# DETECTING DEPRESSION: A COMPARISON BETWEEN SPONTANEOUS AND READ SPEECH

*Sharifa Alghowinem*[1,5], *Roland Goecke*[2,1], *Michael Wagner*[2], *Julien Epps*[3], *Michael Breakspear*[4,3], *Gordon Parker*[3]

[1]Australian National University, Canberra, Australia
[2]University of Canberra, Canberra, Australia
[3]University of New South Wales, Sydney, Australia
[4]Queensland Institute of Medical Research, Brisbane, Australia
[5]Ministry of Higher Education: Kingdom of Saudi Arabia

sharifa.alghowinem@anu.edu.au, Roland.Goecke@ieee.org, michael.wagner@canberra.edu.au, j.epps@unsw.edu.au, mjbreaks@gmail.com, g.parker@blackdog.org.au

## ABSTRACT

Major depressive disorders are mental disorders of high prevalence, leading to a high impact on individuals, their families, society and the economy. In order to assist clinicians to better diagnose depression, we investigate an objective diagnostic aid using affective sensing technology with a focus on acoustic features. In this paper, we hypothesise that (1) classifying the general characteristics of clinical depression using spontaneous speech will give better results than using read speech, (2) that there are some acoustic features that are robust and would give good classification results in both spontaneous and read, and (3) that a 'thin-slicing' approach using smaller parts of the speech data will perform similarly if not better than using the whole speech data. By examining and comparing recognition results for acoustic features on a real-world clinical dataset of 30 depressed and 30 control subjects using SVM for classification and a leave-one-out cross-validation scheme, we found that spontaneous speech has more variability, which increases the recognition rate of depression. We also found that jitter, shimmer, energy and loudness feature groups are robust in characterising both read and spontaneous depressive speech. Remarkably, thin-slicing the read speech, using either the beginning of each sentence or the first few sentences performs better than using all reading task data.

***Index Terms***— Mood detection, clinical depression, voice feature classification, affective sensing.

## 1. INTRODUCTION

Clinical (or major) depression is different from feeling depressed. It is generally acknowledged to be more serious, lasts for long periods and affects a person's functioning. At its most severe, depression is associated with half of all suicides and presents a significant economic burden [1]. For example, more than 180 Australians take their lives in depression related suicide each month [2].

Moreover, current depression diagnosis is limited by assessment methods that rely almost exclusively on patient self-report and clinical judgements of symptom severity [3], risking a range of subjective biases. Recent developments in affective sensing technology will potentially enable an objective assessment. Our goal here is to investigate the general characteristics of depression, which we hope will lead to an objective affective sensing system that assists clinicians in their diagnosis and monitoring of clinical depression. Ultimately, we hope to assist patients with depression to monitor the progress of their illness in a similar way that a patient with diabetes monitors their blood sugar levels with a small portable device.

The main contribution of this paper is an investigation into the discriminative power of read versus spontaneous speech (in an interview / conversation) for the task of detecting depression. We examine the performance of various acoustic features using SVM for classification on a real-world clinically validated dataset of 30 patients with severe depression and 30 healthy control subjects. We also investigate how much speech data is required to give accurate results.

## 2. RELATED WORK

Research into potential bio-markers of central nervous system disorders, e.g. affect and mood disorders, has explored subtle changes in speech characteristics as possible physiologically-based indicators of disease progression, severity or treatment efficacy [3]. Depression patterns within speech have been recognised for many years, with differences found in the pitch, loudness, speaking rate, and articulation [3]. Early studies investigating the vocal affect of depression found

that depressed subjects have a lower dynamic range of the fundamental frequency than normal subjects [4].

Formants are a widely used feature in affective computing [5] and also a significantly distinguishable feature for depression [6, 7] due to the psycho-motor retardation as a symptom for depression leading to a tightening of the vocal tract, which tends to affect formant frequencies [8]. There is evidence that sadness and depression are associated with a decrease in loudness [9], resulting in lower loudness for depressed. Since the loudness is intimately related to sound intensity, both features will be investigated here. Jitter and shimmer voice features have been analysed for depression, finding higher jitter in depression caused by the irregularity of the vocal fold vibrations [9]. On the other hand, shimmer is lower for depressed [10]. Like the jitter feature, the harmonic-to-noise (HNR) feature is higher for depressed due to the patterns of air flow in the speech production differing for depressed and healthy controls [11]. Voice energy is also a widely used distinguishing feature for depression, giving lower energy for depressed patients caused by the glottal pulses. Finally, the pitch features, which have been widely investigated, show a lower range of fundamental frequency (F0) in depressed [3, 12, 13, 14], which increases after treatment [4]. The lower range of F0, indicate a monotone speech [15] and its low variance indicate a lacking of significant expression in depression [7].

