Published in final edited form as:
IEEE Trans Biomed Eng. 2011 March ; 58(3): . doi:10.1109/TBME.2010.2091640.

Detection of Clinical Depression in Adolescents' Speech During Family Interactions

Lu-Shih Alex Low[*], **Namunu C. Maddage**,
School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology, Vic. 3001, Australia (namunu.maddage@rmit.edu.au; margaret.lech@rmit.edu.au)

Margaret Lech,
School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology, Vic. 3001, Australia (namunu.maddage@rmit.edu.au; margaret.lech@rmit.edu.au)

Lisa B. Sheeber, and
Oregon Research Institute, Eugene, OR 97403 USA (lsheeber@ori.org)

Nicholas B. Allen
Orygen Youth Health Research Centre and the Department of Psychological Sciences, University of Melbourne, Vic. 3010, Australia (nba@unimelb.edu.au)

Abstract

The properties of acoustic speech have previously been investigated as possible cues for depression in adults. However, these studies were restricted to small populations of patients and the speech recordings were made during patients' clinical interviews or fixed-text reading sessions. Symptoms of depression often first appear during adolescence at a time when the voice is changing, in both males and females, suggesting that specific studies of these phenomena in adolescent populations are warranted. This study investigated acoustic correlates of depression in a large sample of 139 adolescents (68 clinically depressed and 71 controls). Speech recordings were made during naturalistic interactions between adolescents and their parents. Prosodic, cepstral, spectral, and glottal features, as well as features derived from the Teager energy operator (TEO), were tested within a binary classification framework. Strong gender differences in classification accuracy were observed. The TEO-based features clearly outperformed all other features and feature combinations, providing classification accuracy ranging between 81%–87% for males and 72%–79% for females. Close, but slightly less accurate, results were obtained by combining glottal features with prosodic and spectral features (67%–69% for males and 70%–75% for females). These findings indicate the importance of nonlinear mechanisms associated with the glottal flow formation as cues for clinical depression.

Keywords

Acoustic features; adolescents; clinical depression classification; naturalistic speech

I. Introduction

THE increase in the prevalence of clinical depression in adolescents (i.e., those aged 13–20 years) has been linked to a range of serious outcomes, particularly an increase in the number

[*]*corresponding author*. School of Electrical and Computer Engineering, Royal Melbourne Institute of Technology, Vic. 3001, Australia (lushih.low@student.rmit.edu.au).

of suicide attempts and deaths, making it a public health concern [30], [54]. Therefore, the early detection of depression in adolescence is of primary importance given the fact that there has been a dramatic increase in the incidence of depressive symptoms and disorders in adolescents in recent years [35]. Clinical depression belongs to the group of affective (mood) disorders in which emotional disturbances consist of prolonged periods of excessive sadness marked by reduced emotional expression and physical drive [9]. From a psychological point of view, one of the tell-tale signs of a person being depressed is the way emotions are expressed in his/her speech. This is based on the assumption that the emotional state of a person suffering from a depressive disorder affects the acoustic qualities of their speech, and therefore, depression could be detected through an analysis of perceived changes in the acoustical properties of speech. Due to the significant differences between adult and adolescent speech [20], this study presents an initial attempt to investigate the acoustic correlates of depression in speech of adolescents. An acoustic analysis of speech will provide clinicians with an additional quantitative measure to compliment and strengthen the current diagnostic techniques. Specifically, an automatic, computer-based analysis of speech, indicating the probability of depression that can be used as a mass-screening device, followed by more detailed (and more resource intensive) interview-based clinical diagnosis of depression.

The remaining part of this paper is organized as follows. Section II contains a brief review of existing methods. Section III describes the database formulation and annotation. In Section IV, the detailed description of our methods can be found. The experiments and results are described in Section V followed by the discussion and conclusions in Section VI.

II. Previous Work

The physiological rational for our research is based on a number of studies [48], [49] that gathered a considerable amount of evidence that emotional arousal produces changes in the respiratory, phonatory, and articulatory processes of speech production. Since depression and suicidality manifest themselves through significant emotional changes, these studies are closely related to the studies of emotion recognition in speech. Depressed speech has been consistently characterized by clinicians as dull, monotone, monoloud, lifeless, and “metallic” [43]. Utilizing a subjective assessment, Darby and Hollien [13] conducted a pilot study of severely depressed patients and found that listeners could perceive noticeable differences in prosodic characteristics of speech, such as pitch, loudness, speaking rate, and articulation in depressed patients before and after treatment. Thus, this observation led to numerous studies of objective measurements using speech parameters that reflect these prosodic characteristics, which included fundamental frequency (F0), formants, jitter, shimmer, intensity of the speech signal, and speech rate. Other commonly used speech parameters have been cepstral features [i.e., mel frequency cepstral coefficients (MFCCs)], spectral features [i.e., power spectral density (PSD)], and glottal features. Most of these acoustic speech parameters have already been identified as possible cues to depression [5], [13], [16], [18], [32], [41], [44], [46] in adults and possibly between other life stages. Another useful feature for stress recognition is the Teager energy operator (TEO), which measures the number of additional harmonics due to the nonlinear air flow in the vocal tract that produces the speech signal. However, the complex relationship between emotional stress and clinical depression still remains somewhat unclear [33].

The level of correlation built to recognize complex relationships between speech parameters and depression has customarily been assessed using multivariate analyses [18], [41], [46]. Ozdas *et al.* [46] trained a multivariate maximum-likelihood classifier with combined features calculated from vocal jitter and glottal spectral slope in voice samples of 30 subjects and achieved an overall classification accuracy of 90% between depressed and control patients. Moore *et al.* [41] with a sample data size of 33 subjects, adopted a feature selection strategy by adding one feature at a time to find the highest classification accuracy through quadratic discriminant analysis and concluded that the influence of glottal features were important discriminating factors in improving the detection of clinical depression.

Early studies of acoustic correlates in speech were limited to small databases (very few participants and short audio recordings). To make issues more complicated, there have been discrepancies in the results presented from one study to another. Therefore, further research validating the proposed measures with larger sample sizes is necessary. Furthermore, no specific studies addressing acoustic parameters of speech as indicators of depression in adolescents have been published.

The main focus of our study was to determine the most important acoustic features that can distinguish the speech of nondepressed from the speech of clinically depressed adolescents in a naturalistic environment (i.e., interactions between adolescents and their parents) rather than speech from interviews or a standard reading task. Apart from using traditional features described in many previous studies, additional features derived from the TEO, which have been reported to be successful in stress and emotion recognition [56] were also tested. We also wanted to examine the role of some previously reported phenomena in our data, such as the well-established gender differences in depressive symptoms during early adolescence [45], and the importance of glottal features in the detection of clinical depression [41]. Additionally, we wanted to determine the optimal duration of the speech samples to be analyzed, due to the fact that in past research, there have been variations in the duration, i.e., 20 s in [18] and 30 s in [46].

III. Speech Database Formulation and Annotation

The database obtained from the Oregon Research Institute (ORI), consisted of video and audio recordings of 139 adolescents (93 females and 46 males), with their respective parents participating in three different types of family interactions. It should be noted that no siblings were included in the family interactions. Each of the three interactions was conducted for 20 min, resulting in a total of 60 min of observational data (video recordings) for each family. A brief description of these interactions is provided in the following and a more detailed description can be found in [21] and [22].

1. *Event-planning interaction (EPI):* The family plans a vacation together and reminisces about a fun time they spent together in the past.
2. *Problem-solving interaction (PSI):* The family tries to resolve two topics of disagreement, identified based on a questionnaire completed by the adolescent and parents.
3. *Family consensus interaction (FCI):* This family discussion involves planning the writing of book chapters on the experience of growing up/raising a child that reflects the shared perspective of both the adolescents and their parents.

Adolescents were excluded from this study if they evidenced any substance dependence or conduct disorders or if they were taking any medications that affect the cardiac system. These exclusion criteria were relevant to the collection of cardio-vascular data that is not used in the current report. Based on adolescent interview data [51], ORI research staff determined that 68 adolescents (49 females and 19 males) met the Diagnostic and Statistical Manual of Mental Disorders version IV (DSM-IV) [1] criteria for a current episode of major depressive disorder (MDD). The remaining 71 participants (44 females and 27 males) were healthy, nondepressed (control) adolescents, who did not meet diagnostic criteria for any current psychiatric disorders and had no history of mental health treatment. The adolescents were between 14 and 18 years. For the purposes of the larger study, from which data for this report were derived, it was important to ensure similarity on demographic measures. As such, healthy participants were matched to depressed participants on adolescent age, gender, ethnicity, and the socioeconomic characteristics of their schools [51]. In summary, although the two samples were well matched on many demographic variables, the depressed participants came from households with somewhat lower socioeconomic status, and had mothers with higher levels of depressive symptoms. These differences are not surprising, and reflect well-established associations between adolescent depression, low socioeconomic status, and maternal depression [31].

