INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Unipolar Depression vs. Bipolar Disorder: An Elicitation-based Approach to Short-Term Detection of Mood Disorder

Kun-Yi Huang[1], Chung-Hsien Wu[1], Yu-Ting Kuo[1], and Fong-Lin Jang[2]

[1] Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
[2] Department of Psychiatry, Chi Mei Medical Center, Tainan, Taiwan
chunghsienwu@gmail.com

Abstract

Mood disorders include unipolar depression (UD) and bipolar disorder (BD). In this work, an elicitation-based approach to short-term detection of mood disorder based on the elicited speech responses is proposed. First, a long-short term memory (LSTM)-based classifier was constructed to generate the emotion likelihood for each segment in the elicited speech responses. The emotion likelihoods were then clustered into emotion codewords using the K-means algorithm. Latent semantic analysis (LSA) was then adopted to model the latent relationship between the emotion codewords and the elicited responses. The structural relationships among the emotion codewords in the LSA-based matrix were employed to construct a latent affective structure model (LASM) for characterizing each mood. For mood disorder detection, the similarity between the input speech LASM and each of the mood-specific LASMs was estimated. Finally, the mood with its LASM most similar to the input speech LASM is regarded as the detected mood. Experimental results show that the proposed LASM-based method achieved 73.3%, improving the detection accuracy by 13.3% compared to the commonly used SVM-based classifiers.

Index Terms: Mood disorder, speech emotion recognition, latent affective structure model

1. Introduction

Unipolar depression (UD) and bipolar disorder (BD) are common mental illness. UD experiences two states: euthymia and depression (low). Different from UD, BD experiences two opposite and extreme emotional states: mania (high) and depression (low) through euthymia. The standard of diagnosis is based on Diagnostic and Statistical Manual of Mental Disorders (DSM-V-TR) [1] which defines symptoms, progressions and family medical history and so on. However, a high percentage of bipolar disorder patients are initially misdiagnosed as having unipolar disorder. This is because BD patients seek medical treatment more often when they are in depression state. This misdiagnosis carries significant negative consequences for the treatment of the BD patients. Therefore, it is crucial to distinguish 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, recognition of mood disorder based on speech and facial expression plays an important role in helping clinical diagnosis. UD has attracted many research efforts in different fields in recent years. The results show that these approaches can be used to distinguish depression from a control group in short-term detection [2-9]. However, there is only limited research on BD detection. Most of the studies were focused on long-term mood state monitoring (Mania, euthymia and depression) [10-12].

In short-term detection, the International Affective Picture System (IAPS)-pictures-slideshow intends to provide emotional stimuli of patients, and analyzes their electro dermal response to identify the state of the bipolar patients [13, 14]. Other research tried to identify the state of bipolar patients [15] using video to elicit the emotional state of the patients and collect their facial expressions. Besides facial expression-based approaches, in [16], speech was transformed into lexical information in real-time and used the pre-constructed model from the Alzheimer's database to determine axiological values and time orientation of lexical features. In human-machine communication, speech is the simplest and fastest way. Through speech, patients can express emotion to doctors directly, and doctors can also understand the mood and thought of the patients effectively. Therefore, this work, based on the elicitation from emotional videos, focuses on speech signals of interviews with the patients for distinguishing BD from UD. Firstly, the eliciting emotional videos were used to elicit the patients' emotional responses. Speech responses of the patients were collected through the interviews by a clinician after watching each of six emotional video clips. LSTM is a kind of recurrent neural network which can use their cyclic connections to capture a certain amount of context and it can overcome the vanishing gradient problem. The long-short term memory (LSTM)-based classifier was adopted to obtain the emotion likelihood for each speech response. Finally, based on the emotion likelihoods generated from the LSTM-based emotion classifier followed by emotion codeword mapping, a latent affective structure model (LASM) is proposed to characterize the structural relationship among the emotion codewords for each mood. The mood-based LASMs for BD and UD, respectively, are then used to compare with the input speech LASM for BD and UD detection.