The automatic detection of depression using affective sensing techniques has been investigated lately [16, 17, 18]. While psychology investigations are concerned with the overall patterns of speech using statistical measurements based on functionals from speech prosody, affective sensing approaches classify frame-by-frame using low-level features extracted from speech. The automatic classification from the low-level features results was significant for several features. The first 3 formants gave good classification results (in terms of agreement with clinical opinion) in [17], as well as energy and loudness in [18]. Pitch or F0 classification results were not as good as expected in speaker-independent classification [17, 18], but performed well when comparing data of the same person after treatment [16]. HNR, jitter and shimmer features gave moderate results in [18], though more investigation is needed. Recently, [17] investigated depressed read speech from voiced frames and found that mel-frequency cepstral coefficients (MFCC) and spectral centroid amplitudes were good discriminating features for speaker dependent and independent depression recognition. In previous work, we investigated spontaneous speech and found that MFCC, energy and intensity features gave high recognition rates [18].

In this study, we compare read and spontaneous speech for the recognition of depression. In our dataset (Section 3), the read speech contains emotional sentences and the spontaneous speech contains responses to interview questions designed to elicit emotional responses. Our hypotheses for this study are: (1) correctly classifying clinical depression using spontaneous speech will give more accurate results than using read speech; (2) there are some acoustic features that are robust and give good classification results in both spontaneous and read speech (based on the literature of the depressed speech properties mentioned earlier); and (3) that a 'thin-slicing' approach using different smaller parts of speech data will perform similarly if not better than using the whole speech data. This is based on the physiological slicing theory, where it has been indicated that using a brief observation or 'thin slice' of behaviour can be used to predict the physiological outcome at levels above that expected by chance [19].

In the remainder of the paper, Section 3 describes the methodology, including the dataset, feature extraction and classification methods. Section 4 presents the results and discussion. The conclusions are presented in Section 5.

## 3. METHODOLOGY

### 3.1. Data Collection

For the experiments, we used real-world data collected in an ongoing study at the Black Dog Institute, a clinical research facility in Sydney (Australia) focussing on research in depression and bipolar disorder. Subjects included healthy controls as well as patients who had been diagnosed with severe depression (HAM-D $>$ 15), but no other mental disorders or medical conditions. Control subjects were carefully selected to have no history of mental illness and to match the depressed subjects in age and gender. The audio-video experimental paradigm contains several parts, including a read sentences task and an interview with the subjects [20]. The reading task contained 20 sentences with negative and positive meaning (e.g. "She gave her daughter a slap." , "She gave her daughter a doll."). The interview was conducted by asking specific open questions (in 8 question groups), where the subjects were asked to describe events that had aroused significant emotions. The audio data has been used by [17] to investigate read speech and also in our previous study [18] analysing spontaneous speech. In this paper, we compare the recognition results from different parts of the reading tasks with different parts and durations from the interview.

To date, data from over 40 depressed subjects and over 40 healthy controls (age range 21-75yr, both females and males) has been collected. Before participating, each subject was invited to complete a 'pre-assessment booklet' (general information, e.g. health history), then assessed by trained researchers following the DSM-IV diagnostic rules. Participants who met the criteria for depression were selected. Data were acquired after obtaining informed consent from the participants in accordance with approval from the local institutional ethics committee.

In this paper, only a subset of 30 depressed and 30 control subjects were analysed to achieve a gender balance in each cohort. Only native English speaking participants were selected in this study to reduce the variability that might occur from

