ and length N is defined by the parameter set $\lambda = \{A, B, \pi\}$, where A is the set of observation probabilities conditioned on the state, B is the set of state transition probabilities, and π is the initial state probability parameter set. The HMM follows the Markov properties as the probability of the current state only depends on the previous state, formally $P(s_t|s_{t-1}, s_{t-2}, \ldots, s_1) \approx P(s_t|s_{t-1})$.

The CHMM is an extended version of the HMM. The main difference between the CHMM and HMM is that the CHMM comprises two HMMs that describe the state transitions of two objects, respectively, and the state transitions in each HMM mutually influence each other (Brand et al. 1997). Figure 9 shows the CHMM architecture which couples two HMM chains together. The circles represent the hidden nodes while the squares describe the observable nodes. The state transition probability of CHMM is $P(s_t^A|s_{t-1}^A, s_{t-1}^E)$ and $P(s_t^E|s_{t-1}^A, s_{t-1}^E)$ for AU profile and EP profile HMM, respectively. The current state is dependent on the states of its own chain and that of the other chain at the previous time step. The main feature of the CHMM is that it assesses contextual information between the current state (s_t^A, s_t^E) and its previous state (s_{t-1}^A, s_{t-1}^E). Given an observation sequence $O = [AU, EP]$, where a sequence of AU profile $AU = [au_1, \ldots, au_T]$ and a sequence of EP profile $EP = [ep_1, \ldots, ep_T]$ with T turns, the likelihood function of the state transition sequence with respect to the CHMM is estimated as follows:

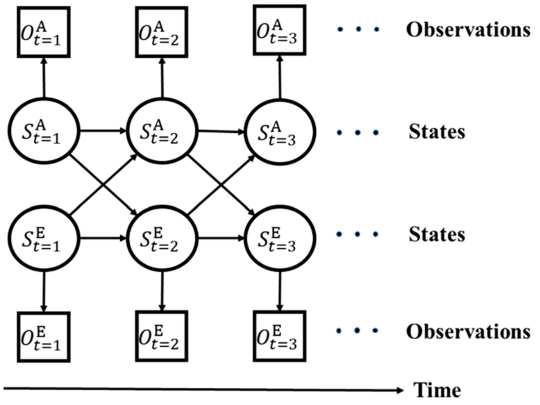

Fig. 9 The graphical topology of two-channel CHMM

$$L(\lambda) = P(s_1^A)P(s_1^E)\prod_{t=1}^{T} P(AU_t|s_t^A)P(EP_t|s_t^E) \times P(s_{t+1}^A|s_t^A, s_t^E)P(s_{t+1}^E|s_t^A, s_t^E) \qquad (10)$$

In Eq. (10), λ contains the parameters of the transition probabilities, the prior probabilities and the observation densities in the CHMM. $P(s_1^A)$ and $P(s_1^E)$ are the prior probabilities of the AU and EP chains, respectively. $P(AU_t|s_t^A)$ is the probability of the output of a given state in the AU chain. Similarly $P(EP_t|s_t^E)$ is the probability of the output of a given state in the EP chain. $P(s_{t+1}^A|s_t^A, s_t^E)$ is the probability of a state in the AU chain given the previous state in the AU chain and the EP chain at time t. Similarly we define $P(s_{t+1}^E|s_t^A, s_t^E)$ for the EP chain. Using the dynamic Bayesian networks to train the maximum likelihood of the CHMM is a well-known technique (Rezek et al. 2000). In this study, two CHMMs are defined and trained to predict the two mood disorders (BD and UD). As mentioned before, each participant totally provided 30 facial image sequences and 30 speech response segments, so the number of turns is T = 30. Therefore, the mood CHMM (BD or UD) with the highest likelihood is determined as the result.

5 Experiments and results

The main idea of this work is to fuse the speech and facial modalities for mood disorder detection. However, the performance of multimodal fusion depends on the contribution of each modality. More precisely, unimodal approach needs to provide unique information beyond others, then multimodal approach can joint those benefits to improve the overall performance. Therefore, each unimodal approach could improve its own performance to some level at least over average. To achieve this purpose, we applied HSC–DAE to adapt the bias data and then obtained more meaningful representation such as AUs and EPs, expecting to improve accuracy for mood disorder detection. After we got the effective method to transform the raw data to meaningful representation, CHMM-based multimodal fusion methods were tested with different classifiers for performance comparison.