2. Mood database design and collection

To the best of our knowledge, there is no speech database for the research on elicitation-based short-term detection of mood disorder currently. In this work, we cooperated with Chi-Mei Medical Center in Taiwan for mood database collection and

http://dx.doi.org/10.21437/Interspeech.2016-620

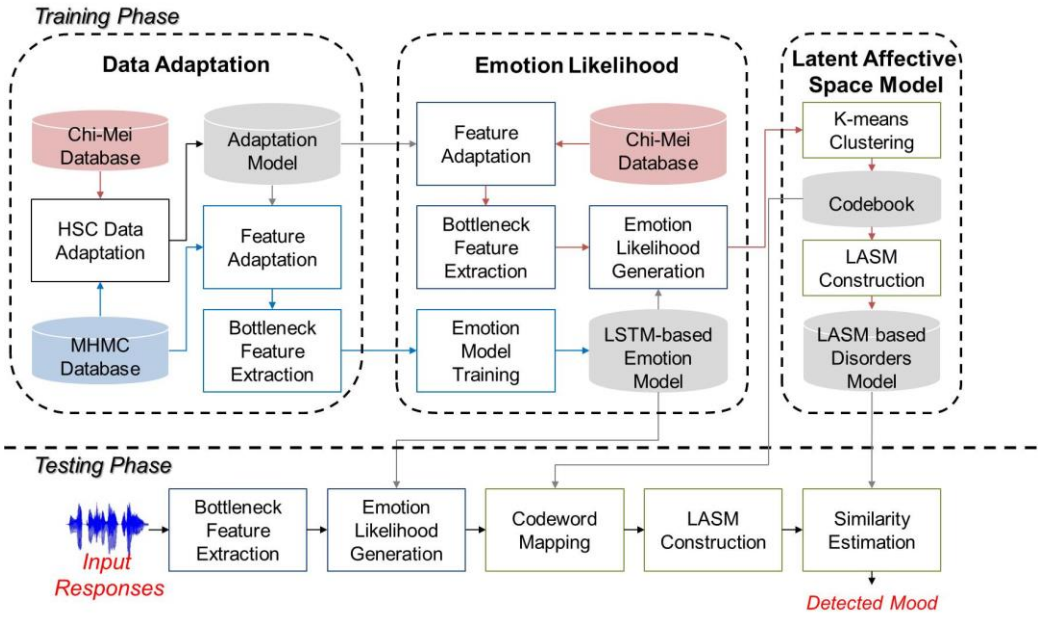

Figure 1: *The system framework*

evaluation. The serial number of the project approved by the Institutional Review Board (IRB) of Chi-Mei Medical Center is 10403-002. This project used six emotional videos, including happiness, fear, surprise, anger, sadness and disgust, to elicit facial expressions and speech responses of the subjects [15, 17]. The speech responses of the subjects in the interviews with a clinician were collected to form the CHI-MEI mood speech database.

Before data collection, each patient was asked to assess his/her mental status based on the criterion presented in [18]-[23]. A basic recording before playing eliciting video was conducted to ensure that the baseline characteristics of the speech responses of the patient can be properly collected. In this step, the clinician explained to the patient the whole recording procedure and asked them the following two questions to collect the baseline speech data of the patient before he/she was elicited by the emotional videos.

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

After baseline recording, each patient will watch six eliciting emotional video clips one by one. After watching each video clip, five pre-recorded questions were played to ask the patient 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?*

3. Proposed method

The system framework of the proposed method is illustrated in Fig. 1. In the training phase, the MHMC emotion database is adapted to the CHI-MEI mood database using the Hierarchical Spectral Clustering (HSC) algorithm [24]. The adapted MHMC database were then used to train the LSTM-based emotion generation model. Finally, an LASM [28] is constructed to model the structural relationships among the emotion codewords for each mood. In the test phase, the input speech responses are fed to the LSTM for emotion likelihood generation. The generated emotion likelihoods are used to construct the input LASM to compare with the LASM of each mood. Finally, the mood with its corresponding LASM most similar to the input LASM is determined as the detected mood for the input speech.

3.1. Data adaptation and bottleneck feature extraction

The MHMC emotion database contained six emotion, including happiness, fear, surprise, anger, sadness and disgust. Each emotion contained 200 emotional utterances from 53 university college students. As the collected CHI-MEI mood database is quite small and does not have emotion annotation, the MHMC emotion database was adopted as a reference database to construct an LSTM-based emotion classifier [26]. In order to solve the database bias problem, the HSC algorithm was used to adapt the data in MHMC emotion database to the data in CHI-MEI mood database. In this work, OpenSMILE was used to extract acoustic features, consisting of 32 dimensions of low-level descriptors (LLDs) with 12 kinds of functionals. Totally, 384 dimensions of acoustic features were extracted for each speech segment from each response.

