DETECTION OF DEPRESSION IN ADOLESCENTS BASED ON STATISTICAL MODELING OF EMOTIONAL INFLUENCES IN PARENT-ADOLESCENT CONVERSATIONS

*Melissa N Stolar, Margaret Lech and Nicholas B Allen**

School of Electrical and Computer Engineering, RMIT University
Melbourne, Australia
*Department of Psychology, University of Oregon, Eugene OR 97 403-1227, USA
s3164969@student.rmit.edu.au, margaret.lech@rmit.edu.au, nallen3@uoregon.edu

ABSTRACT

The current benchmark speech-based depression detection techniques rely on acoustic speech parameters collected from large sets of representative speech recordings. This study for the first time investigates depression detection based on the higher order influence model (HOIM) coefficients and emotional transition parameters derived from a relatively small set of conversational speech recordings representing 63 different parent-adolescent conversations of time duration 20 minutes each. The adolescents included 29 (24 female and 5 male) individuals diagnosed with major depressive disorder and 34 (24 female and 8 male) healthy individuals. The mental state of parents was not assessed. The model-based depression diagnosis was compared with benchmark techniques based on acoustic speech parameters (mel frequency cepstral coefficients (MFCC) and Teager energy operator (TEO)). The classification into depressed on non-depressed categories was performed using the Gaussian Mixture Model (GMM) for the acoustic parameters and the support vector machine (SVM) for the HOIM features. The model based technique led to the highest average classification accuracy of 94% of for the HOIM of order 4, whereas the best benchmark techniques scored 70% for the optimized MFCCs and 71% for the optimized TEO features.

Index Terms— Depression diagnosis, speech classification, conversation modeling, emotional influence model

1. INTRODUCTION

Individuals suffering from clinical depression undergo prolonged periods of excessive sadness, hopelessness, anger, guilt, desperation and loneliness often leading to suicidal thoughts. It has long been known that these specific emotional states experienced by people with clinical depression affect the acoustic qualities of their speech. Depressed speech has been often characterized subjectively as flat, monotone and dull, [20]. These perceptual qualities have been associated with fluctuations of objectively measured acoustic parameters such as speech energy, fundamental frequency (F_0), spectral slope, spectral stationarity, formant frequencies and glottal parameters. The idea of speech being a biomarker for depression has been proven to be feasible in some of the earlier studies investigating this topic [23], [4]. Later works provided further proof of the existing strong correlation between acoustic speech characteristics and the state of depression [6], [23], [24], [30], [18], [17], [19], [1], [2], [27]. Studies aimed specifically at depression in adolescents have extensively compared acoustic features including Teager energy operator (TEO) and mel-frequency cepstral coefficients (MFCC) [14], [13], [15]. Experiments based on speech samples from the Oregon Research Institute (ORI) database of audiovisual recordings, which included 68 depressed and 71 non-depressed adolescents, showed that the TEO features clearly outperformed all other features and feature combinations with accuracy ranging between 81%–87% for males and 72%–79% for females [14]. In [15], the TEO parameters combined with low-level acoustic descriptors led to 78% accuracy in depression detection for males and 75% for females. Manual facial action annotation (FACS) coding, active appearance modeling (AAM) and pitch extraction were used in [4] to measure facial and vocal expression. Classifiers using leave-one-out validation were support vector machine (SVM) for FACS and for AAM and logistic regression for voice. Both face and voice demonstrated moderate concurrent validity with depression. Accuracy in detecting depression was 88% for manual FACS and 79% for AAM. Accuracy for vocal prosody was 79%. Similarly, depression detection from facial images using Gabor wavelet features extracted at the facial landmarks led to about 79% accuracy [16]. Recent studies have shown that acoustic speech analysis based on a multi-channel speech classification approach that combines a number of different acoustic speech parameters, can detect symptoms of depression 2.5 years before the full blown symptoms become apparent [3]. The accuracy of this early prediction of risk for depression was reported to be up to

	Daughter		Son	
	D	ND	D	ND
Mother	21	24	5	8
Father	3	0	0	2

Table 1. Numbers of speech conversations. D - depressed adolescents, ND - non-depressed adolescents.