**Table 1**. Weighted Average Recall (in %) for Acoustic Feature Classification for Different Parts of Speech

| Speech Subset | | Spontaneous Speech | | | Read Speech | | |
|---|---|---|---|---|---|---|---|
| | | Part of Each Question | "Good News" Question | "Sadness Characteristic" Question | All Read Sentences | First few Sentences | Part of Each Sentence |
| Duration (min) | | 40 | 28 | 40 | 40 | 28 | 28 |
| Pitch | F0 | 63.27 | 57.92 | 60.01 | 65.14 | 56.27 | 61.43 |
| | Voice Probability | 60.17 | 68.77 | 60.08 | 59.58 | 55.36 | 60.07 |
| MFCC | MFCC | 69.56 | 68.82 | 66.94 | 56.77 | 61.13 | 56.25 |
| | MFCC, $\Delta$, $\Delta\Delta$ | 70.08 | 71.00 | 73.12 | 65.77 | 51.81 | **73.95** |
| Energy | Log energy | 74.99 | **77.42** | **78.34** | 63.72 | 57.78 | 66.90 |
| | RMS energy | 70.08 | 70.24 | 74.05 | 63.30 | 60.96 | 66.88 |
| Intensity | Intensity | 65.96 | 76.49 | 66.72 | 58.30 | 65.64 | 62.38 |
| | Loudness | 74.87 | 64.13 | 76.57 | 59.95 | 67.75 | 69.29 |
| Formants | 3 Formants | 58.31 | 63.99 | 73.24 | 53.37 | 58.52 | 67.72 |
| Voice Quality | Jitter | **76.55** | 70.14 | 68.94 | 64.79 | 62.95 | 62.44 |
| | Shimmer | 61.63 | 75.84 | 62.51 | **67.56** | **75.84** | 70.52 |
| | Voice Quality | 66.61 | 64.13 | 66.78 | 50.00 | 53.13 | 62.35 |
| | HNR | 66.61 | 71.53 | 66.72 | 62.01 | 63.39 | 59.92 |
| Average | | 67.59 | 69.26 | 68.77 | 60.79 | 60.81 | 64.62 |

different accents. For depressed subjects, the level of depression was a selection criterion, ranging from 13-26 points, with a mean of 19 points of the diagnoses using DSM-IV (where 11-15 points refer to a "Moderate" level, 16-20 points to a "Severe" level, and $\geq$ 21 points to a "Very Severe" level).

We acknowledge that the amount of data used here is relatively small, but this is a common problem [4, 7] in similar studies. As we continue to collect more data, future studies will be able to report on a larger dataset.

### 3.2. Data Preparation

The reading and interview parts were manually labelled to extract pure subject speech. The total pure speech duration for the reading task was almost 40min, while it was 290min for pure spontaneous speech. To have a comparable duration between read and spontaneous speech, we used all read sentences with: (1) the first part of the answer to each question (on average, the first 5s of each of the 8 questions per subject), (2) part of a particular question that has both positive and negative emotions: "Do you get a *characteristic feeling* when you're sad or down and what do you normally do to *cheer yourself up*?". For simplicity, this question will be referred to as the "Sadness characteristic" question. To test the thin-slicing theory, we used a shorter duration of read speech with a question that in a previous study [18] gave the best recognition results. The shorter question is "Can you recall some recent *good news* you had and how did that make you feel?". For simplicity, this question will be referred to as the "Good News" question. We compare it with a similar duration from the reading part by using: (1) the first few sentences, (2) part of each sentence (on average the first 1.4s of each sentence per subject).

### 3.3. Feature Extraction

Voice features could be divided into two categories: Acoustic and linguistic features [21]. However, since we are aiming to find general characteristics for depressed speech regardless of the language used, linguistic features are not being analysed here. Acoustic features could also be categorised into two branches: Low-Level descriptors (LLD), which could be calculated frame-by-frame, and statistical features, which could be calculated based on the LLD over certain units (e.g. words, syllables, sentences, etc.).

Several software tools are available for extracting sound features. Here, we used the open-source software "openSMILE" [22] to extract several LLD features and some functional features from the subject speech labelled intervals (Table 1). The frame size was set to 25ms at a shift of 10ms and using a Hamming window.

### 3.4. Classification and Evaluation

The read and spontaneous speech were classified in a binary speaker-independent scenario (i.e. depressed/non-depressed) using Support Vector Machines (SVM), which are considered current state-of-the-art classifiers since they provide good generalisation properties [23]. To improve the accuracy of the SVM, the cost and gamma parameters were optimised via a wide range grid search for the best parameters using LibSVM [24]. To mitigate the effect of the limited amount of data, a leave-one-out cross-validation was used, without any overlap between training and testing data. That is, 59 different subjects were used in each turn to create a model, which the remaining subject in each turn then was tested against to ensure a valid evaluation [23].