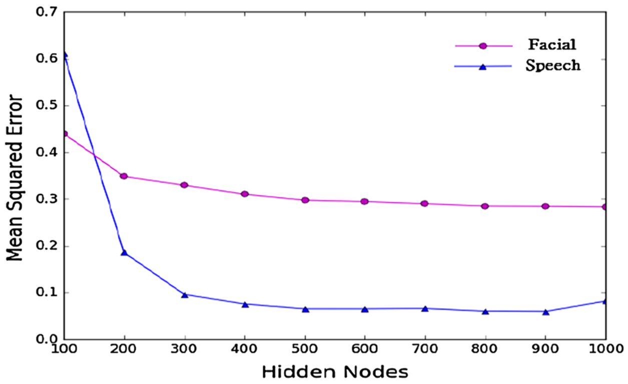

Fig. 10 Relation between MSE and the number of hidden nodes

5.1 Experimental setting

The experiments of this work were evaluated using 13-fold cross validation. Each fold contains the data from 24 subjects for training and 2 subjects (one from each category, i.e., BD or UD) for testing. The evaluation metric defined in Eq. (11) is the detection accuracy.

$$accuracy = (\#TB + \#TD)/(\#TB + \#FD + \#TD + \#FB) \qquad (11)$$

where #TB is the number of True Bipolars; #TD is the number of True Depressions; FB is the number of False Bipolars; and FD is the number of False Depressions. We used the chmmbox toolkit (Rezek et al. 2000) to construct the HMM and CHMM classifiers. Moreover, linearly scaling each attribute to the range (0, 1) for both training and test data was used. Additionally, the preprocess of the data is critical to the training procedure. For speech data, acoustic features of 384 dimensions were extracted for each speech segment. For facial data, we obtained 49 facial points to form a 98-dimensional feature vector. For optimizing the parameters used in the HSC–DAE, the number of hidden nodes should be determined first. As shown in Fig. 10, X-axis denotes the number of hidden nodes and Y-axis represents the Mean Squared Error (MSE) of the HSC–DAE. Speech features and facial point features were used, respectively to train the HSC–DAE by setting the single hidden layer with 100–1000 hidden nodes. The mean squared error shows the difference between the input data and reconstructed data. We should select the number of hidden nodes with smaller MSE. In both cases, we selected the reconstructed data which were trained by 900 hidden nodes with L2 regularization penalty on the activation values and the learning rate decay.

5.2 Experimental results

After HSC–DAE adaptation, adapted facial data and speech were then used for training the SVM-based AU detector and the emotion detector. Therefore, we compressed the facial points (98 dimensions) to the AU profiles (17 dimensions) and the acoustic features (384 dimensions) to the emotion profiles (6 dimensions) for feature representation, respectively. In order to verify the performance of HSC–DAE adaptation, we compared the results of AU and emotion profiles with original facial points and acoustic features. In Table 3, the input features of HMM including the facial points, acoustic features, AU profiles and emotion profiles are denoted by HMM_F, HMM_S, HMM_AU and HMM_EP, respectively. K is the number of states in HMM. For facial features, both HMM_F and HMM_AU achieved the highest accuracy of 53.85 %. However, for speech features, the HMM_S merely achieved 50.00 % accuracy but the HMM_EP could reach 53.85 % accuracy. In brief, using AU and emotion profiles could slightly improve the accuracy of mood disorder and reduce the dimensions of the input features to avoid the curse of dimensionality. Therefore, the AU features transformed from facial features via visual channel and EP features transformed from speech features via audio channel were modeled by two independent component HMMs as unimodal approach.

To evaluate the effects of fusion methods, the performances of the multimodal (feature-level (FF) and model-level (MF)) fusion approaches were compared. In our experiments, two modalities, FF and MF, were used to train an HMM by concatenating these two features as a tightly coupled input stream [*AU*, *EP*]. As a contrast, these two modal features were also adopted to train a CHMM for comparison. Therefore, we used the two methods to make decisions and compared their mood disorder detection results with previous unimodal methods as follows: (1) AU-only HMM (HMM_AU); (2) EP-only HMM (HMM_EP); (3) AU and EP concatenated HMM (HMM_AUEP); (4) CHMM.

