# EARLY PREDICTION OF MAJOR DEPRESSION IN ADOLESCENTS USING GLOTTAL WAVE CHARACTERISTICS AND TEAGER ENERGY PARAMETERS

*Kuan Ee Brian Ooi, Lu-Shih Alex Low, Margaret Lech and Nicholas Allen**

School of Electrical and Computer Engineering, RMIT University, Melbourne 3001, Australia
*ORYGEN Research Centre and Department of Psychology, University of Melbourne, Melbourne 3010, Australia
{k.ooi, lushih.low}@student.rmit.edu.au, margaret.lech@rmit.edu.au, nba@unimelb.edu.au

## ABSTRACT

Previous studies of an automated detection of Major Depression in adolescents based on acoustic speech analysis identified the glottal and the Teager Energy features as the strongest correlates of depression. This study investigates the effectiveness of these features in an early prediction of Major Depression in adolescents using a fully automated speech analysis and classification system. The prediction was achieved through a binary classification of speech recordings from 15 adolescents who developed Major Depression within two years after these recordings were made and 15 adolescents who did not developed Major Depression within the same time period. The results provided a proof of concept that an acoustic speech analysis can be used in early prediction of depression. The glottal features made the strongest predictors of depression with 69% accuracy, 62% specificity and 76% sensitivity. The TEO feature derived from glottal wave also provided good results, specifically when calculated at the frequency range of 1.3 kHz to 5.5 kHz.

***Index Terms***— Clinical depression, prediction, adolescents, glottal parameters, speech acoustics

## 1. INTRODUCTION

Onset depression is one of the most common but also most life-threatening conditions facing young people worldwide. If not detected or treated, it can seriously affect social, emotional, educational and vocational outcomes. It is also the most common precursor of suicide. There is an urgent need for an objective diagnostic test that can recognize early symptoms of depression, or predict the risk of a young person developing depression. Predictive signs of depression would be able to provide early warnings where swift measures can be taken to avoid the possible development of this mental disorder. The problem of an early detection of depression signs has been targeted by various analysis methods from different fields of research and it is an ongoing and very challenging problem. Due to a vast number of confining factors, combined efforts are needed to design efficient diagnostic and preventative strategies. Our experiments aim to contribute to these efforts by investigating the possibility of using a fully automated, computational approach to predict depression in adolescents who are currently not suffering from depression. The approach proposed here builds up on our previous experience with fully automatic detection of depression in adolescents who have already developed depression [8], [9]. Depressed speech has been consistently characterized by clinicians as dull, monotone, monoloud, lifeless, and "metallic" [11]. Some of the most significant studies on automatic detection of depression in speech include works by [10] where prosodic, vocal tract and glottal features were investigated for depression detection in a small group of 15 males (6 depressed and 9 control) and 18 females (9 depressed and 9 control). Speech samples were recorded during reading of a fixed text passage. It was demonstrated that the best performing classification method based on combined glottal and prosodic features and the quadratic discriminant classifier can provide 91% accuracy in males and 93% accuracy in females. In our recent study [8], we investigated acoustic correlates of depression in a large sample of 138 participants (68 clinically depressed and 71 controls). Most importantly for the first time natural speech samples were analysed instead of samples from patients' interviews or text readings. The speech recordings were made during discussions between family members closely approximating a typical family environment. A comparison between a wide range of glottal, prosodic, spectral and Teager energy operator (TEO) based features showed that the TEO features clearly outperformed all other features and feature combinations with classification accuracy ranged from 81% to 87% for male speakers and from 72% to 79% for female speakers.

Current studies aiming to determine the risk of developing symptoms or being diagnosed with clinical depression at the later stage are limited to analysis based on psychological, physiological genetic and socio-economic factors. For example, common psychological assessments for the prediction of depression in adolescents use methods such as self-reports and depression scale ratings [1]. More recent studies combine medical imagery information by investigating the correlation of brain structure with depression [14]. Although important, these studies are costly, time consuming and do not provide unique and quantitative criteria that can be easily validated or used in a mass screening. Our research into the prediction process started from facial image analysis and aimed to determine if a non-depressed person is likely to develop clinical depression within the next 1-2 years. The prediction method described in [12] used a typical classification approach where class models were determined using image data from adolescents who were "at risk" or "not at risk" of depression. The risk factor was confirmed through 2 years of the follow-up data collection. Two feature extraction methods were compared: the eigenface (PCA) feature and the fisherface (PCA+LDA) feature. The best results were obtained for the fisherface yielding a prediction accuracy of 51% with the person independent approach and 61% with the person dependent approach. These results did

not provide strong support for the use of facial images as predictors of depression; therefore the usefulness of vocal features in predicting depression was examined.