74% [23], [21], [22]. The main disadvantages of using acoustic or facial features in depression diagnosis are the reliance on the availability of large datasets of audiovisual recordings. Collection of such recordings can be costly and time consuming.

The current study investigates a new, alternative approach to the more traditional analysis of speech acoustics or facial image features of depressed individuals. The following hypothesis has been tested.

Hypothesis: It has been hypothesized that depressed adolescents exhibit different emotional behavior patterns in conversations compared with healthy non-depressed adolescents. These differences can therefore be used to efficiently detect symptoms of depression.

The proposed new approach creates a statistical model describing speakers' emotional behavior derived from conversational audio recordings that have been labeled with the observed emotional states of the speakers. The model parameters representing intra- and inter-speaker emotional influences, as well as the conditional probabilities of intra- and inter-speaker state transitions, were used as features to classify speech into depressed and non-depressed categories. The proposed modeling process represents an extended, higher order version of the first order emotional influence model for conversation analysis introduced in [28].

2. SPEECH DATA AND ANNOTATION

Speech data used to validate the proposed methodology was part of a larger collection of audio-visual recordings made by the Oregon Research Institute (ORI). It included audio recordings of 63 different dyadic conversations between a single parent (mother or father) and their adolescent child (son or daughter). The adolescents were between 14 and 18 years of age. All conversations had a natural, unscripted character where the participants were asked to discuss a predetermined topic of "planning a family event" The average time duration of each recorded conversation was about 20 minutes and the average ratio of the adolescents' to parents' speech duration across all conversations was 0.73. Based on self-reports and clinical interviews, 29 adolescent participants (24 female and 5 male), included in the validation dataset met the Diagnostic and Statistical Manual of Mental Disorders version IV (DSM-IV) criteria for a current episode of major depressive disorder (MDD) [3]; this group of adolescents is denoted in Table 1 as (D). The remaining 34 adolescent participants (24 female and 8 male) did not meet diagnostic criteria for any current psychiatric disorders and had no history of mental health treatment; this group represents a control set, and is denoted in Table 1 as (ND). The mental state of parents was not assessed. The distribution of participants' gender and the numbers of ND and D adolescents across the 63 dyads are given in Table 1. It shows that the number of available data were biased. Due to this heavily unbalanced gender distribution, the study could provide valid classification results only in a gender-independent case. The speech recordings were synchronized with second-by-second observed affect annotation manually labeled by trained psychologists using the living-in-family-environments (LIFE) coding system developed by the ORI [8]. For the purpose of conversation modeling speakers' states were denoted using four construct labels (positive (+), negative (-), neutral (n) and silence (s)) described in Table 2.

Speaker's Construct State	Speech Activity	Basic Categorical Emotions (observed by marker)
Positive (+):	Yes	Pleasant, happy or caring
Negative (-)	Yes	Contempt, anger, belligerence, anxious, dysphoric or whine
Neutral (n)	Yes	Neutral
Silence (s)	No	Not specified by the code

Table 2. Data annotation labels (speaker's states)

3. METHODOLOGY

3.1. Feature generation

3.1.1. Model based features

The higher order emotional influence model (HOIM) [28] represented a conversation between two people as two interacting Markov chains illustrated in Fig. 1, where each chain was given as a time sequence of speaker's states. All possible speaker's states S_t^i are listed in Table 2. The HOIM equation given in (1) estimated the joint conditional probabilities of a speaker i being in a state S_t given a set of previous states for both speakers.

$$\hat{P}\left(S_t^i | S_{t-n_1}^i, S_{t-n_1}^j, S_{t-n_2}^i, S_{t-n_2}^j, \dots, S_{t-n_N}^i, S_{t-n_N}^j\right) = \theta_{ii} P\left(S_t^i | S_{t-n_1}^i, S_{t-n_2}^i, \dots, S_{t-n_N}^i\right) + \theta_{ij} P\left(S_t^i | S_{t-n_1}^j, S_{t-n_2}^j, \dots, S_{t-n_N}^j\right) \quad (1)$$