Table 3 HMM-based detection performances of mood disorder

K	HMM_F (%)	HMM_S (%)	HMM_AU (%)	HMM_EP (%)
2	46.15	46.15	46.15	**53.85**
3	50.00	**50.00**	50.00	50.00
4	**53.85**	**50.00**	46.15	53.85
5	46.15	46.15	50.00	50.00
6	50.00	**50.00**	46.15	**53.85**
7	50.00	**50.00**	**53.85**	**53.85**
8	50.00	**50.00**	50.00	46.15

The columns shown with bold face are the hightest accuracy of the corresponding methods

The detection performances using previous HMM-based methods for different modalities are shown in Table 4. Here, we tested different hidden state number K from 2 to 8. Between the two unimodal methods, the HMM_EP performed better than the HMM_AU, which shows that EP features might be more informative than AU features. In addition, the HMM_AU obtained the worst result among the four mentioned methods. The main reason might be that the facial expressions are influenced by speaking effect. So the discriminability of AU is decreased.

Theoretically, multimodal approach gets more information which could perform better than the unimodal approach. In our experiments, the CHMM-based fusion method outperformed the unimodal methods. The results support this point. However, the performance of the feature-level HMM_AUEP was approximately equal to those of both unimodal methods (HMM_AU and HMM_EP). In fact, concatenation of the AU and EP features assumes each feature to have equal weight to contribute to mood disorder detection. However, AU and EP cannot provide complementary information for mood disorder detection in feature-level fusion.

In order to capture the dynamics of a signal in terms of some relatively static parameters, we estimated first order derivatives of each feature dimension of AU and EP to extract delta features (ΔAU and ΔEP). Table 5 shows the detection performances with additional delta features. Comparing Table 5 with Table 4, we can see that the delta features improved the performance in CHMM-based fusion method from 61.54 to 65.38 %. However, unimodal

Table 4 HMM-based detection performances of mood disorder

K	HMM_AU (%)	HMM_EP (%)	HMM_AUEP (%)	CHMM (%)
2	46.15	**53.85**	50.00	53.85
3	50.00	50.00	46.15	53.85
4	46.15	53.85	50.00	53.85
5	50.00	50.00	50.00	**61.54**
6	46.15	**53.85**	**53.85**	**61.54**
7	**53.85**	**53.85**	**53.85**	50.00
8	50.00	46.15	50.00	42.31

The columns shown with bold face are the hightest accuracy of the corresponding methods

Table 5 HMM-based detection performances of mood disorder with delta features (ΔAU and ΔEP)

K	HMM_AU (%)	HMM_EP (%)	HMM_AUEP (%)	CHMM (%)
2	46.15	46.15	46.15	57.69
3	46.15	46.15	46.15	50.00
4	50.00	**53.85**	**53.85**	**65.38**
5	46.15	50.00	50.00	61.54
6	**53.85**	**53.85**	**53.85**	**65.38**
7	50.00	46.15	50.00	53.85
8	50.00	50.00	50.00	50.00

The columns shown with bold face are the hightest accuracy of the corresponding methods

methods and feature-level fusion method could not get significant improvements from delta features. This is probably because those methods not like CHMM could share the information with each other. This drawback leads to the low accuracy in higher dimensions when adding delta features to the static features. Regarding to selecting the number of hidden states, K, the state number of the HMM-based method with the highest accuracy ranges from 4 to 7, especially when K is 6. This result shows that the six states coincide with the six elicited emotions mentioned in Sect. 2.

6 Conclusion and future work

This work aims to establish an approach to accurate distinction between BD and UD to make a correct and early diagnosis. We proposed a CHMM-based multimodal fusion approach to modeling the contextual information based on the temporal change of facial expressions and speech responses for mood disorder detection. Moreover, the HSC–DAE method was employed to adapt the facial expression and speech databases to CHI-MEI mood database to reduce the database bias problem. The adapted data were then used to construct the SVM-based detector for AU and emotion profile generation. Finally, the CHMM was applied to fuse the AUs and the EPs for mood disorder detection. We compared several HMM-based methods with the CHMM model fusion method. Experimental results show that the proposed CHMM-based fusion method outperformed the other HMM-based classifiers for mood disorder detection. In the future, combining more modalities is helpful to improve system performance. Considering personal characteristics of an individual patient is also an important factor in mood disorder detection.