The proposed here approach uses speech as a diagnostic signal and assumes that early depression can be manifested through subtle changes of acoustic characteristics as early as 1-2 years before other more obvious symptoms of depression are developed. By detecting these changes a risk factor for depression can be determined. Firstly, this study aimed to provide a proof of the above concept by showing that acoustic characteristics of speech can be used to detect individuals who are likely to develop depression in the near future. Secondly, this study aimed to investigate the effectiveness of the glottal and the Teager energy features in an early prediction of depression in adolescents. These particular features were previously shown to provide excellent results in the diagnosis of depression [8], [10], therefore it was expected that they can also provide good prediction results. The prediction was achieved through a binary classification of speech recordings from 15 adolescents who developed Major Depression within two years after these recordings were made and 15 adolescents who did not developed Major Depression within the same time period. Our results for the first time confirmed that the acoustic speech analysis can be used to identify adolescents likely to develop depression within the next 1-2 years. The remaining parts are organized as follows. Section 2 describes the speech database. Section 3 explains our methodology. Section 4 describes experiments and results and Section 5 contains the conclusions.

## 2. DATABASE

The speech database suitable for developing and testing of the prediction results was of key importance to this research. The prediction process could be validated only if speech data recorded at a certain point of time was complimented with results of clinical examinations conducted at the time when recordings were made and then successively repeated over a period of time sufficient for the participants to develop clear symptoms of depression. Figure 1 provides a graphical summary of the longitudal data collection process conducted by researchers from the ORYGEN-Youth Health Research Centre in Melbourne, Australia. This database was collected during two stages. In the first stage (T1), audio-visual recordings were made of adolescents (12-13 years of age) accompanied by their parents. The recordings captured images and sounds from naturalistic discussions between parents and children on two different topics of family interactions: event-planning interaction (EPI) and problem-solving interaction (PSI). Each interaction was recorded and annotated based on the Living-In-Family-Environments (LIFE) coding system [7] for a separate 20 minutes session. The total of 191 (94 female & 97 male) adolescents participated in this stage and they were all diagnosed as having no symptoms of depression. Two years after T1, the second (follow-up) stage (T2) was conducted with the same participants. During this stage no video recordings were taken but all adolescent participants were tested by psychologists using conventional methods of diagnosis to determine their current state of mental health. Results showed that 15 adolescents (6 male & 9 female) suffered from the Major Depressive Disorder (MDD) and 3 adolescents (1 male & 2 female) had Other Mood Disorders (OMD). The remaining adolescent had no symptoms of depression or other mood disorders.

Using the diagnostic information collected at T2, speech recordings of adolescent participants made in T1 were divided into two classes: "At Risk" (AR) referring to adolescents who were non-depressed in T1 but developed MDD between T1 and T2, and "Not At Risk" (NAR) representing adolescents who were non-depressed in T1 and did not show any symptoms of depression between T1 and T2. Since only data from 15 AR participants was available, speech data set of 30 participants (15 AR and 15 NAR) were used to match the population's sizes for each class. The AR class included 6 male and 9 female participants. This ratio reflected the well known trend in depression epidemics with almost twice as many females as males likely to develop depression during adolescence [5]. In order to match this gender ratio, the NAR class also included 6 male and 9 female participants.