For dimensionality reduction, Gaussian Mixture Models (GMM) with 16 mixture components were created for each subject. The Hidden Markov Model Toolkit (HTK) was used to implement a HMM using only one state to train the GMM

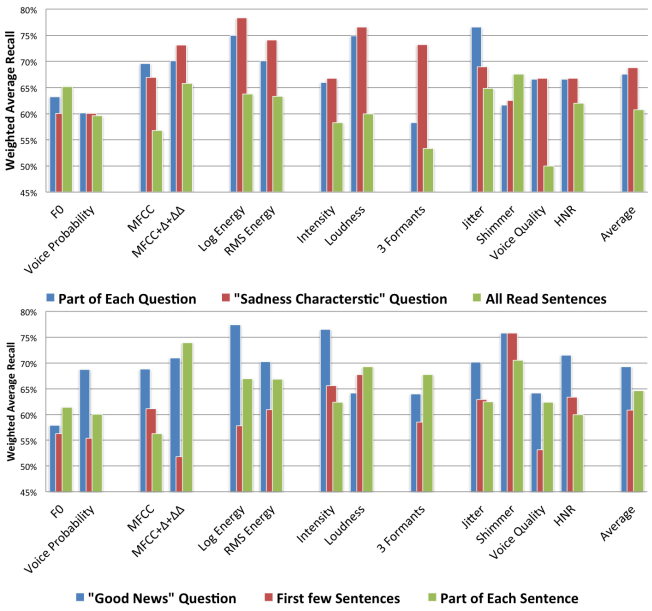


**Fig. 1**. Comparison of WAR results for different parts of speech when using 40min (top) and 28min (bottom) of data

models. The number of mixtures was fixed to ensure consistency in the comparison, acknowledging that some features benefit from more detailed modelling. This approach was beneficial to get the same number of values of the extracted features that formed the input to the SVM regardless of the duration of the subject's speech. The means, variance and weight for the 16 mixtures of GMM formed the super-vector to the SVM classifier.

To measure the performance of the system, several statistical methods could be calculated, such as recall or precision [23]. In this study, the weighted average recall (WAR) was computed and weighted using the duration (see the header of Table 1), in order to give a better comparison.

## 4. RESULTS

Examining the results for the different acoustic features extracted from spontaneous and read speech data, we consider our earlier hypotheses as confirmed. Table 1 and Figure 1 show the classification results of both read and spontaneous speech using different parts and amounts of speech data.

On the first hypothesis: classifying the general characteristic of clinical depression using spontaneous speech gives better results than using read speech. As shown in Table 1, regardless of the duration or part of speech, the overall recognition rate using spontaneous speech was higher than for read speech, indicating that spontaneous speech contains more relevant information about the subject's general characteristics, including their affective state. However, shimmer and F0 features were giving either similar or better results from read speech data for both the 40min and 28min total duration examined. On the other hand, examining the 28min duration results, loudness in read speech performed slightly better than in spontaneous speech.

For the second hypothesis – there are some acoustic features that are robust and that give good classification results in both spontaneous and read speech – we found that jitter, mel-frequency cepstral coefficients (MFCC) with its velocity ($\Delta$) and acceleration ($\Delta\Delta$), and energy (both log and root mean square energy (RMS)) were the common feature groups that gave high WAR results in both spontaneous and read speech, while F0, voicing probability and voice quality (F0 quality) were the worst. The MFCC feature group gave relatively good results for classifying depression in spontaneous speech; however, MFCC with its deltas performed slightly better than using MFCC alone in most of the speech parts. This finding is confirmed with what was found in [17], analysing read speech from the same Black Dog data set. Although formants are a widely used feature in the affect literature [25], their results were not good in either read or spontaneous speech in most of the speech part. However, the formants recognition rate using the "Sadness Characteristic" question, and using part of each sentence, performed better than other parts of speech. That might be caused by the variation of mixed positive and negative emotions in those particular parts of the data. Note that significance tests could not be carried out to compare our results, as we are dealing with decision labels to measure the performance of our system.

Remarkably, while testing the thin-slicing hypothesis by comparing several read speech parts, when using the beginning of each sentence, depression recognition was on average better than using the first few sentences or all sentences. That finding applies to all 13 features except F0, MFCC, intensity, shimmer and jitter, which gave slightly better or similar results. This may indicate that depressed subjects express their depression more strongly at the beginning of utterances before they got involved with the task [26]. On the other hand, slicing the spontaneous speech on average was not giving dramatic differences in depression recognition rate.

## 5. CONCLUSIONS

We have presented work aiming at an objective diagnostic aid supporting clinicians in their diagnosis of depression. The results confirmed our hypotheses by examining and comparing subjects' acoustic features using read and spontaneous speech. We found that using spontaneous speech gave a better result than using read speech for most features. We also found that jitter, shimmer, energy and loudness feature groups were robust in getting general characteristic of depressive speech. Remarkably, we found that the beginning of each sentence in the reading task gives better results than using all reading task acoustic features, indicating that diagnosing depression may be better before the depressed subjects engage in the task.