References

American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders (DSM-5R): American Psychiatric Pub

Barnes TR (1989) A rating scale for drug-induced akathisia. Br J Psychiatry 154:672–676

Bersani G, Polli E, Valeriani G, Zullo D, Melcore C, Capra E, Quartini A, Marino P, Minichino A, Bernabei L (2013) Facial expression in patients with bipolar disorder and schizophrenia in response to emotional stimuli: a partially shared cognitive and social deficit of the two disorders. J Neuropsychiatr Dis Treat 9:1137

Bou-Ghazale SE, Assaleh K (2002) A robust endpoint detection of speech for noisy environments with application to automatic speech recognition. IEEE international conference on acoustics, speech, and signal processing (ICASSP), pp IV-3808–IV-3811

Bozikas VP, Kosmidis MH, Tonia T, Andreou C, Focas K, Karavatos A (2007) Impaired perception of affective prosody in remitted patients with bipolar disorder. J Neuropsychiatry Clin Neurosci 19:436–440

Brand M, Oliver N, Pentland A (1997) Coupled hidden Markov models for complex action recognition. In: Proc. Computer Vision Pattern Recognition, pp. 201–206

Chen CH, Lennox B, Jacob R, Calder A, Lupson V, Bisbrown-Chippendale R, Suckling J, Bullmore E (2006) Explicit and implicit facial affect recognition in manic and depressed states of bipolar disorder: a functional magnetic resonance imaging study. Biol Psychiatry 59:31–39

Cheng C, Chen H, Chan Y, Su Y, Tseng C (2013) Taiwan corpora of Chinese emotions and relevant psychophysiological data—normative data for Chinese jokes. Chin J Psychol 55:555–569

Cohn JF, Kruez TS, Matthews I, Yang Y, Nguyen MH, Padilla MT, Zhou F, La Torre FD (2009) Detecting depression from facial actions and vocal prosody. In: Proc. IEEE international conference on affective computing and intelligent interaction and workshops, pp 1–7

David DP, Soeiro-de-Souza MG, Moreno RA, Bio DS (2014) Facial emotion recognition and its correlation with executive functions in bipolar I patients and healthy controls. J Affect Disord 152:288–294

Ekman P (1999) Basic emotions. Handbook of cognition and emotion. Wiley, New York, pp 45–60

Erguzel TT, Sayar GH, Tarhan N (2016) Artificial intelligence approach to classify unipolar and bipolar depressive disorders. Neural Comput Appl 27:1607–1616

Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the international conference on multimedia, 2010, pp. 1459–1462

Furui S (1989) Unsupervised speaker adaptation based on hierarchical spectral clustering. IEEE Trans Acoust Speech Signal Process 37(12):1923–1930

Giannakopoulos T (2009) A method for silence removal and segmentation of speech signals, implemented in Matlab, Department of Informatics and Telecommunications, University of

Athens, Greece, Computational Intelligence Laboratory (CIL), Insititute of Informatics and Telecommunications (IIT), NCSR DEMOKRITOS, Greece

Glorot X, Bordes A, Bengio Y (2011) Domain adaptation for large-scale sentiment classification: a deep learning approach. In: Proceedings of the 28th international conference on machine learning (ICML-11) pp. 513–520

Goghari VM, Sponheim SR (2013) More pronounced deficits in facial emotion recognition for schizophrenia than bipolar disorder. Compr Psychiatry 54:388–397

Greco A, Valenza G, Lanata A, Rota G, Scilingo EP (2014) Electrodermal activity in bipolar patients during affective elicitation. IEEE J Biomed Health Inform 18(6):1865–1873

Gross JJ, Levenson RW (1995) Emotion elicitation using films. Cogn Emot 9:87–108

Gunes H, Pantic M (2010) Automatic, dimensional and continuous emotion recognition. Int J Synth Emot (IJSE) 1(1):68–99

Guy W (1976) Clinical global impression scale, The ECDEU assessment manual for psychopharmacology-revised. Volume DHEW Publ No ADM 76, vol 338, pp. 218–222

Hamilton M (1960) A rating scale for depression. J Neurol Neurosurg Psychiatry 23:56

Hermansky H, Morgan N (1994) RASTA processing of speech. IEEE Trans Speech Audio Process 2:578–589

Hirschfeld RM, Williams JB, Spitzer RL, Calabrese JR, Flynn L, Keck PE Jr, Lewis L, McElroy SL, Post RM, Rapport DJ (2000) Development and validation of a screening instrument for bipolar spectrum disorder: the mood disorder questionnaire. Am J Psychiatry 157:1873–1875

Howard N (2013) Approach towards a natural language analysis for diagnosing mood disorders and comorbid conditions. In: Proc. mexican international conference on artificial intelligence (MICAI), pp. 234–243

Lanata A, Greco A, Valenza G, Scilingo EP (2014) A pattern recognition approach based on electrodermal response for pathological mood identification in bipolar disorders. In: Proc. IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 3601–3605