The interactions were conducted in a quiet laboratory room at ORI. Family members were seated a few feet apart as would be typical for a discussion between familiars. Lapel wireless microphones (model: *Audio Technica* ATW-831-w-a300) were placed on the participants shirts at the chest level. Although participants were also outfitted with other sensors measuring physiological signals, such as ECG, impedance cardiogram, skin conductance, respiratory, and blood pressure, they did not impede speech behavior. The full 20 min were always used and the order of interactions was fixed: EPI, PSI, and FCI.

The recordings were coded by trained observers using the living-in-family-environments (LIFE) coding system [21]. The LIFE is a behavioral coding system designed to describe the specific timeline of various emotions (called affect codes) and verbal content (called content codes) displayed by the participants during the course of the interaction. The LIFE code is composed of 27 content codes and 10 affect codes. The speech was recorded using two channels and only the audio recordings from the channel belonging to the adolescents' microphone were analyzed. The speech of the adolescents was then segmented from the recordings based on the time annotations containing these emotions and verbal content that were LIFE coded by expert observers. The coding results were positively assessed for an interobserver agreement [22]. The average number of utterances for each adolescent was approximately 278, 251, and 240 for EPI, PSI, and FCI, respectively. The ratio of the adolescents' to parents' speech duration was 0.73, 0.71, and 0.67 for EPI, PSI, and FCI, respectively. The average duration of the speech segments was around 2 to 3 s long, and the sampling rate was 11 kHz.

IV. Methodology

The proposed framework in the modeling and classification of the depressed and control adolescents' speech is illustrated in Fig. 1.

For both the training and testing phases, detection of voiced frames using the linear predictive (LP) technique described in [11] was implemented by first segmenting the normalized speech signal into 25 ms with 50% overlap frames using a rectangular window. From these voiced frames, acoustic features were extracted and normalized within each subject. Statistical analyses were then carried out to discard any acoustic features that were statistically nonsignificant in distinguishing the speech of depressed adolescents from that of control adolescents. Finally, using two different machine-learning techniques of Gaussian mixture model (GMM) and the support vector machine (SVM), the selected extracted acoustic features were modeled into their respective classes (depressed and control class).

In the following Section IV-A–C, the extracted acoustic features, statistical analyses, and modeling techniques are described.

A. Acoustic Features and Feature Grouping

Similar to the procedure in [41], we also proposed the grouping of acoustic features into categories and subcategories that are closely related to the human speech production model. In our study, the acoustic features were grouped into five main feature categories that represented TEO-based, cepstral (C), prosodic (P), spectral (S), and glottal (G) features. Acoustic features grouped into these categories are closely related to the physiological and perceptual components that characterize speech in the human speech production model. The physiological components are related to the feature categories of TEO, prosodic (P), spectral (S), and glottal (G). The TEO feature category is derived from the nonlinear speech production model and measures the nonlinear airflow in the vocal tract, whereas, the feature categories of prosodic (P), spectral (S), and glottal (G) are derived from the linear speech production model of sound propagation along the vocal tract. The feature category of cepstral (C) which is also derived from the linear speech production model, relates to the perceptual aspect.

The acoustic features and their associated categories and subcategories are denoted in the first and second columns in Table II, respectively. The acoustic features are briefly discussed in the following sections.

1) TEO-based features—TEO-based features have shown good performances in stress recognition [56]. In the emotional states of anger or stress, fast air flow causes vortices located near the false vocal folds, which provide additional excitation signals other than pitch [53]. To model this time-varying vortex flow, Teager [52] proposed a nonlinear energy operator called the TEO, which computes an energy profile (also known as the TEO profile). The TEO in a discrete form [26] is defined in (1), where ト[.] is the TEO and *x*(*n*) is the *n*th speech sample point

$$\psi\left[x\left(n\right)\right]=x^{2}\left(n\right)-x\left(n+1\right)x\left(n-1\right). \quad (1)$$

.

Several TEO-based features have been proposed in the literature and we computed the TEO-critical-band-based autocorrelation envelope (TEO-CB-Auto-Env) feature that is based on the method discussed in [56]. Fig. 2(a) depicts the computation of the TEO-CB-Auto-Env feature coefficients. In our implementation, 512-point Gabor bandpass filters for the 15 CBs were used. We followed approximately the same frequency range for our 15 CBs as in [56]. Fig. 2(b) shows an example of the TEO profile waveform and the autocorrelation envelope for an utterance calculated within the 9th critical band (CB-9).

2) Cepstral feature—The MFCC was considered [14] because it has been effectively used in speech content characterization [50]. Optimization of the parameters in the MFCC was carried out to maximize the depressed and control classification accuracy. Based on our previous study [36], the optimized parameters selected were 30 triangular filters in the filter bank to calculate 12 original coefficients in the MFCC.

3) Prosodic features

a) Fundamental frequency (F0) and log energy (LogE): A small-scaled test was conducted on a subset of the data in which three approaches in the F0 extraction was evaluated (i.e., autocorrelation, cepstrum, and average magnitude difference function) using the Roger Jang's audio toolbox [24]. All three approaches yielded comparable results and the autocorrelation method was chosen for the full dataset. The values of F0 were determined on a frame-by-frame basis by finding the maximum values of the autocorrelation function within 40 to 1000 Hz range. The *LogE* [14] of the speech time waveform was calculated for each frame to determine the changes in speaking behavior in response to factors relating to stress, intonation, and emotions.

b) Formants (FMTS) and formant bandwidths (FBWS): A 13th-order LP filter was employed to calculate the formant frequencies. Only values of the first three formants (FMT_1–FMT_3) and formant bandwidths (FBW_1–FBW_3) below its Nyquist frequency were taken. The lower formants have been known to model spoken content [50].

c) Jitter and shimmer: Frequency perturbation also called jitter refers to the short-term (cycle-to-cycle) fluctuations in pitch. It is obtained by measuring the fundamental frequency (F0) of each cycle of vibration, subtracting it from the previous F0 values, and dividing it by the average F0. Shimmer, on the other hand, is calculated in similar fashion; however, the period-to-period variability of the signal peak-to-peak amplitude is calculated instead. In clinical treatment, jitter and shimmer have been widely used to describe the pathological characteristics of voice [42].

4) Spectral features

a) Spectral centroid: Spectral centroid (SC) indicates the center of a signal's spectrum power distribution. It is the calculated weighted mean of frequencies present in the signal, with their magnitudes as weights

$$\mathrm{SC}=\frac{\sum_{n=1}^{M} f(n)\,X(n)}{\sum_{n=1}^{M} X(n)} \quad (2)$$

where $X(n)$ represents the magnitude of frequency bin number n, $f(n)$ represents the center frequency bin, and M is the total number of frequency bins.

b) Spectral flux: Spectral flux (SF) is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the previous frame.

c) Spectral entropy: Spectral entropy (SE) is the means of measuring the amount of information based on Shannon's information theory and it has been applied to emotion recognition in speech [34].

d) Spectral roll-off: Spectral roll-off (SR) is the point, where the frequency that is below some percentage (set as 80% for our experiments) of the power spectrum resides. The equation for SR is as follows:

$$\sum_{n=1}^{K} X(n)=0.80\sum_{n=1}^{M} X(n) \quad (3)$$

where n is the frequency bin index, M is the total number of frequency bins, $X(n)$ is the amplitude of the corresponding frequency bin, and K is the spectral roll-off number.

e) Power spectral density: The one-sided PSD was computed based on the Welch spectral estimator method using a 4096-point fast Fourier transform with a 5-ms nonoverlapping hamming window size. The total power spectral for frequency 0– 2000 Hz, its subbands PSD_1 (0–500 Hz), PSD_2 (500–1000 Hz), PSD_3 (1000–1500 Hz), PSD_4 (1500–2000 Hz), and the ratio of power from each spectral subband to the total power were calculated. The PSD have been effectively used to discriminate between speech of control and depressed adults [18].

5) Glottal features

The glottal pulse and shape have been documented to play an important role in the analysis of speech in clinical depression [41], [46]. For our study, the glottal flow extraction used the TTK Aparat glottal inverse filtering toolbox [2]. The glottal inverse filtering method implemented here was based on an iterative adaptive inverse filtering algorithm (IAIF) [3]. Instead of using a LP filter, a discrete all-pole modeling (DAP) was implemented to model the vocal tract as it is less sensitive to the biasing of formants caused by nearby harmonic peaks. Like for the LPC, the number of formants (or resonances) to model the vocal tract in the DAP was set to 13 (Fs/1000 + 2). This was done to ensure that there was at least one formant within every kilohertz band of the vocal tract transfer function. Once the glottal flow was estimated, quantitative analysis of the glottal flow pulses was performed in the time and frequency domains. It should be noted that glottal waveform extraction is still a matter of study and accurate representations are still difficult to determine and verify. In this study, the glottal timing (GLT) and the glottal frequency (GLF) were used to represent the glottal flow parameters in the time and in the frequency domains, respectively.

