CONTENT BASED CLINICAL DEPRESSION DETECTION IN ADOLESCENTS

Lu-Shih Alex Low, Namunu C. Maddage, Margaret Lech, Lisa Sheeber [1], Nicholas Allen [2]

School of Electrical and Computer Engineering, RMIT University, Melbourne 3001, Australia
[1]Oregon Research Institute, 1715 Franklin Boulevard, Eugene, Oregon 97403
[2]ORYGEN Research Centre and Department of Psychology, University of Melbourne, Melbourne 3010, Australia
lushih.low@student.rmit.edu.au, {namunu.maddage, margaret.lech}@rmit.edu.au, lsheeber@ori.org, nba@unimelb.edu.au

ABSTRACT

This paper studies the effectiveness of speech contents for detecting clinical depression in adolescents. We also evaluated the performances of acoustic features such as Mel frequency cepstral coefficients (MFCC), short time energy (Energy), zero crossing rate (ZCR) and Teager energy operator (TEO) using Gaussian mixture models for depression detection. A clinical data set of speech from 139 adolescents, including 68 (49 girls and 19 boys) diagnosed as clinically depressed, was used in the classification experiments. Each subject participated in three 20 minutes interactions. The classification was first performed using the whole data and a smaller sub-set of data selected based on behavioural constructs defined by trained human observers (data with constructs). In the experiments, we found that the MFCC+Energy feature out performed the TEO feature. The results indicated that using the construct based speech contents in the problem solving interactions (PSI) session improved the detection accuracy. Accuracy was further improved by 4% when the gender dependent depression modelling technique was adopted. By using construct based PSI session speech content, gender based depression models achieved 65.1% average detection accuracy. Also, for both types of features (TEO and MFCC), the correct classification rates were higher for female speakers than for male speakers.

1. INTRODUCTION

The inability to diagnose clinical depression early on in adolescents aged 13-20 years, can have a serious impact on suffers, including the risk for suicidal ideation. Strong evidence demonstrates that most suicides are linked to depressive disorders and symptomatology [13].Teen suicide has become a significant public health concern, seeing as how it is one of the leading causes of death in Australia. Suicide rates among Australian adolescents have increased threefold from the 1960s to the 1990s. Although recent statistics (2006) have shown a dip in the number of youth suicides, it still ranks suicide as the 15th leading cause of death in Australia [1]. Depressed individuals suffer from varying degrees of psychomotor retardation (slowness) or agitation. Sufferers of depression experience prolonged periods of hopelessness, anger, guilt, desperation and loneliness along with, as noted above, a tendency to suicidal thoughts. Dealing with the issue of depression poses a complex and challenging task due to the many potential psychological variables. In an effort to understand and prevent depression and suicide in adolescents, psychologists have carried out studies based on demographic profiles, family self-reports and observational data of a patient in clinical interviews. From these interviews, it has been consistently reported that clinicians observe that the speech of a depressed patient is slow, uniform, monotonous and expressionless with the patient having the fear of expressing him or herself [12]. During listening tests [4], listeners could perceive differences in pitch, loudness, speaking rate and articulation of speech recorded from depressed patients before and after treatment. This has led to considerable amount of interest in combining psychological assessments with acoustic speech analysis to objectively measure behavioural changes in a patient over time. Any improvement in objective diagnosis would translate into relevant clinical applications including the early detection of depressive condition and the evaluation of treatment outcome. Therefore, this is the basis of our research. In turn, this could lead to the possibility of developing a computerized healthcare system that would assist mental health professionals by providing early warning-signs indicating whether a patient is likely to be depressed through their voice patterns.

As early as the 19th century, attempts have been made to analyse vocal acoustic parameters by finding potential indicators of depression [16]. Since then, there have been numerous efforts to empirically determine the physical and mental health of individuals through their vocal speech patterns. In fact, designing an automatic computerized system for depressive screening in speech is not a novel idea [21]. Currently however, there has not been any computerized vocal diagnosis tools that can provide accurate results to assist psychologists in detecting clinical depression among adolescents. The most commonly studied parameters in speaker characterization pertinent to the literature have been the measures relating to prosodic (i.e. Fundamental frequency (F0), speaking rate, energy) and the vocal tract (i.e. formants) [6], [11], [20], [14], [5], [15]. This is due to the fact that they have the closest relation to human perception [11]. Unfortunately, to make issues more complicated, there have been discrepancies in results presented from one researcher to another. Although most researchers [20], [14] found that F0 correlated well with depression, France *et al.* (2000) [5] experiments on severely depressed and near-term suicidal subjects with gender separation found that F0 was an ineffective discriminator for both depressed male and female patients. Instead, formants and power spectral density measurements proved to be the better discriminators. This could be due to the many different variables such as recording conditions, number of participants and the level of participant's depression ratings.