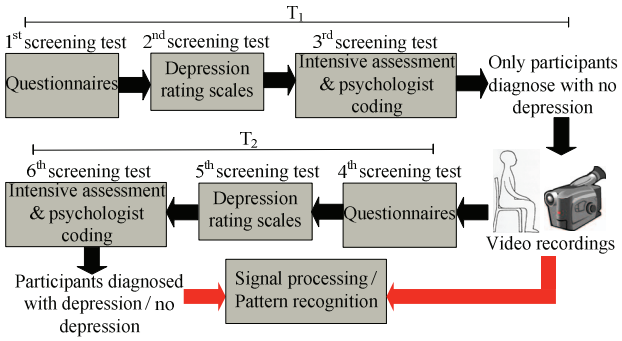


**Figure 1:** Longitudal data collection procedure.

## 3. METHODOLOGY

The experimental framework used in this study was based on the modeling and classification of training and testing datasets. During the supervised training process, training speech samples from known classes were pre-processed and analyzed in order to generate class-characteristic feature parameters. The feature parameters were then used to generate statistical class models. During the testing stage, testing speech samples from unknown classes were passed thought the same pre-processing and feature extraction process as in training stage. The resulting feature vectors were compared by the classifier with the class models, and the probable class to which the tested speech sample was most likely to belong to was determined.

### 3.1. Pre-processing

All speech samples were extracted from the audio-visual recordings database provided by the ORGEN-YH Research centre. The speech samples was processed on a frame-by-frame basis with frame length equal to 25 milliseconds and 50% overlap between frames. Frames that did not contain voiced speech were discarded, whereas frames containing voiced speech were concatenated and used in the subsequent feature extraction process. The voiced/unvoiced detection was based on the linear prediction technique calculated using the speech processing and synthesis toolboxes [4]. The $13^{th}$ order linear prediction coefficients were calculated for each frame where the energy of the prediction error and the first reflection coefficient are calculated with a threshold set to detect voiced samples and discard the unvoiced regions.

### 3.2. Feature extraction

*3.2.1. Glottal (G) features*

Glottal waveform represents speech component generated based on the air flow during the opening and closing of the vocal folds. An Adaptive Inverse Filtering algorithm (IAIF) [3] using a discrete all-pole modeling (DAP) from the TTK Aparat toolbox [2] was applied to generate the glottal wave and its parameters. The glottal timing (GT) was represented by 9 different parameters describing timing of the opening and closing phases of the vocal folds. The glottal frequency (GF) was represented by 3 different frequency domain parameters representing ratios between amplitudes of the harmonic components.

*3.2.2. Teager Energy Operator based features derived from the speech waveform (TEO)*

Teager energy operator (TEO) based feature represents a nonlinear airflow along the vocal tract of speech production which was suggested by Teager [13]. Zhou et al. [16] investigated the use of TEO features and found that the critical band TEO autocorrelation features was most effective in stress classification. Low et al. also successfully tested the TEO features in detection of clinical depression [8]. Following steps described in [8], [16], the TEO-CB-Auto-Env parameters were calculated for speech waveform within 15 critical bands given by the Gabor filters.

*3.2.3. Teager Energy Operator based features derived from the glottal waveform (TEO_G)*

Depression detection studies by Moore et al. [10] and Low et al. [8] pointed to the high effectiveness of glottal features. It was suggested that the high effectiveness was due to the sensitivity of subtle voice changes and the glottal domain is linked to the source of voice production which is the onset of expression. The TEO-CB-Auto-Env features were calculated directly on the glottal waveform providing TEO_G parameters within 15 critical bands. An example showing the processing stages involved in calculating the TEO_G parameters is illustrated in Figure 2.

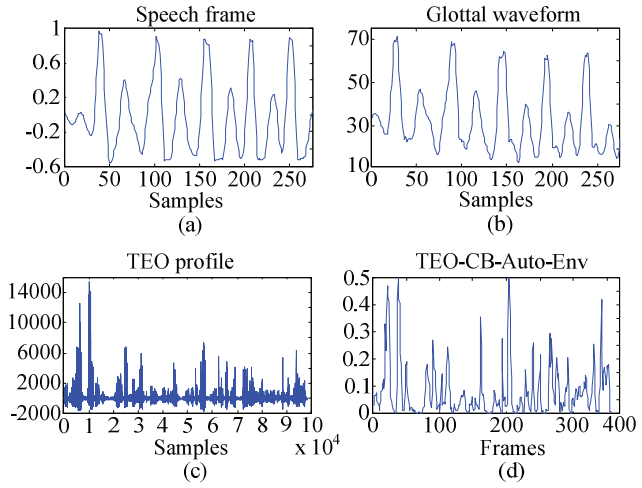


**Figure 2:** TEO_G feature extraction: (a) 25ms speech frame. (b) Glottal flow waveform. (c) TEO profile of the glottal waveform. (d) TEO features derived from glottal waveform.