The glottal flow can be divided into a few phases that are illustrated from the glottal flow pulse in Fig. 3(b). This is shown by the mark boundaries from the dotted lines indicating the GLT interval for the opening phase (OP), closing phase (CP), and closed phased (C) that describes the glottal pulse shape. It has been suggested that the glottal OP can be subdivided into two timing instances referred to as the primary opening (▴) and secondary opening (●) [47]. The duration of the primary and secondary openings of OP is denoted by T_{o1} and T_{o2}, respectively. The duration of CP is denoted by T_c and the period of the glottal cycle is denoted by T. Once these instances are acquired, several timing and frequency parameters can be easily calculated. In GLT, the timing parameters used is the open quotients (OQ_1 and OQ_2), approximation of the open quotient (OQ_a), quasi-open quotient (QOQ), speed quotients (SQ_1 and SQ_2), closing quotient (CIQ), amplitude quotient (AQ), and normalized amplitude quotient (NAQ). For the GLF, the frequency parameters used are the difference of the first and second harmonics [labeled H_1 and H_2 in Fig. 3(d)] in decibels of the glottal flow power spectrum (DH12), harmonic richness factor (HRF), and the parabolic spectral parameter (PSP). Table I shows a brief summary of their parameters and an in-depth description of them can be found in [2].

6) Delta (Δ) and Delta–delta (Δ-Δ) coefficients

The inclusion of the first- and second-order derivatives (delta and delta–delta), which can capture the temporal information among neighboring frames are calculated as follows:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta \left(c_{t+\theta} - c_{t-\theta} \right)}{2 \sum_{\theta=1}^{\Theta} \theta^2} \qquad (4)$$

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 window size is set $\Theta = 9$ to obtain both delta and delta–delta coefficients. The same formula (4) is applied to the delta coefficients to obtain the delta–delta coefficients. The delta and delta–delta were incorporated in all the acoustic features.

B. Statistical Analysis and Feature Selection

Table II presents the extracted acoustic features grouped into categories and subcategories for both male and female subjects from each interaction. A total of 14 acoustic features comprising of 186 feature coefficients, which included their delta and delta–delta coefficients, were statistically examined for significance in characterizing speech of depressed and control adolescents. This was done as a preliminary step to ensure that feature coefficients that gave a statistically nonsignificant result were removed in the modeling of depressed and control speech. Assumptions of parametric testing were examined for each feature coefficient to check if they were normally distributed within each of the depressed and control classes. The Kolmogorov–Smirnov (KS) test [17] indicated that the entire feature coefficients were normally distributed ($p > 0.05$). In order to identify relationships that might exist between the feature coefficients, a multivariate analysis of variance (MANOVA) procedure was conducted on pairwise comparison of the depressed and control classes. Instead of combining all the feature coefficients in MANOVA, which is considered a suboptimal approach, unless there is a good theoretical basis for doing so [17], individual subcategories representing the acoustic features along with their delta and delta–delta coefficients were examined with MANOVA. The reason behind this approach was that incorporation of delta and delta–delta coefficients in previous work [36], [37] has shown to result in an increase in classification results, and therefore, correlations should exists between the feature coefficients in each subcategorical feature in MANOVA. In MANOVA, multivariate group tests were performed on each subcategorical feature using *Wilks's lambda*. Features in the subcategory that met a significance level of $p < 0.05$ were retained. Otherwise, the feature was then followed up with a one-way analysis of variance (ANOVA) on each feature coefficient. Each feature coefficient that also met a significance level of $p < 0.05$ for ANOVA were kept in that subcategory. Otherwise, if all the feature coefficients in the subcategorical feature were still nonsignificant, the subcategorical feature was discarded.

Selected acoustic features in the subcategories and the number of coefficients are listed in Table II The plus sign indicates that the subcategorical feature produced a statistically significant result and the minus sign indicates that the result was statistically nonsignificant. We found that all the features (in the subcategory) were statistically different between depressed and control speech of female adolescents in all three interactions. However, for speech in the male adolescents, a few features were not significant, as indicated by a minus sign in Table II.

C. Modeling and Classification

1. *Gaussian mixture model:* GMM has been effectively used in speech information modeling tasks, such as speaker recognition and spoken language identification [7]. We used the HTK toolbox [55] for GMM-based depressed and control content modeling. In the implementation, expectation-maximization (EM) algorithm was used for estimating parameters of mean, covariance, and mixture weight of each Gaussian component in the GMMs. For computational efficiency, diagonal covariance matrices were used in the Gaussian component instead of full covariance.

2. *Optimized parallel SVM:* In recent years, SVM has also been widely used in speech content analysis [8]. To increase the computational efficiency when working with large training vectors, we followed a similar approach to that described in [12], where a single SVM was replaced by a parallel configuration of SVMs. The training of each SVM involves tuning the parameter of the kernel γ and the penalty error C [10]. The aim is to achieve the best (γ, C) pair in obtaining the optimal classification accuracy in predicting unknown data from the testing set. The

optimized parallel SVM (OPSVM) was implemented using the LIBSVM toolbox [10]. In our implementation, we first divided the training data of each class into random subsets. Each subset was then scaled to be [0, 1]. We then empirically modeled each training data subset using a SVM with radial-based function in the kernel. Searching for the most appropriate (γ, C) pair was performed through a grid search using fivefold cross validation on the training dataset.

In the classification, a hyperbolic tangent of a weighted sum of outputs from each individual SVM was taken. Instead of the neural network approach proposed in [12], a global optimization algorithm based on simulated annealing [39] was used to determine the weight associated with each SVM.

V. Experiments and Results

Experiments with the framework outlined in Section IV were carried out using the database described in Section III. Data from approximately 50% of the adolescents, including 33 depressed (23 females and 10 males) and 34 control subjects (21 females and 13 males) were used for testing, and the remaining data were used to train the depressed and control models. A series of experiments are briefly explained in the following (EXP1 to EXP 6). The results of these experiments are discussed in the following sections. The main objective was to correctly classify the test adolescents (alternatively called subjects) as either depressed or control. The subject-based correct classification accuracy (SBCCA) was calculated as described in the following equation:

$$\text{SBCCA}=\frac{\begin{pmatrix}\text{Number of correctly}\\ \text{classified subjects}\end{pmatrix}\times 100\%}{\text{Total number of subjects}}. \quad (5)$$

To determine the number of correctly classified subjects in (5), the utterance-based correct classification accuracy (UBCCA) was first calculated using the following formula:

$$\text{UBCCA}=\frac{\begin{pmatrix}\text{Number of correctly}\\ \text{classified utterances}\end{pmatrix}\times 100\%}{\text{Total number of utterances}}. \quad (6)$$

If UBCCA for a given subject was greater than 50% for the depressed class, then the subject was classified as depressed. Since, the predicted classes of the test subjects were known; the total number of correctly classified subjects could therefore be calculated and used in (5) to determine the SBCCA values. In addition, the correct classification of depressed and control subjects were measured in terms of sensitivity, specificity, and the overall accuracy defined as follows: True positive (TP) = Number of depressed subjects classified as depressed False negative (FN) = Number of depressed subjects classified as control True negative (TN) = Number of control subjects classified as control False positive (FP) = Number of control subjects classified as depressed

$$\begin{aligned}
&\text{True positive (TP)}\\
&=\text{Number of depressed subjects classified as depressed}\\
&\text{False negative (FN)}\\
&=\text{Number of depressed subjects classified as control}\\
&\text{True negative (TN)}\\
&=\text{Number of contorl subjects classifed as control}\\
&\text{False positive (FP)}\\
&=\text{Number of control subjects classifeid as depressed}\\
&\text{Sensitivity}=\tfrac{\text{TP}}{\text{TP+FN}}\times 100\%\\
&\text{Specificity}=\tfrac{\text{TN}}{\text{TN+FP}}\times 100\%
\end{aligned} \tag{7}$$

$$\text{Overall accuracy}=\frac{\text{TP+TN}}{\text{TP+FN+TN+FP}}\times 100\%. \tag{8}$$

.

In general, the objective was to achieve the highest overall classification accuracy by obtaining an optimal sensitivity to specificity ratio (ideally > 1), and at the same time, keeping the ratio between sensitivity and specificity at a reasonable margin (to avoid class skews). However, in some cases, it was not possible to achieve reasonably high accuracy without making the sensitivity to specificity ratio <1.

1. *EXP1:* Using two feature categories, i.e., TEO and cepstral (C), the effectiveness of gender-independent (GIM) and gender-dependent modeling (GDM) techniques for depressed and control adolescent classification was first examined.

2. *EXP2:* Next, testing on different lengths of utterances from the testing set was examined.

3. *EXP3:* Using the best gender modeling strategy and optimal test utterance length found in EXP1 and EXP2, the effectiveness of other feature categories of prosodic (P), spectral (S), glottal (G), and their category combinations for depressed and control adolescent classification was investigated.