Where N is the model order, n_i for $i = 1, \dots, N$ are time delays such that $n_1 < n_2 < \cdots < n_N$. The HOIM influence coefficients (ICs), θ_{ij} were constrained such that $\sum_j \theta_{ij} = 1$ and $\theta_{ij} > 0$ for $1 \le i, j \le M$, with M denoting the number of speakers. For each pair of speakers (dyad, M=2), the HOIM of order *N*, generated a set of four ICs (θ_{ij}). For *i=j* these coefficients represented the amount of emotional influence the speakers had on themselves (θ_{11} − parent→parent and θ_{22} − adolescent→adolescent) and for i≠j these parameters represented the amount of emotional influences the speakers had on their counterparts (θ_{12} − parent→adolescent and θ_{21} − adolescent→parent). The values of the ICs were determined using the expectation

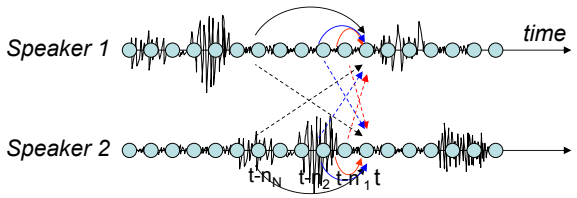

Fig. 1. Conversation between two speakers represented as two time-chains of speakers states. The arrows indicate intra- and inter-speaker transitions from previous states that affect the current states (at time t).

maximization (EM) data fitting method combined with data smoothing techniques described in more detail in [28]. The classification features were then composed as a *conversation signature* vector {**Φ**, **P**}, where **Φ** was a vector of the ICs $\mathbf{\Phi}=[\theta_{11},\theta_{12},\theta_{21},\theta_{22}]$ and **P** was a vector of the corresponding prior (estimated from the data) conditional probabilities of intra- and inter-speaker state transitions $\mathbf{P}=[p_{11},p_{12},p_{21},p_{22}]$ estimated as

$$p_{ij} = P\left(S_t^i | S_{t-n_1}^j, S_{t-n_2}^j, \ldots, S_{t-n_N}^j\right) \quad (2)$$

Where, N is the model order and i,j=1,2. The time delays used in the HOIM experiments were $n_k = k$ seconds for k=1, ...,N.

3.1.2. Acoustic (benchmark) features

Preprocessing:

The speech data sampled at 11 kHz was segmented into 16 ms frames with 50% overlap using the Hamming window. Within each frame, the audio signal amplitude was normalized to ± 1 range and the voiced speech segments were extracted and concatenated for feature extraction. The acoustic features including the mel frequency cepstral coefficients (MFCC) and the parameters derived from the Teager energy operator (TEO) were extracted on the frame-by-frame basis using 16 ms frames with 50% overlap between frames.

Mel-frequency cepstral coefficients (MFCCs):

The MFCCs [26] were calculated as the spectral amplitudes resulting from the discrete cosine transform of the log power spectrum mapped to the mel-scale. Initially, 13 coefficients were generated for each frame and after discarding the first one, the remaining 12 MFCCs were used as features.

Teager energy operator (TEO) parameters:

Acoustic speech features derived from the Teager Energy Operator (TEO) have been previously applied in emotion [7], stress [29] and depression [23] classification systems. The process of calculating the TEO parameters followed the frame-based method introduced in [32], which calculates the area under the TEO autocorrelation envelope within 17 frequency bands. The frequency bands were obtained through the Wavelet Packet analysis of speech signals as close estimates of the critical bands characterizing the human auditory system. Given signal samples x[m] for each 16 ms frame, the TEO (instantaneous energy) values were calculated as [10], [32]

$$\Psi(x[m]) = x^2[m] - x[m-1]x[m+1] \quad (3)$$

The instantaneous energy was then used to evaluate the TEO autocorrelation function given as [32]

$$R_{\Psi(x)}[\tau] = \frac{1}{2K+1}\sum_{m=-K}^{K}\Psi(x[m])\Psi(x[m+\tau]) \quad (4)$$

Where K is the number of samples in the given frame. After smoothing with cubic splines, the area under the autocorrelation contour was estimated for each frame within each of the 17 frequency bands leading to 17 feature parameters/frame.