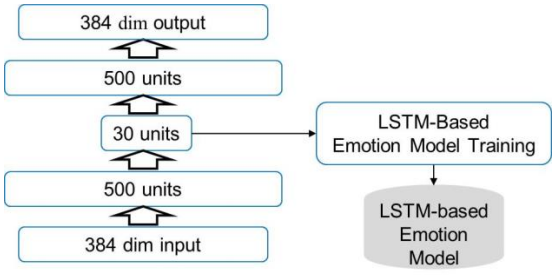

Figure 2: *Typical 3-hidden layer bottleneck feature extraction architecture used in this work.*

Fig 2 shows the bottleneck features directly from 384 features. This is different from most bottleneck features that are generated from alternative features. The 384-dimensional feature vector was used as the input to the deep neural networks [27] with 30 hidden units in the bottleneck layer and 500 hidden units for the other hidden layers. We used a learning rate of 0.2 for the remaining 200 epochs during fine-tuning.

3.2. Latent affective space model

For mood modeling, the emotion likelihoods obtained from the LSTM-based classifier were used to construct the LASM to model the structural relationship among the emotion codewords of the emotion likelihoods transformed using the K-means algorithm. Fig. 3 shows the constructed matrix for *D* codewords and *R* elicited responses.

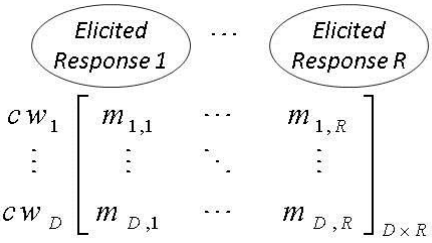

Figure 3: *The D-by-R matrix for D codewords and R elicited responses.*

Each speech response can be segmented into a speech segment sequence by using a fixed window size of 1 second with 50% overlap. The emotion likelihoods were obtained from the speech segment sequence using the LSTM-based emotion model. All the emotion likelihood vectors were used to construct an emotion codebook $CW = \{cw_1, cw_2, \ldots, cw_D\}$ with *D* codewords using the K-means algorithm.

In this work, two LASMs are constructed for the two moods: UD and BD, respectively. In order to estimate the importance of each emotion codeword for each response, the Emotion Codeword Frequency–Inverse Elicited Response Frequency (*ECF-IERF*) is defined as the entry value in the matrix. $m_{d,r}^{M}$ denotes the *ECF-IERF* of the *d*-th emotion codeword for the *r*-th response calculated as follows.

$$m_{d,r}^{M} = ECF^{M}(d,r) \times IERF^{M}(d,r) = \frac{cw_{d,r}^{M}}{\sum_{d} cw_{d,r}^{M}} \times \log \frac{|R|}{\left|\left\{r : cw_{d}^{M} \in R_r\right\}\right|} \quad (1)$$

where $M \in \{UD, BD\}$, $R = \{R_1, R_2, ..., R_6\}$ represents the set of responses and R_r is the *r*-th response. $|R|$ is the number of responses and the value is 6, representing six responses in this work. The LASM is proposed to characterize the structural relationship between the emotion codewords and the responses. Fig. 4 shows the process for LASM construction. First, the emotion likelihoods are mapped to the emotion codewords, which are then used to construct the LASM using *ECF-IERF*.