Recent study by Hansen et al. [6] pointed to the frequency-dependent nature of TEO features in stress classification. We investigate the effectiveness of the TEO_G feature at different frequency levels by analyzing 5 frequency sub-bands, each spanning across 3 critical bands.

*3.2.4. Combined features (Glottal + TEO)*

Combined feature vectors including the glottal features and the TEO features (G+TEO) were used to test the suggestions given in [8], [10] and indicating that addition of glottal features to other types of features can enhance their performance.

### 3.3. Modeling and classification method based on the Gaussian mixture model (GMM)

The Gaussian mixture model from the HTK toolbox [15] was implemented to generate class models and perform the classification tasks into "At Risk" (AR) and "Not At Risk" (NAR) groups. Each Gaussian mixture is trained with 3 states using diagonal covariance matrices in the classification process.

## 4. EXPERIMENTS & RESULTS

Experiments were conducted on the dataset of 30 participants, with 15 participants for each class (AR and NAR). Around 50% of the dataset, i.e. 8 AR participants and 7 NAR participants were used for training and the remaining 50% was used for testing. All of the speech samples came from the Problem Solving Interaction (PSI) which was previously shown to be particularly effective in depression classification [8], [12]. The classification task was person dependent (PD) and aimed to correctly identify the tested adolescents as either AR or NAR where the number of the correctly classified utterances for a given participant was greater than 50%. This predictive classification results were assessed based on three statistical parameters: specificity, sensitivity and accuracy defined as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100\% \quad (1)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100\% \quad (2)$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (3)$$

where, TP=number of true positive outcomes (the number of AR adolescents classified as AR), FP=number of false positive outcomes (the number of NAR adolescents classified as AR), TN=number of true negative outcomes (the number of NAR adolescents classified as NAR), and FN=number of false negative results (the number of AR adolescents classified as NAR). The results were averaged over 3 cross validation runs of the training and testing datasets independently across different features.

The performance summary of the baseline TEO and Glottal (G) features in Table 1 shows that glottal features performed at higher prediction accuracy (69%) when compared to the TEO features (52%). These results lead to further investigation on the effects of glottal features reflected in the application of TEO_G features as well as combination of glottal and TEO features (G+TEO). Results of these investigations are illustrated in Table 2 which show very close performance for both types of features (63-64%) and a significant increase when compared to the baseline TEO feature. However, these approaches did not exceed the 69% of accuracy of the glottal features applied alone. The contribution of different

frequency ranges to the classification accuracy is illustrated in Figure 3, where 15 critical bands (CB) from the TEO_G feature were subdivided into 5 sub-bands (SB) with 3 CB in each SB. It was observed that the higher frequency range (1.3-5.5 kHz) provided better classification accuracy for both the AR and NAR classes. For SB with frequency below 1.3 kHz, the classification accuracy is lower than 60% whereas for SB with frequency above 1.3 kHz, the classification accuracy is higher than 60%.

| Baseline feature classification | | | |
|---|---|---|---|
| Training/Testing Features | Classification Accuracy in % | | |
| | Sensitivity | Specificity | Accuracy |
| G | 76.19 | 62.5 | 69.35 |
| TEO | 43.45 | 60.71 | 52.08 |

**Table 1.** Average classification accuracy for baseline features

| Feature combination classification | | | |
|---|---|---|---|
| Training/Testing Features | Classification Accuracy in % | | |
| | Sensitivity | Specificity | Accuracy |
| TEO_G | 68.45 | 60.71 | 64.58 |
| G + TEO | 60.12 | 65.48 | 62.8 |