Performing multivariate analyses on vocal features extracted from a patient's speech has been the main focus in recent studies in order to increase the accuracy of classification in clinical depression [5], [15], [11]. Highest classification accuracy achieved up to date has been presented by Moore *et al.* (2008) [11]. On a sample data size of 33 subjects (15 major depress, 18 controls), Moore adopted a feature selection strategy by adding one feature at a time to find the highest classification accuracy through quadratic discriminant analysis and obtained a classification accuracy of 90% and 96% with the combination of prosodic and glottal features for male and female patients respectively. However, the sample data size for the experiments conducted may be deemed too small for creating statistically significant results for clinical application.

In this paper, we tested the use of features such as mel frequency cepstral coefficients (MFCC), energy, zero-crossing, delta MFCCs and delta-delta- MFCCs in automatic detection of clinical depression in adolescents. Although MFCCs have been widely used in speaker recognition, limited studies have been performed in the area of stress or clinical depression classification. It was shown that stress classification using these features provided better classification accuracy in text-dependent models [22]. For comparison purposes, Teager energy operators (TEO) which has shown good accuracy and reliability in emotional stress classification is also explored.

2. ANALYSIS METHOD

Figure 1 depicts the proposed framework that models the contents of depressed and control subjects. For both training and testing phases, we first detect the voiced frames in the pre-processing stage. Secondly, MFCC feature coefficients are extracted from these voiced frames. In the training phase, two Gaussian mixture models (GMMs) are then trained using the extracted features belonging to depressed and control subjects.

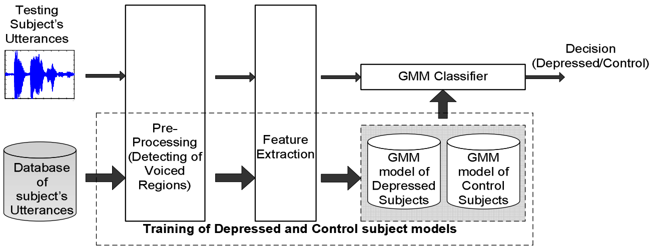
Testing Subject's Utterances
Database of subject's Utterances
Pre-Processing (Detecting of Voiced Regions)
Feature Extraction
GMM Classifier
Decision (Depressed/Control)
GMM model of Depressed Subjects
GMM model of Control Subjects
Training of Depressed and Control subject models

Figure 1: Modelling of depressed and control subjects.

In the testing phase, the decision on whether the subject is depressed or control is taken when more than 50% of utterances from the subject belongs to that particular class. The following subsections explain the pre-processing, feature extraction and depressed/control content modelling steps.

2.1 Pre-Processing

We use linear prediction based technique explained in [3] to detect voiced regions. First, the speech signal is normalized based on the maximum amplitude and then segmented into 25msec, 50% overlapped frames using a rectangular window. Then 13th order linear prediction coefficients (LPCs) are calculated per frame. Energy of the prediction error and the first reflection coefficient r_1 are calculated and a threshold is empirically set to detect voiced frames.

$$r_1 = \frac{\frac{1}{N}\sum_{n=1}^{N-1} s(n)s(n+1)}{\frac{1}{N}\sum_{n=1}^{N} s(n)s(n)} \qquad (1)$$

Eq. (1) explains the calculation of the first reflection coefficient r_1, where N is the number of samples in the analysis frame and s(n) is the speech sample. Silence frames which are considered as unvoiced, are also removed in the pre-processing. Detected voiced frames are then concatenated for feature extraction.