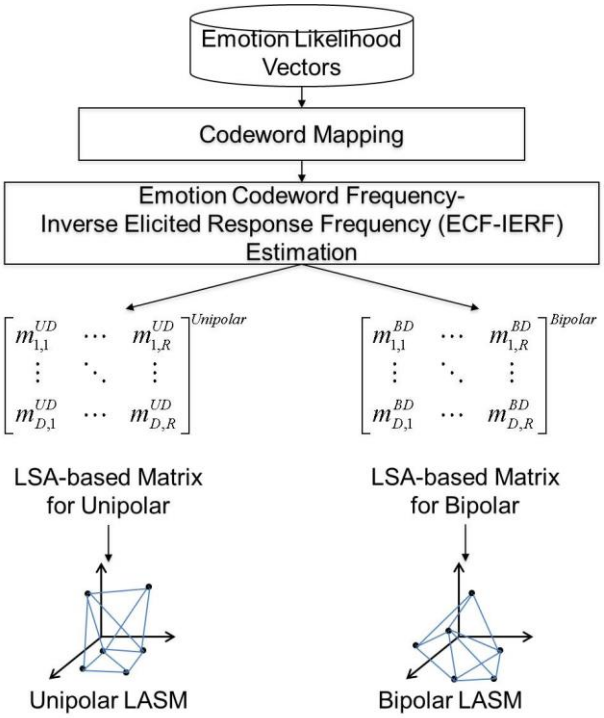

Figure 4: *Process of LASM construction.*

For mood detection, the input feature vector sequence EL^{O} is obtained from observation *O* based on the LSTM-based emotion classifier. cw^{O} is the obtained emotion codeword sequence. The Euclidean Distance (EUD) and Cosine Angle Distance (CAD) are used to calculate the similarity between the input speech LASM L^{O} and the LASM L^{M} of mood *M*. In order to balance the similarity from the EUD-based and the CAD-based, a linear combination of the similarities of these two terms is employed as follows.

$$\begin{aligned} M^{*} &= \arg\max_{M \in \{UD,BD\}} P(M \mid O) \\ &= \arg\max_{M \in \{UD,BD\}} P(M \mid EL^{O}) \\ &= \arg\max_{M \in \{UD,BD\}} P(M \mid cw^{O}) \\ &\approx \arg\max_{M \in \{UD,BD\}} Score(L^{M} \mid cw^{O}) \\ &\approx \arg\max_{M \in \{UD,BD\}} Sim(L^{O}, L^{M}) \\ &\approx \arg\max_{M \in \{UD,BD\}} \alpha \underset{EUD}{Sim}(L^{O}, L^{M}) + (1-\alpha) \underset{CAD}{Sim}(L^{O}, L^{M}) \end{aligned} \quad (2)$$

The calculation of the EUD-based similarity and CAD-based similarity between the LASMs of input *O* and mood *M* is defined in Eq. (3) and Eq. (4), respectively.

$$\underset{EUD}{Sim}(L^O, L^M) = \frac{1}{1 + e^{\sqrt{\sum_d \sum_r \left(m_{d,r}^O - m_{d,r}^M\right)^2}}} \quad (3)$$

$$\underset{CAD}{Sim}(L^O, L^M) = \frac{L^O \cdot L^M}{\|L^O\| \|L^M\|} = \frac{\sum_d \sum_r m_{d,r}^O \times m_{d,r}^M}{\sqrt{\sum_d \sum_r \left(m_{d,r}^O\right)^2} \times \sqrt{\sum_d \sum_r \left(m_{d,r}^M\right)^2}} \quad (4)$$

where $m_{d,r}$ is the d-th emotion codeword with respect to the r-th response.

4. Result and discussion

In this work, the speech responses were collected from 30 subjects, including 15 UDs and 15 BDs to construct the CHI-MEI mood database. The average response time of UD and BD is 58.89s and 79.68s. Table 1 show the each average response time in UD and BD at CHI-MEI mood database.

Table 1. Statistics of the CHI-MEI mood database

Response	UD	BD
R1	97.85s	57.27s
R2	85.29s	66.89s
R3	80.49s	30.66s
R4	82.48s	92.23s
R5	76.71s	74.94s
R6	55.26s	31.37s

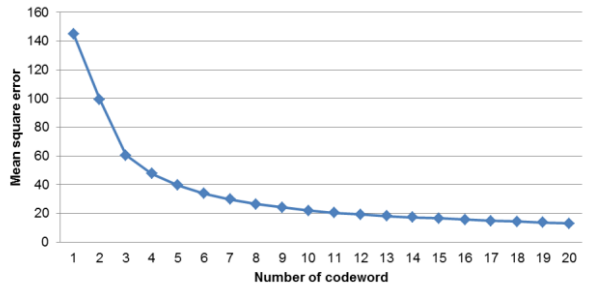

Figure 5: *Scree plot of vector quantization using K-means algorithm.*

Fig. 5 shows the scree plot of vector quantization using K-means algorithm. X-axis represents the number of emotion codewords and y-axis represents the mean square error under different number of codewords. In this work, 10 emotion codewords was selected for the experiments.