4. *EXP4:* From our database described in Section III, the study of feature categories proposed in recent published work by others was carried out.

5. *EXP5:* Next, the top feature category out of the TEO and C categories that yielded the highest classification accuracy was selected based on their performances in EXP1 and combined with the other feature categories (P, S, and G) as described in EXP3.

6. *EXP6:* Due to high performance in modeling speech contents, GMM was employed for modeling speech of depressed and control adolescents in EXP1–EXP5. In the final experiment, the best feature category combination obtained in EXP5 was compared with the SVM classifier because of the advantageous properties of its generalization capabilities for solving two-class problems.

All classification results were cross validated based on four turns using different training and testing sets.

A. Effectiveness of GIM and GDM (EXP1)

As noted in Section III, although participants were matched on a range of demographic variables, we only considered gender in our analyses because the development of gender differences in depressive symptoms has been documented to occur during early adolescence [45]. Accordingly, the examination of how gender differences might affect classification accuracies in clinical depression was analyzed. For the purpose of this experiment, two feature categories (see Table II) of TEO and cepstral (C) were selected as a starting point for the analysis. These two types of features have been previously reported as effective discriminators of stress and emotion in speech [56] and also in speaker and language detection [50]. Since the depression is often characterized as an affect (emotion) regulation [51] disorder, it was expected that the cepstral and TEO-based features could also provide good results in the depression detection in speech. The GDM, depressed, and control class models were generated separately from the feature vectors of male and female subjects using the GMM-based training procedure. The GIM depressed and control class models on the other hand, were trained by combining together feature vectors from both male and female subjects. Similar testing length as in [46] of 0.5 min for each utterance was achieved by concatenating only the voiced sections from the utterances belonging to each adolescent. Depressed and control classification using GDMs were carried out assuming that the gender of the test adolescent is known.

Fig. 4 shows the overall classification performances based on SBCCAs for both GDMs and GIMs. In all the interactions, as can be observed in Fig. 4, GDMs outperformed the GIMs. It was also observed in Fig. 4 that the TEO-based features consistently outperformed the cepstral and cepstral + TEO combination in both GDMs and GIMs. Compared to TEO with GIMs, TEO with GDMs improved the SBCCA by 3.8%, 9.3%, and 2.9% for EPI, PSI, and FCI, respectively.

Table III shows the sensitivity and specificity of the SBCCA with feature category TEO for GIM and GDM. From the sensitivity results in the table, it can be observed that GDMs performed better than GIMs in detecting depressed subjects when both males and females are modeled separately in all interaction contexts. GDMs for the males resulted in an accuracy improvement in sensitivity of 1.8%, 30.2%, and 8.8% for EPI, PSI, and FCI, respectively, compared to GIMs. For the females, GDMs also resulted in an accuracy improvement in sensitivity of 2.4%, 33.5%, and 22.2% for EPI, PSI, and FCI, respectively, compared to GIMs.

B. Optimization of the Test Utterance Length (EXP2)

In previous research, test utterances of 20 s [18] and 0.5 min [46] in length have been used for depressed and control subject classification. To determine the optimal duration of speech samples to be analyzed, we carried out EXP2 to examine SBCCA with different duration (i.e., 0.5, 1, 2, and 3 min) of concatenated voiced utterances. In these experiments, we used the feature category of TEO with the GDMs because this setting outperformed the others in previous experiments in EXP1. Experimental results are shown in Fig. 5. With reference to the accuracies using 1 min utterances, we noticed around 7.1%, 5.1%, and 7.0% average SBCCA drops (of all interactions) for the male subjects and 2.6%, 5.6%, and 6.6% SBCCA drops for the female subjects using 0.5, 2, and 3 min utterance durations, respectively. It can be observed in Fig. 5 that the SBCCA measure was consistently achieving the highest value for utterance length of 1 min; this length of the test utterances was, therefore, chosen as a length to be used in our subsequent experiments.

C. Effectiveness of Prosodic, Spectral, and Glottal Feature Categories, and Their Combinations (EXP3)

Further analyses were conducted on the other different feature categories representing prosodic (P), spectral (S), and glottal (G) (see Table II) using the GDMs. Table IV presents overall accuracy results based on SBCCA for feature categories P, S, G, and their different combinations for male and female subjects in all the interactions (EPI, PSI, and FCI). From the table, a few key findings can be observed.

First, for the males, the influence of G on individual categories P and S (i.e., P + G and S + G) improved classification accuracy compared to P and S alone. Compared to P alone, P + G increased SBCCA 0.8%, 15.1%, and 6% in the EPI, PSI, and FCI, respectively. Also for males, compared to S alone, S + G increased accuracy by 1.4%, 2.6%, and 6.7% for the EPI, PSI, and FCI, respectively.

Second, for the females, combining G with other feature categories also improved classification rates in the EPI and FCI. In the EPI and FCI for females, P + G showed a 2.6% and 6.3% improvement over P alone. In the EPI, PSI, and FCI, S + G showed a 6.7%, 8.6%, and 0.6% improvement over S alone. However, in PSI, a slight decrease of 1.8% was shown for P + G when compared to P alone.

Third, for both males and females, combining G with P + S also improved the classification accuracy for all the interactions. In the case for males, P + S + G improved accuracy over P + S by 4.6%, 8.9%, and 11.2% for EPI, PSI, and FCI, respectively. For the females, P + S + G also improved accuracy over P + S by 2.8%, 7%, and 4.4% for EPI, PSI, and FCI, respectively.

Fourth, for the EPI and FCI, P, S, G, and their different combinations were better discriminators in the female than the male samples. A similar trend emerged for the PSI, with the exception that G and P + G for females showed a decreased in SBCCA rates over the males of 11.8% and 0.4%, respectively.

Based on the aforementioned observations, it is more likely that feature category G and its combination with P and S can improve SBCCA for both male and female subjects.

D. Study of Feature Category Combinations Proposed in [41] (EXP4)

The classification performances of the proposed feature categories of prosodic (P_o), vocal tract (V_o), and glottal (G_o) defined in [41] was examined with our database. This examination was aimed at determining if we could establish any similar trends or performances in defining our own feature categories and subcategories with those proposed in [41]. Results are presented in Table V. Note that although the same groupings of feature categories as [41] were implemented, the subcategorical features were slightly different. Also, in [41], the subcategorical features of formants and formant bandwidths were partitioned to have another feature category that represented measurements of the vocal tract shape and length (denoted V_o in Table V). However, for our grouping (see Table II), the prosodic feature category contained the subcategorical feature of formants and formant bandwidths. Therefore, it is not surprising that the results for P_o + V_o and P_o + V_o + G_o in Table V are the same as P and P + G in Table IV.

Similar to results based on the combinations of glottal features, an improvement in classification rates is seen for P_o + G_o compared to P_o alone for both male and female adolescents (see Table V). Comparing Table IV and Table V, it is observed that using our entire feature category combinations of P + S + G (see Table IV) yielded better classification accuracies when compared to the grouping of P_o + V_o + G_o (see Table V) for both male and female adolescents.

For the males, P + S + G gave a 5.9%, 1.2%, and 4.2% classification accuracy increase in the EPI, PSI, and FCI, respectively, as compared to P_o + V_o + G_o. For the females, P + S +

G gave a 3.2%, 9.7%, and 1.9% classification accuracy increase in EPI, PSI, and FCI, respectively, as compared to $P_o + V_o + G_o$.

E. Performance Analysis by Combining TEO Category With Prosodic, Spectral, and Glottal Categories (EXP5)

In EXP1, the TEO-based features showed the best overall performance, therefore, in the next stage of our experiments, the effect of combining the TEO-based features with the P, S, and G features was investigated. Table VI summarizes the influences in percentage increase or decrease in accuracies when the feature category of TEO was added to P, S, and G, compared with having P, S, and G and their different combinations alone, as shown in Table IV. For the males, when TEO feature category was combined, a significant accuracy increment was observed in all the interactions. To determine if these increments were statistically significant, a McNemar's test was conducted on paired feature categories (i.e., P and P + TEO) between the fourfold cross-validation results of each category. As highlighted in bold in Table VI, statistical significance in accuracy increments ($p < 0.05$) were obtained for all the males (highlighted in bold) when TEO was added to the other feature categories. Interestingly, in most cases, the TEO feature category itself, showed higher classification accuracy for both male and female adolescents throughout all the interactions. This is shown in Table VII, where the classification results are presented in terms of sensitivity, specificity, and overall accuracy for the TEO feature category.

F. Comparison With SVM (EXP6)

The results that have been discussed so far were based on the GMM. To examine whether the classification accuracies were biased with respect to GMM, the best results obtained with GMM, i.e., TEO with GMM were compared with the results of TEO with SVM.