2.2 Feature Extraction

Mel frequency cepstral coefficients (MFCC) which have been effectively used for speech content characterization are based on the linear speech production models which assume that airflow propagates in the vocal tract as a linear plane wave. According to Teager [19] on the other hand, this assumption may not hold since the true source of sound production is actually coming from the vortex-flow interactions that are non-linear. For example, in stress speech, the fast air flow causes vortices to be located near the false vocal fold area and this provides additional excitation signals other than pitch [22], [19]. Thus, we further investigate both MFCCs and TEOs strength for modelling clinical depression. The following subsections briefly discuss the computation of MFCC and TEO.

2.2.1 Mel Frequency Cepstral Coefficients

The mel frequency cepstral coefficients have been widely used in speech processing. A mel is a unit of measure of perceived pitch. It does not correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive the pitch in a linear manner. The relationship between linear frequency scale (F_{linear}) and the log frequency scale (F_{log}) is explained in Eq. (2) where C is a scaling factor which controls the span of Log_{10} frequency scale. When C is set to 700, the log frequency scale then becomes mel scale or ($F_{log} = F_{mel}$).

$$F_{\log} = \frac{C \log_{10}}{\log_{10} 2}[1 + \frac{F_{linear}}{C}] \qquad (2)$$

To calculate the positions of filters to cover spectral regions, first filters are linearly positioned in the mel frequency scale and transformed back to linear frequency scale. Using this filter bank we compute the MFCCs. The output Y(i) of the ith filter in the linear frequency scale is defined in Eq. (3), where S(.) is the signal spectrum, H_i(.) is the ith filter, and m_i and n_i are boundaries of the ith filter.

$$Y(i) = \sum_{j=m_i}^{n_i} \log_{10}[S(j)]H_i(j) \qquad (3)$$

Eq. (4) describes the computation of nth MFCC, k_i is the center frequency of the ith filter, and N and N_{cb} are number of frequency sample points and number of filters, respectively.

$$C(n) = \frac{2}{N}\sum_{i=1}^{N_{cb}} Y(i)\cos(k_i \frac{2\pi}{N} n) \qquad (4)$$

2.2.2 Delta and Delta-Delta-MFCC

Another popular feature arrangement which can capture the temporal information is the inclusion of the first and second order derivatives (Delta and Delta-Delta) of MFCCs combined with the original MFCCs. The Delta and the Delta-Delta MFCCs are calculated using the following formula:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \qquad (5)$$

where d_t is a delta coefficient at time t and it is computed in terms of the corresponding static coefficients from $c_{t-\Theta}$ to $c_{t+\Theta}$. The Θ is the size of the window which provides information about spectral changes in the neighbouring frames respective to current frame c_t. The same formula is applied to the Delta coefficients to obtain the Delta-Delta coefficients. For our experiments, the window size is set Θ=2 to obtain both Delta and Delta-Delta coefficients.

2.2.3 Energy and Zero-Crossing

The logarithmic of short-term energy E_s (m) and the number of zero crossings Z_s (number of times the signal sequence changes sign), within a frame are also important features in speech analysis which can be easily obtained using the following two formulas:

$$E_s(m) = \log \sum_{n=m-N+1}^{m} s^2(n) \qquad (6)$$

$$Z_s = \ell\{\frac{|\operatorname{sgn}\{s(n)\} - \operatorname{sgn}\{s(n-1)\}|}{2} \; where, \operatorname{sgn}\{s(n)\} \begin{cases} +1, & s(n)\geq 0 \\ -1, & s(n)<0 \end{cases} \qquad (7)$$

Hence, we add both short-term log energy and zero-crossing as extra features to MFCCs.

2.3.1 Teager Energy Operator (TEO)

Modelling the time-varying vortex flow is a challenging task and Teager [19] devised a simple algorithm which uses a non-linear energy-tracking operator called the Teager energy operator (TEO). The Teager energy operator (TEO) in a discrete form [9] is defined as

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1) \qquad (8)$$

where Ψ[.] is the Teager energy operator (TEO) and x(n) is the sampled speech signal.