3.2. Optimal feature selection

Large sets of raw features generated on the frame-by-frame basis were likely to contain a significant amount of redundancies that could act like noise and reduce the classification accuracy. Therefore, selecting a small subset of "*optimal*" features that show the strongest correlation with the state of depression was essential for accurate detection of depression and reduction of data dimensionality. The reduction of data dimensionality was of particular importance in the case of model-based classification, as the number of features increased exponentially with the model order. The optimal feature selection was achieved using the minimum redundancy and maximum relevance (mRMR) filter approach based on calculation of the mutual information quotient (MIQ) [5]. Features selected by the mRMR method were further optimized using the wrapper procedure suggested in [25]. The wrapper algorithm conducted an iterative K-fold cross-validation search (with 80% training and 20% testing data) leading to a set of features that gave a minimum value of the classification error.

3.3. Classification

The choice of classifiers providing efficient discrimination into two classes "*depressed*" and "*non-depressed*" was determined experimentally. It was found that the HOIM features were best suited to the support vector machine classifier (SVM) [31], whereas the TEO and MFCC and TEO features worked well with the Gaussian mixture model (GMM) classifier with 5 Gaussian mixtures [31]. The training and testing procedures for both GMM and SVM followed the leave-one-out cross-validation (LOOCV) process with 80% of data used in training and 20% in testing. The classification results were described in terms of an average accuracy, sensitivity and specificity, where the sensitivity was defined as

$$sensitivity = \frac{true\ depressed}{true\ depressed - false\ nondepressed} \quad (5)$$

and the specificity was defined as

$$specificity = \frac{true\ nondepressed}{true\ nondepressed - false\ depressed} \quad (6)$$

Acoustic Features	Average Depression Classification Results		
	Accuracy	Sensitivity	Specificity
MFCC	58%	60%	57%
MFCC optimal	70%	59%	80%
TEO	56%	54%	60%
TEO optimal	**71%**	**71%**	**72%**

Table 3. Average results of the depression detection based on raw and optimized acoustic speech parameters.

When assessing the performance, a well performing system would have high values for all of these three parameters; however, if a compromise had to be made, it was desired for the sensitivity to be slightly higher than the specificity. This way, a safer depression screening assessment could be made without missing on too many "*true depressed*" cases.

4. EXPERIMENTS AND RESULTS

4.1. Depression detection based on acoustic speech parameters (benchmark approach)

The benchmark results are shown in Table 3. It can be observed that the use of optimal feature selection led to a significant improvement of the classification outcomes for all types of acoustic features. The best performance of 71% accuracy and good 71/72 specificity to sensitivity ratio was given by the optimized TEO parameters. A close accuracy of 70% was given by the optimal MFCC, however in this case there was a very low sensitivity value of only 59%. This relatively high performance of the TEO parameters compared with the MFCC is consisted with similar results reported in [13], [15], [23].

4.2. Depression detection based on the HOIM parameters (proposed approach)

Like in the case of acoustic features, selection of optimal HOIM based features led to significant improvement of the classification outcomes (see Tables 4&5). The HOIM based depression classification results for the optimized features for HOIM of order 1-5 are presented in Table 5. While models of order 1-3 and 5 showed quite similar performance of 75-82% accuracy, the model of order 4 led to an outstandingly high performance of 94% accuracy and equally high sensitivity to specificity ratio of 93/94. Given that the time delay between the current and the previous state used in the modeling process was 1 second, the high performance observed in the of model order 4 indicates that the time duration needed to form distinct emotional interaction patterns showing clear differences between depressed and non-depressed family environment was about 4 seconds. This appears to be consistent with [11], [12], where the same adolescent's data was analyzed showing that during the time delay of up to 5 seconds the depressed adolescents were likely to remain in the same state, whereas their non-depressed counterparts change their state more frequently. This phenomenon known as the "*emotional inertia*" is one of the emotional characteristics attributed to the depressed adolescents.