SVM was implemented in the form of OPSVM discussed in Section IV-C2. The number of subsets in OPSVM to be trained was varied from 10 to 30 with a step size of 10. The number of subsets (or SVMs) in OPSVM that gave the highest classification was chosen in the model selection. The optimal number of subsets (or SVMs) in OPSVM, which maximized the SBCCA with TEO was 20. Table VIII shows the results of equal weights and optimal weights calculated from the global optimization algorithm in the OPSVM. In the male sample, compared to using equal weights, the optimal weights in OPSVM increased the SBCCA by approximately 5.4%, 10.3%, and 3.9% in the EPI, PSI, and FCI, respectively. In the female sample, SBCCA increments in using optimal weights were approximately 10.2%, 6.6%, and 8.9% in the EPI, PSI, and FCI, respectively. Although OPSVM yielded very similar results as compared with the GMM modeling technique, the computational time required in training the models of our dataset was less efficient compared to the GMM.

VI. Discussion and Conclusion

Speech is known to contain important information regarding a person's psychological state [43]. Thus, a speech-based depression detection system could serve as a screening tool to assist mental health professionals in identifying clinically depressed persons. As this system is intended as the first stage of a diagnostic process and not a definitive identifier (i.e., detection via this type of system would normally be followed by full clinical evaluation of these screened as potentially depressed), the objective was to identify more depressed subjects (higher sensitivity) rather than to screen out negative cases (higher specificity).

Using speech recorded during interactions between adolescents and their parents, this paper reports an investigation of five acoustic feature categories (i.e., the TEO, cepstral (C), prosodic (P), spectral (S), and glottal (G) features) for detecting clinical depression in adolescents. Experiments were performed on speech recorded during three interaction tasks that were designed to create different types of interactional contexts (see Section III). The proposed acoustic feature categories were formed based on the physiological and perceptual similarities of the speech production model. The TEO and C feature categories were selected as our starting point, since they have been effectively employed in speaker, language, and emotion content modeling [50], [56].

Previous psychological studies reported significant differences in depressive symptoms between adolescent males and females [45]. Therefore, the influence of gender differences in depression detection was first investigated using feature categories of TEO and C. (see EXP1 in Section V-A). Experimental results in Fig. 4 indicated higher subject-based detection rate (average of 5.3%) of depressed subjects with GDM than with GIM. This is consistent with those previous psychological studies that have suggested significant variations in depressive symptoms based on the gender [45].

In order to build an accurate screening system for clinically depressed subjects, it is important to study how the test utterance length affects the overall performances of the system. Our experiments based on the TEO feature category (EXP2 in Section V-B) indicated that utterances with 1 min of speech content improved the subject-based classification accuracy.

Experiments with feature categories P, S, and G (EXP3, Section V-C) yielded accuracy improvements when G was combined with P, S, or P + S for the male subjects in all the interactions (see Table IV). For the female subjects, this trend was similar except for the P + G combination in the PSI, whereby there was a 1.8% average accuracy drop compared to P alone.

In searching for a relatively independent reference point that would allow us to verify these findings, feature categories similar to recent published research [41] were examined in our database (EXP 4 of Section V-D). Consistent with past research [41], implementation of both our proposed feature categories and feature categories from [41] demonstrated that the critical role of the glottal feature category, which when added to the other stand-alone feature categories, increased the overall discrimination between speech of depressed and control classes. However, in EXP3 for females, the increase in accuracy for P + G was not shown during the PSI. These findings could be due to the fact that the PSI is the interaction that is most likely to elicit conflictual behavior, which in turn could contribute to an increase in pitch during angry and loud speech in these stressful scenarios. The rapid motion of the glottis caused by the increased in pitch does not always yield complete closure. Therefore, increased pitch could yield difficulties in obtaining reliable information about the changes in the glottal waveform. These difficulties are especially pronounced for females as they tend to exhibit higher pitch [40].

Comparing Table IV and Table V, it can be observed that our final combinations of P + S + G gave higher classification results compared to the combinations of proposed feature categories of $P_o + V_o + G_o$ in [41]. One possible reason for the increase is that in the spectral (S) feature category, the acoustic subcategorical feature of PSD was included and it has been noted in past research that PSD provides a superior discrimination between the speech of control and depressed adults [18].

In EXP5 of Section V-E, it was found that by adding the TEO feature category to different combinations of the P, S, and G features listed in Table VI, the classification accuracy increased in all cases for the males and in some cases for the females. Most interestingly, the TEO-based features, when used on their own, clearly outperformed all other features and their combinations. This pattern held for all three interactions across both genders. In Table VII, it can be seen for the males that the TEO feature category yielded correct classification scores of 81.36%, 82.96%, and 86.64%, respectively. For the females, the TEO feature category yielded correct classification scores of 78.87%, 75.70%, and 72.01%, respectively. Looking across the different interactions in Table VII, it can also be observed that although the overall accuracy and the specificity measures did not provide consistent results, there is a clear pattern within the sensitivity measure, which shows that the PSI provides consistently higher results for both male and female subjects. This again can be attributed to the fact that the PSI evokes situations most likely to elicit conflicting behavior, and therefore, produces more pronounced changes in speech acoustics in identifying depression.

A. Why do TEO and Glottal Features Significantly Improve the Detection Accuracies of Clinically Depressed Subjects?

It was observed that the glottal features boosted the accuracy of discrimination between speech of depressed and control adolescents. TEO-based features also appear to be powerful discriminants of depression in speech. Both observations maybe closely related to the physical impact of depression on the speech production processes through the vocal folds and tract (tube extending from vocal folds to the lips). In order to explore this further, it is helpful to briefly discuss the main processes in speech production.

Assuming that speech is an amplitude and frequency (AM–FM) modulated signal, the TEO parameter represents a measure of instantaneous energy calculated not only as a function of signal amplitude but frequency as well [38], [56]. This indicates that the TEO values contain information about spectral distribution of the signal energy and show sensitivity to the presence of additional harmonics and cross harmonics in the speech signal [56].

The experimental studies of the vocal flow formation [6], [27]–[29], [52], on the other hand, provide strong evidence that the glottal air flow has a nonlinear character with a laminar flow component as well as additional turbulent components called vortices. In [27]–[29] two types of vortices were identified; each occurring in a specific part of the vibration cycle, and at a certain location relative to the glottis. During the early opening phase of the vocal folds, when the glottis is convergent, supraglottal vortices occur above the vocal folds. During the latter part of the vocal fold closing, when the glottis is divergent, intraglottal vortices are formed between the vocal folds. The intraglottal vortices can alter the vibration of vocal folds, whereas, the supraglottal vortices provide additional sound sources when hitting hard surfaces of the vocal tract or interacting with each other. It was demonstrated in [28] that the level of symmetry in the vocal fold vibration has a strong effect on the glottal energy distribution across the frequency spectrum. It has been postulated that these additional sound sources [25], [56] generate extra harmonics and cross harmonics in speech.

As indicated in [56], the number of supraglottal vortices is likely to be related to the level of emotional stress. Moreover, the tension of laryngeal muscles responsible for the stiffness of the vocal folds (and hence, the vortices) is controlled by the sympathetic nervous system. Hence, it is likely that different patterns of the glottal wave formation reflect different emotional or mental states of a speaker, and therefore, contain important cues for the recognition of depression in speech.

Fig. 6 shows the average normalized area of the autocorrelation envelope for all the speech frames in the TEO feature category in both the depressed (marked with “X”) and control class (marked with “O”). The normalized area measurement was plotted for all the CBs in the TEO feature category. The normalized area details the strengths of the produced additional harmonics within the CB, which further indicates the turbulent air flow occurring during the phonation process. This area parameter in TEO has also been documented to provide useful assessment in vocal fold pathology [19]. It is evident from Fig. 6 that the average normalized areas of the CBs for the speeches of depressed subjects are higher than the speeches of control subjects. This pattern indicates that higher additional harmonics are generated in the depressed speech than in the control speech. Therefore, the result suggests that more vortices appear in the air flow during the phonation process for the depressed subjects than for the control subjects. How the glottal feature, TEO-based feature, and dynamics of the air flow are linked to the regulatory psychophysiological processes occurring during depression remains to be investigated. Though not conclusive, recent studies [15], [23] have suggested that the speech production systems show physical manifestations of the psychological difficulties of depressed persons (vocal folds and vocal tract). In such cases, patterns (laminar and vortices) of air flow in the speech production system of depressed subjects differ from the air flow of control subjects. For example, clinical depression may have a significant effect on vocal fold dysfunction. This could explain why the glottal features are effective in differentiating depressed from control speech. The TEO-based feature detects the presence of the extra harmonics and cross harmonics generated by the vortices, making it an effective feature for discrimination between depressed and control speeches.

The plots of the area under the normalized autocorrelation envelope in Fig. 6 provided clearer distinction between the depressed and control classes for the male subjects than for the female subjects. This observation indicates that there is a higher variation in the number of additional harmonics between the depressed and control male subjects than between the depressed and control female subjects. Therefore, it appears that the effects of depression on the voice characteristics of male subjects are more profound than the effects on the voice characteristics of female subjects.