2.3.2 TEO Critical Band Based Autocorrelation Envelope (TEO-CB-Auto-Env)

As proposed by Zhou [22], it is more convenient to break the bandwidth of the speech spectrum into smaller bands (also known as critical bands) before calculating the TEO profile (Eq.8) for each independent band. Gabor bandpass filter [10] as shown in Figure 2(b) is implemented to separate voiced utterances into 16 critical bands. The TEO profile is then calculated for each of the bands. The Gabor-filtered TEO stream is then segmented into frames. The autocorrelation of the TEO output is computed and the area under the normalized autocorrelation envelope is calculated to give the *TEO-CB-Auto-Env* features. We follow the same frequency range for the 16 critical bands as in [22]. As an example, Figure 2(c) shows the TEO profile waveform from critical band 9 (between 1080 Hz -1270 Hz) and figure 2(d) shows the normalized autocorrelation of one frame from the TEO profile output waveform.

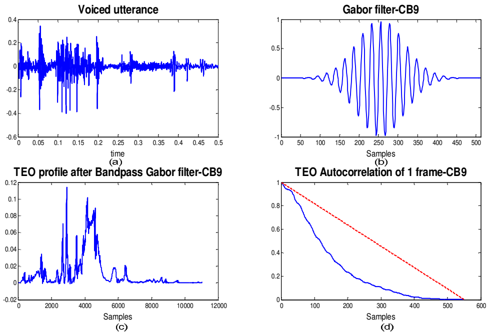

Figure 2: TEO-Auto-Env feature extraction for Critical Band 9. (a) Voiced part of utterance (b) Gabor band-pass filter (c) TEO profile after Gabor band-pass filter (d) TEO autocorrelation of one frame.

2.3 Gaussian Mixture Model (GMM)

Gaussian mixture model has been effectively used in speech information modelling tasks such as speaker recognition and spoken languages identification. A Gaussian mixture model is a simple linear superposition of Gaussian components, aimed at providing a richer class of density models than a single Gaussian as given in Eq. (9), where x is a D- dimensional random vector (i.e. $X=\{x_1,x_2 \ldots x_N\}$) and K is the number of Gaussian densities. Each Gaussian density $N(x|\mu_k, \sum_k)$ is called a component of the mixture and has its own mean μ_k and covariance $\sum_k$.

$$p(x) = \sum_{k=1}^{K} \pi_k N(x \mid u_k, \Sigma_k) \qquad (9)$$

The parameters π_k are called mixing coefficients. If we integrate both sides of Eq. (9) with respect to x, and note that both p(x) and the individual Gaussian components are normalized, we will obtain the following expression:

$$\sum_{k=1}^{K} \pi_k = 1 \qquad (10)$$

In the training process, the maximum likelihood (ML) estimation used by the GMM shown in Figure 1 is used to compute model parameters which maximize the likelihood of GMM using expectation-maximization algorithm for the given training data. Each model is trained with 3 iterations every time the number of Gaussian mixture components is increased. Diagonal covariance matrices are used instead of full covariance for computational efficiency purposes. We use HTK toolbox[1] to implement the Gaussian mixture models in the classification of clinical depression.

3. EXPERIMENTS & RESULTS

The database described here was obtained as a result of collaboration with the Oregon Research Institute, USA (ORI). This database consists of video and audio recordings of 139 adolescents (93 girls and 46 boys) with their respective parents participating in three different types of 20-min interactions: 1) Event planning interaction (EPI), 2) Problem-solving interaction (PSI) and 3) Family consensus interaction (FCI). Through self-report and interview measures of the adolescence's depression evaluated by research staff from ORI [17], 68 (49 girls and 19 boys) were diagnosed as suffering from Major Depressive Disorder (MDD), and the remainder (44 girls and 27 boys) were healthy controls (i.e., no current or lifetime history of MDD). The depressed and healthy groups were matched on their demographic data which included their sex, race and age. The adolescents were between 12 and 19 years old. As will be discussed in the later part of this section, the PSI, which shows higher separation results over the other interactions or combined interactions, is described here in detail. A more detailed description on the other interactions can be found at [7]. Upon starting the interaction for the PSIs, parents and adolescents mutually agreed on two topics of disagreement that were completed from a questionnaire. Each family unit was then asked to discuss the problems, one at a time, and try to come to some resolution that was mutually agreeable to all parties. Each problem was discussed for 10 minutes, resulting in a total of 20 minutes of observational data for each family. The video recordings were manually coded by psychologists using the Living-In-Family-Environments (LIFE) coding system. The LIFE coding system [7], [8] is an event-based coding system recorded in real-time and was developed to accurately capture specific behaviours representative of depressed individuals and that would effectively discriminate between their behaviour and those of non depressed individuals. It is designed to annotate the specific timeline of various emotions (called affect codes) and verbal components (called content codes) displayed by the subject's speech during the whole course of the interaction. The LIFE code is composed of 27 content codes and 10 affect codes. The audio segments displaying these emotions and verbal content are extracted from the video recordings based on the time annotations. The average speech utterance extracted is around 2 to 3 seconds long and its original sampling rate is decimated by a factor of 4. This is done by first using an anti-aliasing filter, followed by downsampling the audio signal from 44.1 kHz to 11 kHz sampling rate. For our experiments, the audio was converted from stereo to mono channel format.