**Table 2.** Average classification accuracy for glottal influenced features

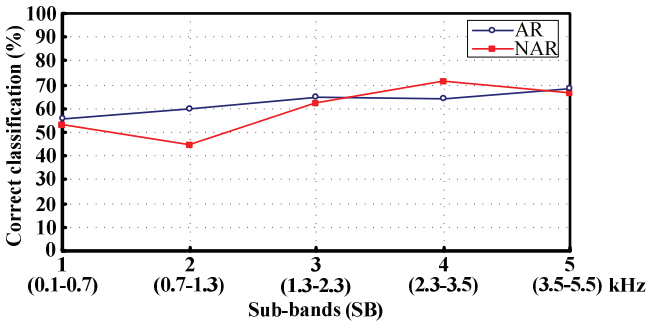


**Figure 3:** Comparison of classification percentage across frequency bands.

## 5. CONCLUSION

The first of its kind automatic prediction of clinical depression in adolescents was presented. Prediction was achieved through a person dependent binary classification of speech samples into two categories: at risk of developing depression within the next 1-2 years depression (AR) and not at risk (NAR). The study lead to the following general conclusions: 1. It was demonstrated that speech analysis can be used to provide early indication of risk for depression (accuracy > 60%) as early as 1 to 2 years before other symptoms detectable using current diagnostic methods can be determined. 2. The results provide support to theories assuming that either a gradual development of symptoms of depression or existence of specific psychophysical characteristics in depression-prone individuals is reflected in their acoustic characteristics of speech. 3. It was observed that the glottal parameters provide the best prediction results with accuracy up to 69%. The authors are currently conducting further experiments validating these conclusions and expanding the prediction methodology into an integrated speech and facial image analysis.

## 6. ACKNOWLEDGEMENTS

The authors would like to thank the ORYGEN Youth Health Research Centre, Australia for their support in this research.

## 7. REFERENCES

[1] M. Aebi, et al., "Prediction of major affective disorders in adolescents by self-report measures," *Journal of Affective Disorders*, vol. 115, pp. 140-149, 2009.

[2] M. Airas, "TKK Aparat: An environment for voice inverse filtering and parameterization," *Logopedics Phoniatrics Vocology*, vol. 33, no. 1, pp. 49-64, 2008.

[3] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2-3, pp. 109-118, 1992.

[4] D. G. Childers, Speech processing and synthesis toolboxes, *Wiley*, New York, Chichester, 2000.

[5] J. Cyranowski, et al., "Adolescent onset of the gender difference in lifetime rates of major depression," *Arch General Psychiatry*, vol. 57, pp. 21-27, 2000.

[6] J. H. L. Hansen, et al., "Robust Emotional Stressed Speech Detection Using Weighted Frequency Subbands," *EURASIP Journal on Advances in Signal Processing*, vol. 2011.

[7] H. Hops, et al. "Living in family environments (LIFE) coding system: Reference manual for coders," *Oregon Research Institute*, Eugene, OR, Unpublished manuscript, 2003.

[8] L. S. Low, et al., "Detection of Clinical Depression in Adolescents' Speech During Family Interactions," *IEEE Transactions, Biomedical Engineering*, vol. 58, no.3, pp.574-586, 2011.

[9] N. C. Maddage, et al., "Video-based detection of the clinical depression in adolescents," in *Engineering in Medicine and Biology Society, EMBC* , pp. 3723-3726, 2009.

[10] E. Moore, et al., "Critical Analysis of the Impact of Glottal Features in the Classification of Clinical Depression in Speech," *IEEE Transactions, Biomedical Engineering*, vol. 55, pp. 96-107, 2008

[11] P. J. Moses, The voice of neurosis, *New York: Grune & Stratton*, 1954.

[12] K. E. B. Ooi, et al., "Prediction of clinical depression in adolescents using facial image analysis," *Image Analysis for Multimedia Interactive Services, WIAMIS*, 2011.

[13] H. Teager, " Some observations on oral air flow during phonation," *IEEE Transaction, Acoustics, Speech and Signal Processing*, vol.28, pp. 599-601, 1980.

[14] M. B. H. Yap, et al., "Interaction of Parenting Experiences and Brain Structure in the Prediction of Depressive Symptoms in Adolescents," *Arch Gen Psychiatry*, vol. 65, pp. 1377-1385, December 1, 2008.

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

[16] G. Zhou, et al., "Nonlinear feature based classification of speech under stress," *IEEE Transactions, Speech and Audio Processing*, vol. 9, pp. 201-216, 2001.