HOIM Order	Average Depression Classification Results		
	Accuracy	Sensitivity	Specificity
1	62%	55%	68%
2	70%	62%	77%
3	64%	70%	59%
4	68%	65%	71%
5	70%	66%	74%

Table 4. Average results of the depression detection based on raw HOIM parameters.

HOIM Order	Average Depression Classification Results		
	Accuracy	Sensitivity	Specificity
1	81%	72%	88%
2	75%	69%	79%
3	82%	79%	85%
4	**94%**	**93%**	**94%**
5	78%	79%	76%

Table 5. Average results of the depression detection based on optimized HOIM parameters.

5. CONCLUSIONS

The results presented here strongly confirmed the initial hypothesis assuming the existence of significant differences in emotional interaction patterns between parents and their adolescent children in families with depressed adolescents and in families with healthy non-depressed adolescents.

The study showed that the HOIM of emotional interactions in conversations provided very efficient detection of major depression in adolescents. The HOIM depression detection method outperformed the benchmark approaches based on acoustic speech parameters (MFCCs and TEO parameters). The HOIM results were also higher than the results provided by most of the speech-based depression detection techniques mentioned in Section 1.

The well-recognized strong relation between depression disorder and the emotional state of a person [9] indicates that both the acoustic and model based depression recognition systems make their decisions by detecting emotional differences between depressed and non-depressed speakers. The stronger performance of the model based technique can be attributed to the fact that it relies on intra- and inter-speaker emotional interactions, whereas the acoustic speech analysis is limited to the intra-speaker observations only.

Some of the most important limitations of this study that need to be addressed in future research include: the lack of gender-dependent results, the scope being limited to parent-adolescent conversation and to one specific topic only, and the lack of mental health assessment of parents.

6. ACKNOWLEDGEMENT

The authors would like to thank Dr Lisa Sheeber from the Oregon Research Institute and all volunteers who participated in the collection of data used in this study.

12. REFERENCES

[1] S Alghowinem, R Goecke, M Wagner, J Epps, T Gedeon, M Breakspear, and G Parker, "A comparative study of different classifiers for detecting depression from spontaneous speech," Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 26-31 May 2013, 8022 – 8026.

[2] S Alghowinem, R Goecke, , M Wagner, J Epps, M Breakspear, and G Parker, "Detecting depression: A comparison between spontaneous and read speech", Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 26-31 May 2013, 7547 - 7551.

[3] American Psychiatric Association, "Diagnostic and Statistical. Manual of Mental Disorders", 4th ed., American Psychiatric Association, Washington, DC, 1994.

[4] JF Cohn, TS Kruez, I Matthews, Y Yang, MH Nguyen, M Padilla, and F De la Torre, "Detecting depression from facial actions and vocal prosody". Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, 10-12 Sept. 2009, 1 - 7.

[5] C Ding and H Peng, "Minimum redundancy feature selection from microarray gene expression data," J of Bioinformatics and Computational Biology, 2005, 3(2), 185-205.

[6] DJ France, RG Shiavi, S Silverman, M Silverman, "Acoustical properties of speech as indicators of depression and suicidal risk," IEEE Trans Biomed Eng., 2000, 47(7), 829-837.

[7] A Georgogiannis, V Digalakis, "Speech Emotion Recognition using non-linear Teager energy based features in noisy environments", Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European, 27-31 Aug. 2012, 2045 – 2049.

[8] H Hops, B Davis, N Longoria, "Methodological Issues in Direct Observation-Illustrations with the Living in Familial Environments (LIFE) Coding System," Journal of Clinical Child Psychology, vol. 24, no. 2, pp. 193-203, 1995.

[9] R Ingram (Editor), The International Encyclopedia of Depression, Springer New York, 2009.

[10] JF Kaiser, "On a simple algorithm to calculate the energy of a signal," ICASSP 1990.

[11] P Kuppens , LB Sheeber, MB Yap, S Whittle, JG Simmons, "Emotional Inertia Prospectively Predicts the Onset of Depressive Disorder in Adolescents", Emotion, vol. 12, no. 2, pp. 283-289, 2012.