The speech corpus was extracted in two alternative ways:

- Annotating the separate interactions and combined interactions without using any behavioural constructs
- Annotating the separate interactions and combined interaction using behavioural constructs

In the first annotation method, interactions prepared without using any constructs can be described as low-level coding. This is where

[1] http://htk.eng.cam.ac.uk/

all content codes with their respective affect codes are extracted. Conversely, for the second annotation method, interactions prepared based on behavioural constructs can be described as high-level coding. The constructs methodology was developed by behavioural researchers from ORI [8]. The constructs were created by pairing up selected content codes and affect codes into four main classes that represent *aggressive, depressive, facilitative*, and *problem solving* behaviours. The following describes the experimental setup in modelling adolescents' depression.

3.1 Feature Selection

It was found in the literature that the Mel frequency cepstral coefficient (MFCC) and Teager Energy Operator (TEO) features performed well characterizing depressed speech contents. Thus, its strengths were also examined in our experiments. Due to high performances in modelling speech contents, GMM were employed for modelling depressed and control classes. First, optimization of the parameters in the MFCC was carried out to maximise the detection accuracy of depressed and control subjects. The correct classification accuracy was plotted as a function of 1) number of filters (from 10 to 60 with a step size of 10; and 2) number of coefficients (from 6 to 30 with a step size of 6). Based on these plots we selected the number of filters and the number of coefficients that were giving the highest value of classification accuracy. For the MFCC parameter optimization, half of the utterances per subject (10 min of speech) were used and divided equally for training 1024 Gaussian mixtures by iteratively increasing the number of Gaussian mixture components by a factor of two (i.e. 2, 4, 8, 16, 32) for depressed and control classes. It was found that 30 filters in the filter bank and 12 coefficients maximized the correct classification accuracy. Optimized MFCC parameters and full speech corpus were then used in further experiments. Around 50% of the total number of subjects containing 33 depressed (23 girls and 10 boys) and 34 control subjects (21 girls and 13 boys) were used as our testing data. Subject utterances were tested based on the length of the utterances extracted from the behavioural construct coding. Additional features such as short-term energy and zero-crossing rates were then added to observe if they would introduce additional improvements in the correct classification rates. Although it was found that adding zero-crossing to MFCCs (with energy included) did not improve results, adding short-term energy alone improved the classification rate by 2%. Incorporating velocity (delta) and acceleration (delta-delta) coefficients to the original 13 coefficients (MFCCs+Energy) also improved the correct classification accuracy by around 3%. Thus we combined both the MFCC and Energy features and computed 39 coefficients per feature vector by including both velocity and acceleration coefficients.