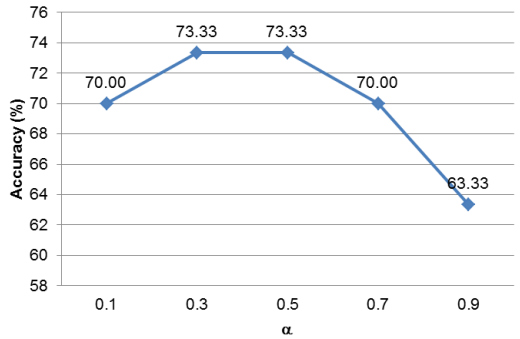

Figure 6: *Experimental results on the weighting factor α.*

Five-fold cross validation was employed for the following experiments. Fig. 6 shows the results as a function of the value of α. The x-axis is the value of α, and y-axis is the recognition accuracy. The best performance achieved 73.3% for the values of α from 0.3 to 0.6.

This work compared the performances of the proposed method with the commonly used classifiers support vector machines (SVMs). The libSVM was used for implementation and the RBF kernel was used. The results are shown in Fig. 7. The raw feature recognition using the 384-dimensional features by the SVM achieved 60.00%, while the emotion likelihood (EL) based on SVM was only 53.33%. The reason may rely on that the EL dimension was smaller than that of the raw features. The proposed method is better than the SVM-based method because the proposed method considers the structural relationship between the emotion codewords and the responses.

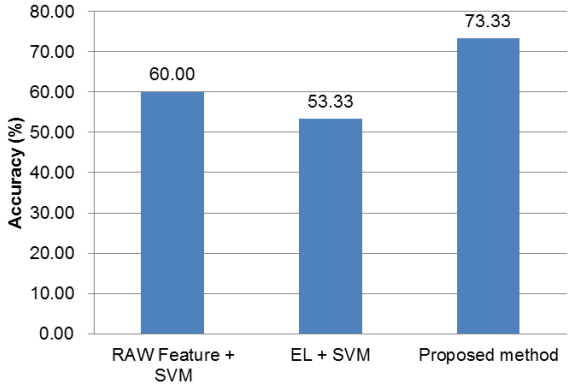

Figure 7: *Comparisons among the traditional SVM-based methods and the proposed method.*

5. Conclusions

This work proposed an LASM-based approach using the affective structure to model the structural relationships among the codewords in the six responses. For the database bias problem, the emotion database MHMC was used for adaptation using HSC for emotion likelihood generation. Two LASMs were constructed based on the *ECF-IERF* which represents the importance and relevance between the emotion codewords and the responses. Experimental results show that the proposed LASM-based method outperformed the commonly used SVM-based methods for mood disorder detection.

6. Acknowledgements

This paper was supported by the Ministry of Science and Technology of Taiwan under Contract NSC 102-2221-E- 006-094-MY3 and the Headquarters of University Advancement at the National Cheng Kung University, which is sponsored by the Ministry of Education, Taiwan.

7. References

[1] A. P. Association, *Diagnostic and statistical manual of mental disorders (fifth edn)*, American Psychiatric Association, 2011.

[2] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. D. La Torre, "Detecting depression from facial actions and vocal prosody," in Proc. IEEE International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, pp. 1-7, 2009.

[3] D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, and D. M. Wilkes, "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE Transactions on Biomedical Engineering, vol. 47, no. 7, pp. 829-837, 2000.

[4] R. Horwitz, T. F. Quatieri, B. S. Helfer, B. Yu, J. R. Williamson, and J. Mundt, "On the relative importance of vocal source, system, and prosody in human depression," in Proc. International Conference on Body Sensor Networks (BSN), pp. 1-6, 2013.

[5] 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 Proc. IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 5154-5157, 2010.

[6] L. S. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents' speech during family interactions," IEEE Transactions on Biomedical Engineering, vol. 58, pp. 574-586, 2011.