[12] P Kuppens, NB Allen, LB Sheeber, "Emotional Inertia and Psychological Maladjustment," Psychological Science, vol. 21, no. 7, pp. 984-991, 2010.

[13] A Low, MC Maddage, M Lech, N Allen, "Mel Frequency Cepstral feature and Gaussian mixtures for modelling clinical depression in adolescents", Cognitive Informatics, 2009. ICCI '09. 8th IEEE International Conference on, 15-17 June 2009, 346 - 350.

[14] LS Low, NC Maddage, M Lech, LB Sheeber, NB Allen, "Detection of clinical depression in adolescents speech during family interactions", IEEE Transactions, Biomedical Engineering, 2011, 58(3), 574-586.

[15] A Low, MC Maddage, M Lech, L Sheeber, N Allen, "Influence of acoustic low-level descriptors in the detection of clinical depression", Acoustics Speech and Signal Processing 2010 IEEE International Conference on, 5154 - 5157.

[16] NC Maddage, R Senaratne, LS Low, M Lech, N Allen, "Video-based detection of clinical depression in adolescents", Conf Proc IEEE Eng Med Biol Soc. 2009;2009:3723-6.

[17] E II Moore, MA Clements, JW Peifer, L Weisser, "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression", IEEE Trans Biomed Eng., vol.55, no.1, pp.96-107, 2008.

[18] E II Moore, M Clements, J Peifer, L Weisser, "Comparing objective feature statistics of speech for classifying clinical depression",Conf Proc IEEE Eng Med Biol Soc.2004;1:17-20.

[19] E II Moore and M. Clements, "Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information", Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on, 17-21 May 2004, 101-4 vol.1.

[20] P Moses, The Voice of Neurosis. New York: Grune & Stratton, 1954.

[21] KEB Ooi M Lech M., NB Allen, "Multi-Channel Weighted Speech Classification System for Prediction of Major Depression in Adolescents", IEEE Transactions on Biomedical Engineering 2013, 60(2), 497-506.

[22] KEB Ooi, M Lech and NB Allen, "Prediction of major depression in adolescents using an optimized multi-channel weighted speech classification system", Elsevier, Biomedical Signal Processing and Control, Vol. 14, Nov 2014, pp. 228-239.

[23] KEB Ooi, L Low, M Lech, N Allen. "Early prediction of major depression in adolescents using glottal wave characteristics and Teager rnergy parameters", Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on, 25-30 March 2012, 4613 - 4616.

[24] A Ozdas, RG Shiavi, SE Silverman MK Silverman, DM Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near suicidal risk", IEEE Trans Biomed Eng, 2004, 51(9), 1530-1540.

[25] H. Peng, L. Berkeley; F. Long F; C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevancy, and min-redundancy", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(8), 1226-1238.

[26] T Quatieri, Discrete-Time Speech Signal Processing: Principles and Practice, Prentice Hall, 2002.

[27] S Scherer, J Pestian, LP Morency, "Investigating the speech characteristics of suicidal adolescents", Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 26-31 May 2013, 709 - 713.

[28] MN Stolar, M Lech, LB Sheeber, I Burnett and NB Allen, "Introducing Emotions to the Modelling of Intra- and Inter-Personal Influences in Parent-Adolescent Conversations", IEEE Trans on Affective Computing 2013, 4(4), 372-385.

[29] JM Susskind, GE Hinton, J Movellan, AK Anderson, Generating Facial Expressions with Deep Belief Nets. In V. Kordic Affective Computing, Emotion Modelling, Synthesis and Recognition. ARS Publishers, 2008.

[30] Y Yang, C Fairbairn, JF Cohn, "Detecting depression severity from vocal prosody", Affective Computing, IEEE Transactions on, 2013, 4(2), 142 - 150.

[31] S Young, "HTK: The Hidden Markov Model Toolkit V3.4," 1993, http://htk.eng.cam.ac.uk

[32] G Zhou, JHL Hansen, JF Kaiser, "Nonlinear feature based classification of speech under stress", IEEE Transactions, Speech and Audio Processing, 2001, vol. 9, pp. 201-216.