## 6. REFERENCES


[1] Y Lecrubier, "Depressive illness and disability," *European Neuropsychopharmacology*, vol. 10 Suppl 4, pp. S439–S443, 2000.

[2] ABS Australian Bureau of Statistics, *Causes of death 2006*, Number 3303.0. ABS: Canberra, 2008.

[3] James C Mundt, Peter J Snyder, Michael S Cannizzaro, Kara Chappie, and Dayna S Geralts, "Voice acoustic measures of depression severity and treatment response collected via interactive voice response technology," *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[4] A. Ozdas, R.G. Shiavi, S.E. Silverman, M.K. Silverman, and D.M. Wilkes, "Analysis of fundamental frequency for near term suicidal risk assessment," *IEEE Conf. Systems, Man, Cybernetics*, pp. 1853–1858, 2000.

[5] Shashidhar G Koolagudi and K Sreenivasa Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 37–40, 2011.

[6] Alistair J. Flint, Sandra E. Black, Irene Campbell-Taylor, Gillian F. Gailey, and Carey Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 309–319, July 1993.

[7] Elliot Moore, Mark Clements, John W Peifer, and Lydia Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech.," *IEEE Transactions on Bio-medical Engineering*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

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

[9] K R Scherer, *Vocal assessment of affective disorders*, pp. 57–82, Lawrence Erlbaum Associates, 1987.

[10] Ana Nunes, Lidia Coimbra, and Antonio Teixeira, "Voice quality of european portuguese emotional speech corresponding author," *Computational Processing of the Portuguese Language Lecture Notes in Computer Science*, vol. 6001/2010, pp. 142–151, 2010.

[11] Lu-Shih Alex Low, Namunu C Maddage, Margaret Lech, Lisa B Sheeber, and Nicholas B Allen, "Detection of clinical depression in adolescents speech during family interactions," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 574–586, 2011.

[12] A Nilsonne, "Speech characteristics as indicators of depressive illness," *Acta Psychiatrica Scandinavica*, vol. 77, no. 3, pp. 253–263, 1988.

[13] S Kuny and H H Stassen, "Speaking behavior and voice sound characteristics in depressive patients during recovery," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 289–307, 1993.

[14] H Ellgring and K R Scherer, "Vocal indicators of mood change in depression," *Journal of Nonverbal Behavior*, vol. 20, no. 2, pp. 83–110, 1996.

[15] Elliot Moore, Mark Clements, John Peifer, and Lydia Weisser, "Comparing objective feature statistics of speech for classifying clinical depression," *Proc. 26th Ann. Conf. Eng. Med. Biol.*, vol. 1, pp. 17–20, Jan. 2004.

[16] Jeffrey F. Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre, "Detecting depression from facial actions and vocal prosody," *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–7, Sept. 2009.

[17] Nicholas Cummins, Julien Epps, Michael Breakspear, and Roland Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in *Proc. Interspeech 2011*, 2011, pp. 2997–3000.

[18] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Julien Epps, Michael Breakspear, and Gordon Parker, "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech," in *Proc. FLAIRS-25*, 2012, pp. 141–146.

[19] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis.," *Psychological Bulletin*, vol. 111, no. 2, pp. 256, 1992.

[20] Gordon McIntyre, Roland Goecke, Matthew Hyett, Melissa Green, and Michael Breakspear, "An Approach for Automatically Measuring Facial Activity in Depressed Subjects," in *Proc. ACII2009*, 2009, pp. 223–230.

[21] Tim Polzehl, Alexander Schmitt, Florian Metze, and Michael Wagner, "Anger recognition in speech using acoustic and linguistic cues," *Speech Communication*, vol. 53, no. 910, pp. 1198 – 1209, 2011.

[22] Florian Eyben, Martin Wöllmer, and Björn Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM'10)*, Oct. 2010, pp. 1459–1462.

[23] Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. February, pp. 1062–1087, 2011.

[24] C C Chang and C J Lin, "Libsvm: a library for svm," *Computer*, pp. 1–30, 2001.

[25] Shashidhar Koolagudi and K. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, pp. 1–19.

[26] Susan Nolen-Hoeksema, "Sex differences in unipolar depression: Evidence and theory," *Psychological Bulletin*, vol. 101, no. 2, pp. 259–282, 1987.