[7] K. E. B. Ooi, M. Lech, and N. B. Allen, "Multichannel weighted speech classification system for prediction of major depression in adolescents," IEEE Transactions on Biomedical Engineering, IEEE, vol. 60, no. 2, pp. 497-506, 2013.

[8] M. H. Sanchez, D. Vergyri, L. Ferrer, C. Richey, P. Garcia, B. Knoth, and W. Jarrold, "Using Prosodic and Spectral Features in Detecting Depression in Elderly Males," in Proc. INTERSPEECH, pp. 3001-3004, 2011.

[9] T. Yingthawornsuk and R. G. Shiavi, "Distinguishing depression and suicidal risk in men using GMM based frequency contents of affective vocal tract response," in Proc. International Conference on Automation and Systems, IEEE, pp. 901-904, 2008.

[10] A. Grunerbl, A. Muaremi, V. Osmani, G. Bahle, S. Ohler, G. Troster, O. Mayora, C. Haring, and P. Lukowicz, "Smartphone-based recognition of States and state changes in bipolar disorder patients," IEEE Journal of Biomedical and Health Informatics, vol. 19, no. 1, pp. 140-148, 2015.

[11] O. Schleusing, P. Renevey, M. Bertschi, S. Dasen, J. M. Koller, and R. Paradiso, "Monitoring physiological and behavioral signals to detect mood changes of bipolar patients," in Proc. International Symposium on Medical Information & Communication Technology (ISMICT), pp. 130-134, 2011.

[12] Z. N. Karam, E. M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, and M. G. Mcinnis, "Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4858-4862, 2014.

[13] A. Greco, G. Valenza, A. Lanata, G. Rota, and E. P. Scilingo, "Electrodermal activity in bipolar patients during affective elicitation," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 6, pp. 1865-1873, 2014.

[14] A. Lanata, A. Greco, G. Valenza, and E. P. Scilingo, "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, 2014.

[15] G. Bersani, E. Polli, G. Valeriani, D. Zullo, C. Melcore, E. Capra, A. Quartini, P. Marino, A. Minichino, and L. Bernabei, "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," Journal of Neuropsychiatric disease and treatment, vol. 9, pp. 1137, 2013.

[16] N. Howard, "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, 2013.

[17] P. Ekman, "Basic Emotions," in Handbook of Cognition and Emotion, New York, NY, John Wiley & Sons Ltd, pp. 45-60, 1999.

[18] M. Hamilton, "A rating scale for depression," Journal of neurology, neurosurgery, and psychiatry, vol. 23, no. 1, pp. 56, 1960.

[19] R. M. Hirschfeld, J. B. Williams, R. L. Spitzer, J. R. Calabrese, L. Flynn, P. E. Keck Jr, L. Lewis, S. L. McElroy, R. M. Post, and D. J. Rapport, "Development and validation of a screening instrument for bipolar spectrum disorder: the Mood Disorder Questionnaire," American Journal of Psychiatry, vol. 157, no. 11, pp. 1873-1875, 2000.

[20] R. Young, J. Biggs, V. Ziegler, and D. Meyer, "A rating scale for mania: reliability, validity and sensitivity," The British Journal of Psychiatry, vol. 133, no. 5, pp. 429-435, 1978.

[21] S. Leucht, G. Pitschel-Walz, D. Abraham, and W. Kissling, "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," Schizophrenia research, vol. 35, no. 1, pp. 51-68, 1999.

[22] T. R. Barnes, "A rating scale for drug-induced akathisia," The British Journal of Psychiatry, vol. 154, no. 5, pp. 672-676, 1989.

[23] W. Guy, "The clinical global impression scale," The ECDEU Assessment Manual for Psychopharmacology-Revised. Volume DHEW Publ No ADM 76, vol. 338, pp. 218-222, 1976.

[24] S. Furui, "Unsupervised speaker adaptation based on hierarchical spectral clustering," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, no. 12, pp. 1923-1930, 1989.

[25] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in Proc. International Conference on Data Engineering Workshops, pp. 8-8, 2006.

[26] S. Hochreiter, J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[27] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 30-42, Jan. 2012.

[28] C. H. Wu, H. P. Shen and C. S. Hsu, "Code-Switching Event Detection by Using a Latent Language Space Model and the Delta-Bayesian Information Criterion," IEEE Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1558-1570, Oct. 2014.