3.2 Content based depression detection

Experiments based on increasing the number of Gaussians by a factor of two was carried out using data from separate interactions (EPI, PSI and FCI) and combined interactions (EPI+PSI+FCI) without and with behavioural construct annotations. Table 1 summarizes the average classification accuracy (subject level) which was cross validated using 4 turns of training and testing data sets for depressed and control classes. Results in the first row (EXP 1) explain the rate of correct classification accuracies for all combined interactions (EPI+PSI+FCI) when gender information is not taken into consideration at modelling level, whereas the results in EXP 2, were based on the gender based models. EXP 3 and EXP 4 presents the same format as described in EXP1 and EXP 2 but this time results were based on each individual session (EPI, PSI, and FCI). EXP 1 to EXP 4 didn't select the utterances based on the behavioural constructs for training and testing. However, results in both EXP 5 and EXP 6 were based on the behavioural construct annotated utterances. It was initially expected that higher classification rates should be given when using the data without behavioural constructs as it contained all the affect and content codes which reflects detailed information about the mental status of the subjects. However, it was found that the behavioural construct-based reduction of the data using problem solving interaction (PSI) produced higher classification rates (see EXP 5 and EXP 6), which suggests the importance in using only certain emotion based speech contents in detecting depression. This is consistent with psychologist evaluations in [17], that calculate single statistical measures such as mean, variance and z-scores on the timing durations of PSI of each subject using the behavioural construct based methodology. This finding is predictable given that the PSI is the interaction task that is most likely to elicit conflictual interactions, and that many previous studies have shown that levels of conflict in family environments are a particularly strong correlate of adolescent's depression [18]. As such it is not unexpected that the interaction task that elicits the most conlflictual interactions is the one that most accurately distinguishes the depressed and control participants.

Table 1: Correct classification rates using MFCC + energy for different interactions and combined interactions.

EXP No	Gender	Constructs	EPI	PSI	FCI
1	independent	No	56.97		
2	dependent	No	60.69		
3	independent	No	57.96	55.8	50.66
4	dependent	No	56.97	55.86	60.10
5	independent	Yes	48.93	59.88	56.29
6	dependent	Yes	54.81	62.11	54.41

From the results in EXP 5 and EXP 6, we found ~4% accuracy improvement with the gender based depression modelling in PSI session. This clearly indicates presence of gender oriented differences in the complexity of depression symptoms which is consistent with conclusions presented in [2] & [17]. Since PSI was giving relatively higher results for EXP5 and EXP6, the two experiments were re-simulated but this time we iteratively increased the number of Gaussian mixtures sequentially ranging from 1 to 200. This was done to see if further improvement in classification accuracy could be achieved. As shown in table 2, by sequentially increasing the number of Gaussian mixtures, it gave a ~5% increase in average classification rate for the gender dependent model as compared to EXP6. In both of Tables 2&3, adolescent females performed better in detecting depression over their male counterparts when using MFCC+Energy features and TEO based features, with an improve accuracy of ~5% and ~17% respectively. This could be due to the fact that psychologists have observed that female adolescents experience substantially higher levels of depressive symptoms than do males [2].

Table 2: Correct classification results – MFCC + Energy features.

Problem Solving Interaction (PSI) with constructs				
Testing dataset	**Training dataset**	**Number of Gaussian Mixtures**	**Classification accuracy in %**	
			Depressed	**Control**
Male + Female subjects	Male +Female subjects	19	60.24	59.64
Male subjects only	Male subjects only	182	61.67	67.99
Female subjects only	Female subjects only	27	64.75	65.91

Due to previously reported higher performances, the performances of MFCC+Energy feature setup with TEO based feature were compared. Table 3 show the classification accuracies when using the TEO based feature as described in Section 2.3.1. The TEO based feature was only applied to the data with behavioural constructs since as it was shown in Tables 1&2 that the behavioural construct data yields in general higher classification rates. When comparing Tables 2&3, it can be seen that in both gender independent and gender dependent models, the TEO-based feature provided slightly lower classification results when compared with the MFCC+Energy feature setup. In the case of female subjects the MFCC+Energy feature provided 4-5% higher accuracy, and in the case of male subjects, it gave up to 10% higher accuracy. However, in the gender independent case, the difference between TEO and MFCC+Energy performance was very small (1%).

Table 3: Correct classification results - TEO based features.

Problem Solving Interaction (PSI) with constructs				
Testing dataset	**Training dataset**	**Number of Gaussian Mixtures**	**Classification accuracy in %**	
			Depressed	**Control**
Male + Female subjects	Male +Female subjects	128	58.72	60.38
Male subjects only	Male subjects only	94	51.94	56.59
Female subjects only	Female subjects only	32	60.71	60.23

The reason for the comparatively lower performance in the TEO is due to the fact that the speech utterances are heavily corrupted with crosstalk of parents, resulting in errors in calculation of the vortex flow of subject's speech.