These observations are consistent with the results in Table VI showing that the addition of TEO features to the other types of features provided statistically significant (McNemar's test, $p < 0.05$) improvement of the classification accuracy only in the case of male subjects. The observed small increase of the classification accuracy for the female subjects was found to be statistically insignificant. It is possible that these differences are related to the fact that there are clear differences between types of depression most frequently exhibited in males and females, however, further investigations are needed.

In summary, our study showed that clinical depression can be detected in adolescents using naturalistic speech samples. The classification accuracy strongly depended on the gender and on the type of acoustic features. The nonlinear approach of the TEO-based feature category provided the highest correlation with depression in the speech of both male and female adolescents. However, clinical depression detection still remains a challenging task due to the large number of potential genetic, psychological, social, cultural, and environmental factors that contribute to the development of this condition [51]. In addition, a potential limitation of this study is that the speech may contain some features that are specific to the family context, or that are primarily elicited by parental behavior. Therefore, in future studies, we plan to verify our findings on a different database and also investigate different nonlinear approaches for modeling depressive speech characteristics in order to improve discrimination between depressed and control subjects.

Acknowledgments

The authors would like to thank the families, who participated in the data collection sessions. They would also like to thank the Editor-in-Chief Prof. B. Wheeler, the Associate Editor Prof. P. Loizou, and the reviewers for their valuable feedback and suggestions.

This work was supported by the Australian Research Council Linkage Project LP0776235 and Oregon Research Institute under Grant R01 MH65340.

References

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

[2]. Airas M. TKK Aparat: An environment for voice inverse filtering and parameterization. Logopedics Phoniatrics Vocology. 2008; vol. 33(no. 1):49–64.

[3]. Alku P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. Speech Commun. 1992; vol. 11(no. 2–3):109–118.

[4]. Alku P, Strik H, Vilkman E. Parabolic spectral parameter—A new method for quantification of the glottal flow. Speech Commun. 1997; vol. 22(no. 1):67–79.

[5]. Alpert M, Pouget ER, Silva RR. Reflections of depression in acoustic measures of the patient's speech. J. Affect. Disorders. 2001; vol. 66(no. 1):59–69. [PubMed: 11532533]

[6]. Barney A, Shadle CH, Davies P. Fluid flow in a dynamic mechanical model of the vocal folds and tract. I. Measurements and theory. J. Acoust. Soc. Amer. 1999; vol. 105(no. 1):444–455.

[7]. Bishop, CM. Pattern Recognition and machine learning. Springer; New York: 2006.

[8]. Burges CJC. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Disc. 1998; vol. 2(no. 2):121–167.

[9]. Cavenar, JO.; Keith, H.; Brodie, H.; Weiner, RD. Signs and Symptoms in Psychiatry. Lippincott Williams & Wilkins; Philadelphia: 1983.

[10]. Chang, CC.; Lin, CJ. LIBSVM: A library for support vector machines. 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[11]. Childers, DG. Speech Processing and Synthesis Toolboxes. Wiley; Chichester, New York: 2000.

[12]. Collobert R, Bengio S, Bengio Y. A parallel mixture of SVMs for very large scale problems. Neural Comput. 2002; vol. 14(no. 5):1105–1114. [PubMed: 11972909]

[13]. Darby JK, Hollien H. Vocal and speech patterns of depressive patients. Folia Phoniatrica. 1977; vol. 29(no. 4):279–291. [PubMed: 604242]

[14]. Deller, JR.; Proakis, JG.; Hansen, JH. Discrete Time Proc. Speech Signals. Prentice Hall PTR; Upper Saddle River, NJ: 1999.

[15]. Dietrich M, Abbott KV, Schmidt JG, Rosen CA. The frequency of perceived stress, anxiety, and depression in patients with common pathologies affecting voice. J. Voice. Jul.2008 vol. 22(no. 4):472–488. [PubMed: 18395419]

[16]. Ellgring H, Scherer KR. Vocal indicators of mood change in depression. J. Nonverbal Behav. 1996; vol. 20(no. 2):83–110.

[17]. Field, AP. Discovering Statistics Using SPSS: (and sex, drugs and rock 'n' roll). 2nd ed. Sage; London: 2005.

[18]. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. Jul.2000 vol. 47(no. 7): 829–837. [PubMed: 10916253]

[19]. Hansen JHL, Gavidia-Ceballos L, Kaiser JF. A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment. IEEE Trans. Biomed. Eng. Mar. 1998 vol. 45(no. 3):300–313. [PubMed: 9509746]

[20]. Hollien H, Green R, Massey K. Longitudinal research on adolescent voice change in males. J. Acoust. Soc. Amer. 1994; vol. 96(no. 5):2646–2654. [PubMed: 7983270]

[21]. Hops H, Biglan A, Arthur J, Sherman L, Tolman A, Longoria N. Living in family environments (LIFE) coding system: Reference manual for coders. 2003 [Online]. Available: http://www.ori.org/projects/LifeCode/Introduction.html.

[22]. Hops H, Davis B, Longoria N. Methodological issues in direct observation-illustrations with the living in familial environments (LIFE) coding system. J. Clin. Child Psychol. 1995; vol. 24(no. 2):193–203.

[23]. Husein OF, Husein TN, Gardner R, Chiang T, Larson DG, Obert K, Thompson J, Trudeau MD, Dell DM, Forrest LA. Formal psychological testing in patients with paradoxical vocal folds dysfunction. J. Laryngoscope. Apr.2008 vol. 118:740–747.

[24]. Jang, R. Audio Processing Toolbox. 1996. [Online]. Available: http://neural.cs.nthu.edu.tw/jang/

[25]. Kaiser, JF.; Titze, I.; Scherer, R. Vocal Folds Physiol.: Biomech. Acoust. Phonatory Control. Iowa Univ. Press; 1983. Some observations on vocal tract operation from a fluid flow point of view; p. 358-386.

[26]. Kaiser JF. On a simple algorithm to calculate the 'energy' of a signal. Proc. IEEE Int. Conf. Acoustic, Speech, Signal Process. 1990:381–384.

[27]. Khosla S, Murugappan S, Gutmark E. What can vortices tell us about vocal fold vibration and voice production. Curr. Opin. Otolaryngo. 2008; vol. 16(no. 3):183–187.

[28]. Khosla S, Murugappan S, Paniello R, Ying J, Gutmark E. Role of vortices in voice production: Normal versus asymmetric tension. Laryngoscope. 2009; vol. 119(no. 1):216–221. [PubMed: 19117305]

[29]. Khosla S, Muruguppan S, Gutmark E, Scherer R. Vortical flow field during phonation in an excised canine larynx model. Ann. Otol. Rhinol. Laryngol. 2007; vol. 116(no. 3):217–228. [PubMed: 17419527]

[30]. Klerman GL. The current age of youthful melancholia – Evidence for increase in depression among adolescents and young adults. Brit. J. Psychiatry. 1988; vol. 152:4–14. [PubMed: 3167377]

[31]. Klein DN, Lewinsohn PM, Seeley JR, Rohde P. A family study of major depressive disorder in a community sample of adolescents. Arch. Gen. Psychiatry. 2001; vol. 58(no. 1):13–20. [PubMed: 11146753]

[32]. Kuny S, Stassen HH. Speaking behavior and voice sound characteristics in depressive patients during recovery. J. Psychiatric Res. 1993; vol. 27(no. 3):289–307.

[33]. Langlieb AM, DePaulo JRJ. Etiology of depression and implications on work environment. J. Occupat. Environ. Med. 2008; vol. 50(no. 4):391–395.

[34]. Lee W-S, Roh Y-W, Kim D-J, Kim J-H, Hong K-S. Speech emotion recognition using spectral entropy. Intell. Robotics Appl. 2008; vol. 5315:45–54.

[35]. Lewinsohn PM, Rohde P, Seeley JR. Major depressive disorder in older adolescents: Prevalence, risk factors, and clinical implications. Clin. Psychol. Rev. 1998; vol. 18(no. 7):765–794. [PubMed: 9827321]

[36]. Low LSA, Maddage NC, Lech M, Sheeber LB, Allen NB. Content based clinical depression detection in adolescents. Proc. Eur. Signal Process. Conf. 2009:2362–2365.

[37]. Low LSA, Maddage NC, Lech M, Sheeber LB, Allen NB. Influence of acoustic low-level descriptors in the detection of clinical depression in adolescents. Proc. IEEE Int. Conf. Acoustic. 2010:5154–5157. Speech, Signal Process.

[38]. Maragos P, Quatieri T, Kaiser JF. On amplitude and frequency demodulation using energy operators. IEEE Trans. Signal Process. Apr.1993 vol. 41(no. 4):1532–1550.