Langenecker SA, Bieliauskas LA, Rapport LJ, Zubieta JK, Wilde EA, Berent S (2005) Face emotion perception and executive functioning deficits in depression. J Clin Exp Neuropsychol 27:320–333

Leucht S, Pitschel-Walz G, Abraham D, Kissling W (1999) Efficacy and extrapyramidal side-effects of the new antipsychotics olanzapine, quetiapine, risperidone, and sertindole compared to conventional antipsychotics and placebo. A meta-analysis of randomized controlled trials. Schizophr Res 35:51–68

Low LSA, Maddage NC, Lech M, Sheeber L, Allen NB (2010) Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. In: Proc. IEEE international conference on acoustics speech and signal processing (ICASSP), pp. 5154–5157

Low LSA, Maddage NC, Lech M, Sheeber L, Allen NB (2011) Detection of clinical depression in adolescents' speech during family interactions. IEEE Trans Biomed Eng 58(3):574–586

Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW), pp. 94–101

Martin O, Kotsia I, Macq B, Pitas I (2006) The eNTERFACE'05 audio-visual emotion database. In: Proceedings of the 22nd international conference on data engineering workshops (ICDEW'06), Atlanta, USA, 3–7 April 2006

Ooi KEB, Lech M, Allen NB (2013) Multichannel weighted speech classification system for prediction of major depression in adolescents. IEEE Trans Biomed Eng 60(2):497–506

Perlis RH (2005) Misdiagnosis of bipolar disorder. Am J Managed Care 11:S271–S274

Rattani A, Kisku DR, Bicego M, Tistarelli M (2007) Feature level fusion of face and fingerprint biometrics. In: Proceedings of the first IEEE international conference on biometrics: theory, applications, and systems, 2007, BTAS 2007, IEEE, pp 1–6

Rezek L, Sykacek P, Roberts SJ (2000) Learning interaction dynamics with coupled hidden Markov models. IEE proceedings—science, measurement and technology, vol 147. no. 6

Saha G, Chakroborty S, Senapati S (2005) A new silence removal and endpoint detection algorithm for speech and speaker recognition applications. In: Proceedings of the 11th national conference on communications (NCC), pp. 291–295

Sanchez MH, Vergyri D, Ferrer L, Richey C, Garcia P, Knoth B, Jarrold W (2011) Using prosodic and spectral features in detecting depression in elderly males. In: Proc. INTERSPEECH, pp. 3001–3004

Schuller B, Valstar M, Eyben F, McKeown G, Cowie R, Pantic M (2011) AVEC 2011 the first international audio/visual emotion challenge. In: Proc. First int. audio/visual emotion challenge and workshop (ACII), pp. 415–424

Schuller B, Valstar M, Eyben F, Cowie R, Pantic M (2012) AVEC 2012—the continuous audio/visual emotion challenge. In: Proc. of int. audio/visual emotion challenge and workshop (AVEC), ACM ICMI

Summers M, Papadopoulou K, Bruno S, Cipolotti L, Ron MA (2006) Bipolar I and bipolar II disorder: cognition and emotion processing. Psychol Med 36:1799–1809

Surguladze SA, Young AW, Senior C, Brébion G, Travis MJ, Phillips ML (2004) Recognition accuracy and response bias to happy and sad facial expressions in patients with major depression. Neuropsychology 18:212

Tzimiropoulos G, Pantic M (2014) Gauss-newton deformable part models for face alignment in-the-wild. In: Computer vision and pattern recognition (CVPR), 2014 IEEE conference, pp. 1851–1858

Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S et al (2013) AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the 3rd ACM international workshop on audio/visual emotion challenge, ACM, pp 3–10

Vederman AC, Weisenbach SL, Rapport LJ, Leon HM, Haase BD, Franti LM, Schallmo MP, Saunders EF, Kamali MM, Zubieta JK (2012) Modality-specific alterations in the perception of emotional stimuli in bipolar disorder compared to healthy controls and major depressive disorder. Cortex 48:1027–1034

Wu CH, Lin JC, Wei WL (2014) A survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA transactions on signal and information processing, vol. 3, e12

Young R, Biggs J, Ziegler V, Meyer D (1978) A rating scale for mania: reliability, validity and sensitivity. Br J Psychiatry 133:429–435

Zeng Z, Pantic M, Roisman G, Huang TS (2009) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31.1 pp. 39–58