4. DISCUSSION & CONCLUSION

In this paper we examine the effectiveness of speech contents of different subject interactions and acoustic features for clinical depression detection. The MFCC + short time energy coefficients with their velocity and acceleration coefficients out performed the Teager energy operator (TEO) feature in the detection. The content specific depression detection experiments indicated that selecting the behavioural construct based speech contents from the problem solving interactions (PSI) session instead of event planning interaction (EPI) and family consensus interaction (FCI) gave higher detection rates. Average detection accuracy further improved by 4% when gender based depression modelling technique is adopted for the behavioural construct based PSI speech contents. We achieved 65.1% average detection accuracy using the gender based depression models trained using the MFCC+ energy coefficients with their velocity and acceleration coefficients extracted from the behavioural construct based PSI speech content.

We observed that subject's speech utterances are heavily corrupted by the crosstalk's of parents, which we believe is the main reason for the poor performances of both the features and the GMM. Thus in the future experiments, we plan to implement techniques to suppress the parent's crosstalk and enhance the subjects speech in the utterances. In the future, we will also, perform classification experiments using longer speech utterances (i.e. 30 or 60 seconds) and test larger variety of acoustic features and modelling techniques.

5. ACKNOWLEDGMENT

This research was supported by the Australian Research Council Linkage Grant LP0776235. The authors would like to thank the ORYGEN Research Centre, Australia, the Oregon Research Institute USA for their support in this research.

REFERENCES

[1] Australian Bureau of Statistics, "*Causes of Death, Australia, 2006,"* ABS catalogue number 3303.0, 2008.

[2] W. R. Avison and D. D. McAlpine, "Gender differences in symptoms of depression among adolescents," *Journal of Health and Social Behavior*, vol. 33, pp. 77-96, 1992.

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

[4] J. K. Darby and H. Hollien, "Vocal and speech patterns of depressive patients," *Folia Phoniatrica*, vol. 29, pp. 279-291, 1977.

[5] D. J. France, et al. "Acoustical properties of speech as indicators of depression and suicidal risk," *IEEE Transactions, Biomedical Engineering*, vol. 47, pp. 829-837, 2000.

[6] W. A. Hargreaves and J. A. Starkweather, "Voice quality changes in depression," *Language and Speech*, vol. 7, pp. 84-88, 1964.

[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] H. Hops, B. Davis, and N. Longoria, "Methodological issues in direct observation-illustrations with the living in familial environments (LIFE) coding system," *Journal of Clinical Child Psychology*, vol. 24, pp. 193-203, 1995.

[9] J.F. Kaiser, "On a simple algorithm to calculate the `energy' of a signal," *in Proc. Int. Conf. Acoustic, Speech, Signal Processing '90*, Albuquerque, NM, USA pp. 381-384, 1990

[10] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions, Signal Processing,* vol. 41, pp. 3024-3051, 1993

[11] 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.

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

[13] N.H.M.R.C, "Depression in young people: a guide for mental health professionals," *National Health and Medical Research Council*, Canberra, Australia, 1997.

[14] A. Nilsonne, et al. "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," *The Journal of the Acoustical Society of America*, vol. 83, pp. 716-728, 1988.

[15] A. Ozdas, et al. "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk," *IEEE Transactions, Biomedical Engineering*, vol. 51, pp. 1530-1540, 2004.

[16] E.W. Scripture, "A study of emotions by speech transcription," Vox 31, 179-183, 1921

[17] L. Sheeber, et.al. "Interactional processes in families with depressed and non-depressed adolescents: reinforcement of depressive behaviour," *Behaviour Research and Therapy*, vol. 36, pp. 417-427, 1998.

[18] L. Sheerber, H. Hops, B. Davis, " Family processes in adolescent depression," *Clin Child Fam Psychol Rev,* vol. *4*(1), 19-35, 2001

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

[20] F. Tolkmitt, et al. "Vocal indicators of psychiatric-treatment effects in depressives and schizophrenics," *Journal of Communication Disorders*, vol. 15, pp. 209-222, 1982.

[21] F. J. Tolkmitt, , K. R. Scherer, "Effect of experimentally induced stress on vocal parameters," *J. Experimental Psychology: Human Perception and Performance*, vol 12(3) pp. 302-313, Aug, 1986

[22] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions, Speech and Audio Processing,* vol. 9, pp. 201-216, 20