[39]. Mitchell A, Lech M, Kokotoff DM, Waterhouse R. Search for high performance direct contact stacked patches using optimization. IEEE Trans. Antennas Propag. Feb.2003 vol. 51(no. 2):249–255.

[40]. Moore E, Clements M. Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. Proc. IEEE Int. Conf. Acoustic. 2004; vol. 1:101–104. Speech, Signal Process.

[41]. Moore E, Clements MA, Peifer JW, Weisser L. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. IEEE Trans. Biomed. Eng. Jan.2008 vol. 55(no. 1):96–107. [PubMed: 18232351]

[42]. Moran RJ, Reilly RB, de Chazal P, Lacy PD. Telephony-based voice pathology assessment using automated speech analysis. IEEE Trans. Biomed. Eng. Mar.2006 vol. 53(no. 3):468–477. [PubMed: 16532773]

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

[44]. Nilsonne A, Sundberg J, Ternstrom S, Askenfelt A. Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression. J. Acoust. Soc. Amer. 1988; vol. 83(no. 2):716–728. [PubMed: 3351130]

[45]. Nolenhoeksema S, Girgus JS. The emergence of gender differences in depression during adolescence. Psychol. Bull. 1994; vol. 115(no. 3):424–443.

[46]. Ozdas A, Shiavi RG, Silverman SE, Silverman MK, Wilkes DM. Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk. IEEE Trans. Biomed. Eng. Sep.2004 vol. 51(no. 9):1530–1540. [PubMed: 15376501]

[47]. Pulakka, H. Master's thesis. Helsinki University of Techn.; Espoo, Finland: 2005. Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography.

[48]. Scherer KR. Expression of emotion in voice and music. J. Voice. 1995; vol. 9(no. 3):235–248. [PubMed: 8541967]

[49]. Scherer KR, Zei B. Vocal indicators of affective disorders. Psychother. Psychosom. 1988; vol. 49(no. 3–4):179–186. [PubMed: 3070621]

[50]. Schuller B, Batliner A, Seppi D, Steidl S, Vogt T, Wagner J, Devillers L, Vidrascu L, Amir N, Kessous L, Aharonson V. The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals. Proc. Interspeech. 2007:2253–2256.

[51]. Sheeber LB, Allen NB, Leve C, Davis B, Shortt JW, Katz LF. Dynamics of affective experience and behavior in depressed adolescents. J. Child Psychol. Psychiatry. 2009; vol. 50(no. 11):1419–1427. [PubMed: 19702661]

[52]. Teager HM. Some observations on oral air flow during phonation. IEEE Trans. Acoust. Speech, Signal Process. Oct.1980 vol. 28(no. 5):599–601.

[53]. Teager HM, Teager SM. Evidence for nonlinear sound production mechanisms in the vocal tract. NATO Adv. Study Inst., Ser. D. 1990; vol. 55:241–262.

[54]. Tonge B. Depression in young people. Australian Prescrib. 1998; vol. 21(no. 1):20–22.

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

[56]. Zhou GJ, Hansen JHL, Kaiser JF. Nonlinear feature based classification of speech under stress. IEEE Trans. Speech Audio Process. Mar; 2001 vol. 9(no. 3):201–216.

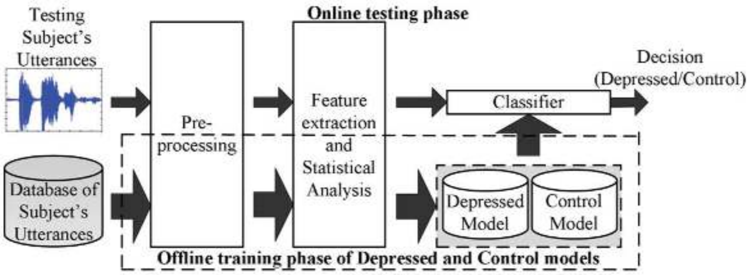

Fig. 1.
Block diagram in modeling speech of depressed and control adolescents.

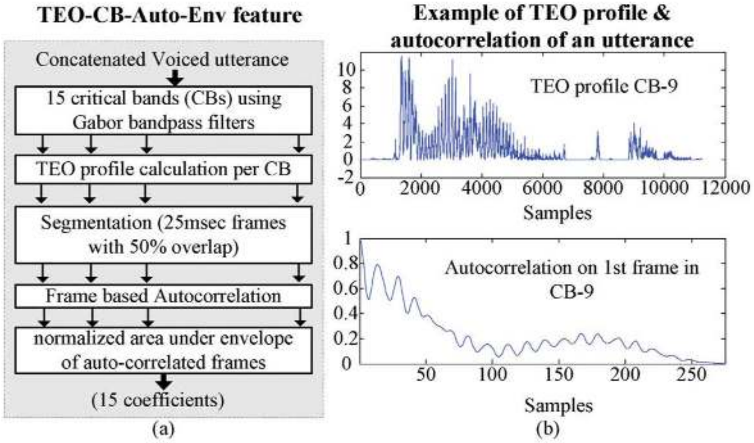
TEO-CB-Auto-Env feature
Concatenated Voiced utterance
15 critical bands (CBs) using Gabor bandpass filters
TEO profile calculation per CB
Segmentation (25msec frames with 50% overlap)
Frame based Autocorrelation
normalized area under envelope of auto-correlated frames
(15 coefficients)
(a)
Example of TEO profile & autocorrelation of an utterance
TEO profile CB-9
10
8
6
4
2
0
-2
0
2000
4000
6000
8000
10000
12000
Samples
Autocorrelation on 1st frame in CB-9
1
0.8
0.6
0.4
0.2
0
50
100
150
200
250
Samples
(b)

Fig 2.

TEO-CB-Auto-Env feature (a) Feature extraction implementation (b) Example of the TEO profile and the autocorrelation envelope for an utterance within the CB-9.

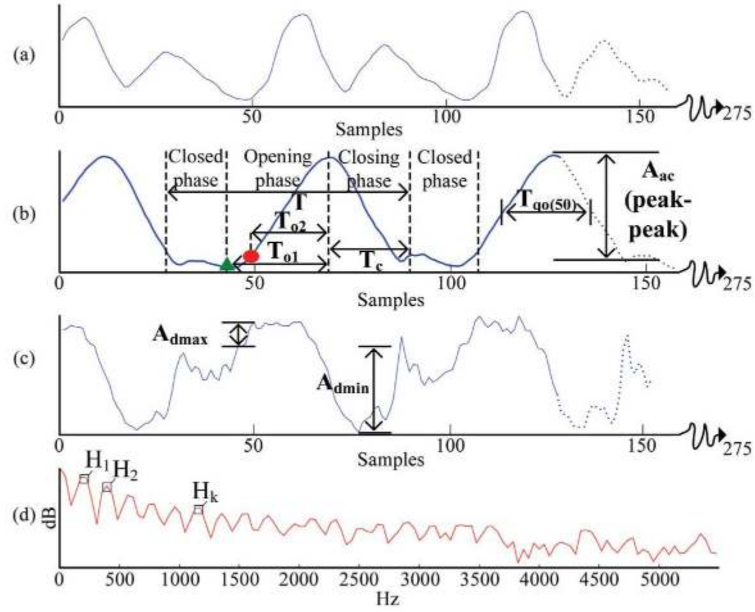

Fig 3.

Glottal inverse filtering (a) Speech frame of 25 ms (b) Glottal flow estimate (c) Glottal flow derivative (d) Glottal flow spectrum.

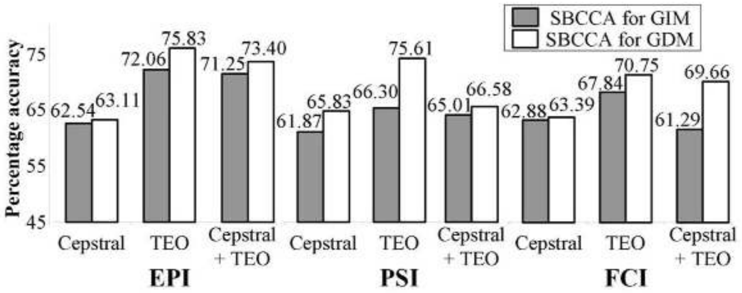

Fig 4.

SBCCA for the TEO and cepstral features using GIM and GDM.

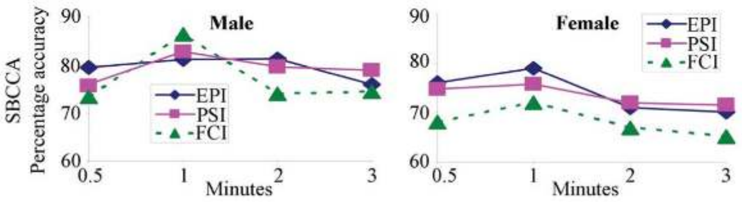

Fig 5.

Classification accuracies using different concatenated test utterances length for TEO feature category.

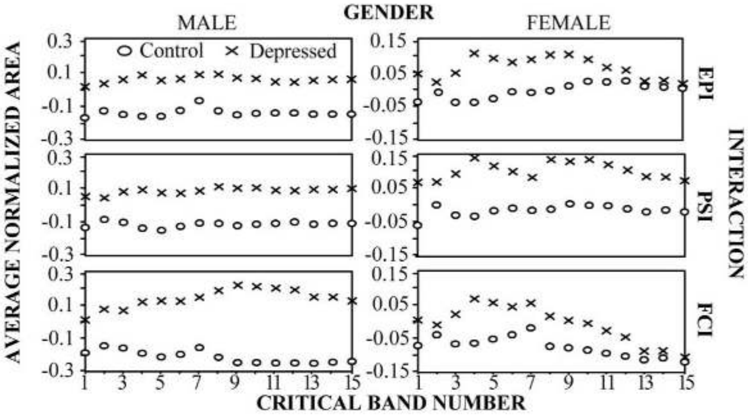

Fig 6.

Average frames (25 ms) normalized area under the autocorrelation envelope for the TEO feature category for each of the 15 CBs in all adolescents within the depressed and control classes.

TABLE I

Glottal Features—Timing Parameters (GLT) and Frequency Parameters (GLF)

	Feature parameter	Symbo	Description	Calculation method
GLT	1	OQ$_1$	The ratio of the primary opening phase to the length of glottal cycle duration.	$\frac{T_{o1} + T_c}{T}$
	2	OQ$_2$	The ratio of the seconday opening phase to the length of glottal cycle duration.	$\frac{T_{o2} + T_c}{T}$
	3	OQ$_a$	Approximates the opening to OQ for an ideal LF pulse.	$A_{ac}\left(\frac{\pi}{2A_{d\ \max}} + \frac{1}{A_{d\ \min}}\right)f_o$
	4	QOQ	The time of the open phase duration that is 50% above the peak to peak amplitude of the glottal flow.	$\frac{T_{qo}(50)}{T}$
	5	SQ$_1$	The ratio of timing duration of the primary opening phase to the closing phase.	$\frac{T_{o1}}{T_c}$
	6	SQ$_2$	The ratio of timing duration of the secondary opening phase to the closing phase.	$\frac{T_{o2}}{T_c}$
	7	AQ	The ratio of timing duration of the closing phase to the length of the glottal cycle.	$\frac{T_c}{T}$
	8	CIQ	The ratio of the peak-to-peak amplitude of the glottal flow to the minimum peak of the pulse derivative.	$\frac{A_{ac}}{A_{d\ \min}}$
	9	NAQ	Normalized AQ by dividing it by the length of the glottal cycle duration.	$\frac{AQ}{T}$
GLF	10	PSP	Fits a second-order polynomial to the glottal flow spectrum on a logarithmic scale computed over a single glottal cycle.	Refer to [4]
	11	DH12	Difference of the first and second harmonics in decibels.	$HI\text{-}H2$
	12	HRF	The ratio of the sum of harmonics magnitude above the first harmonic (HI) to the magnitude of the first harmonic.	$\frac{\Sigma_{k>2} H_k}{H_1}$

TABLE II

MANOVA and ANOVA Analysis on the Subcategory Features for Both Male and Female Adolescents

Category	Sub-category features[a]	No. of feature coeff.	Significance (male)			Significance (female) – In all interactions of EPI, PSI and FCI, all the feature sub-categories are used; i.e. 186 coefficients from all the features for each interaction
			EPI	PSI	FCI	
TEO	TEO-CB-Auto-Env	45	+	+	+	
Cepstral (C)	MFCC	36	+	+	+	
Prosodies (P)	F_0	3	−	+	+	
	LogE	3	+	+	+	
	FMTS & FBWS	18	+	+	+	
	Jitter	3	−	+	−	
	Shimmer	3	+	+	+	
Spectral (S)	Centroid	3	+	−	−	
	Flux	3	+	+	+	
	Entropy	3	+	+	+	
	Roll-off	3	+	+	+	
	PSD	27	+	+	+	
Glottal (G)	GLT	27	+	+	+	
	GLF	9	+	+	+	
	Total	186	180	183	180	

[a] All features include their delta (Δ) and delta-delta (Δ-Δ)

TABLE III

TEO Category Classification Performance Using SBCCA With 0.5 min Test Utterances on GIM and GDM—Sensitivity and Specificity Results

Training & Testing feature: TEO							
Modeling Strategy		**EPI**		**PSI**		**FCI**	
		Sensitivity	**Specificity**	**Sensitivity**	**Specificity**	**Sensitivity**	**Specificity**
GIM (Male & Female)		74.59	69.52	51.17	81.43	53.42	82.26
GDM	Male	76.39	82.97	81.39	71.57	62.22	84.89
	Female	76.95	75.00	84.71	64.77	75.65	60.23

TABLE IV

Classification Performance of Prosodic, Spectral, and Glottal Feature Categories Using SBCCA With 1 min Test Utterances

Training/Testing Features		**Overall Accuracy %**					
		EPI		**PSI**		**FCI**	
		MALE	**FEMALE**	**MALE**	**FEMALE**	**MALE**	**FEMALE**
P		59.83	66.60	50.87	67.33	58.87	62.28
S		61.26	64.96	51.64	58.08	51.07	69.38
G		59.59	71.91	74.56	62.77	66.03	73.46
P + s		61.95	69.61	58.31	68.22	57.91	65.99
P	+ G	60.65	69.19	65.96	65.56	64.90	68.55
s	+ G	62.65	71.67	54.21	66.69	57.77	70.01
P + s	+ G	66.50	72.40	67.18	75.25	69.10	70.41

TABLE V

Classification Performance of Feature Categories Proposed in [41] Using SBCCA With 1 min Test Utterances

Training/Testing Features		Overall Accuracy %					
		EPI		PSI		FCI	
		MALE	FEMALE	MALE	FEMALE	MALE	FEMALE
P_o		57.44	68.79	58.02	61.70	61.61	64.71
$P_o + V_o$		59.83	66.60	50.87	67.33	58.87	62.28
P_o	$+ G_o$	66.64	76.54	64.68	64.76	65.25	66.26
$P_o + V_o$	$+ G_o$	60.65	69.19	65.96	65.56	64.90	68.55

TABLE VI

Influence of TEO Category in Percentage Accuracy Improvement and Their Statistical Significance (Compared to Table IV) When Added to Prosodic, Spectral, and Glottal Categories

Training/Testing Features		**Overall Accuracy increase (+) / decrease (−)**					
		EPI		**PSI**		**FCI**	
		MALE	**FEMALE**	**MALE**	**FEMALE**	**MALE**	**FEMALE**
P	+ TEO	**+21.60%**	+2.43%	**+31.35%**	−0.92%	**+25.13%**	+4.61%
S	+ TEO	**+17.32%**	+0.82%	**+27.83%**	+4.30%	**+29.29%**	−5.08%
G	+ TEO	**+22.41%**	+7.07%	**+6.80%**	+4.73%	**+13.37%**	−5.90%
P + S	+ TEO	**+15.67%**	+1.19%	**+11.35%**	−6.34%	**+19.17%**	−0.79%
P + G	+ TEO	**+19.00%**	−0.90%	**+17.68%**	+7.14%	**+15.75%**	+0.62%
S + G	+ TEO	**+16.93%**	−0.30%	**+24.83%**	+2.65%	**+19.99%**	−2.38%
P + S + G	+ TEO	**+13.75%**	−2.01%	**+14.47%**	−0.38%	**+10.59%**	−0.76%

[*] Accuracies highlighted in bold indicate the McNemar’s test results for the statistically significant accuracy increments ($p<0.05$).

TABLE VII

TEO Category Classification Performance Using SBCCA With 1 min Test Utterances—Sensitivity, Specificity, and Overall Accuracy Results

Training/ Testing Features: TEO	Event Planning Interaction (EPI)					
	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
	81.67	81.04	81.36	80.64	77.27	78.87
	Problem Solving Interaction (PSI)					
	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
	86.94	78.98	82.96	81.38	70.02	75.70
	Family Consensus Interaction (FCI)					
	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
	80.83	92.45	86.64	72.08	71.94	72.01

TABLE VIII

OPSVM Classification Results for TEO Feature Category BASED ON SBCCA

Event Planning Interaction (EPI)						
SVM weights	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Equal	100	46.15	73.08	58.33	81.82	70.08
Optimal	83.50	73.36	78.43	78.75	81.82	80.29
Problem Solving Interaction (PSI)						
SVM weights	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Equal	93.88	30.77	62.33	54.17	81.82	68
Optimal	75.10	70.24	72.67	58.33	90.91	74.62
Family Consensus Interaction (FCI)						
SVM weights	Male			Female		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Equal	77.78	84.62	81.2	70.83	68.18	69.51
Optimal	77.78	92.31	85.05	75	81.82	78.41