From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech.

Conference Paper · January 2012

Sharifa Alghowinem
Australian National University

Roland Goecke
University of Canberra

Julien Epps
UNSW Sydney

Michael Breakspear
Queensland Institute of Medical Research

From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech

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

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

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

Abstract

Depression and other mood disorders are common and disabling disorders. We present work towards an objective diagnostic aid supporting clinicians using affective sensing technology with a focus on acoustic and statistical features from spontaneous speech. This work investigates differences in expressing positive and negative emotions in depressed and healthy control subjects as well as whether initial gender classification increases the recognition rate. To this end, spontaneous speech from interviews of 30 subjects of each depressed and controls was analysed, with a focus on questions eliciting positive and negative emotions. Using HMMs with GMMs for classification with 30-fold cross-validation, we found that MFCC, energy and intensity features gave highest recognition rates when female and male subjects were analysed together. When the dataset was first split by gender, root mean square energy and shimmer features, respectively, were found to give the highest recognition rates in females, while it was voice quality for males. Overall, correct recognition rates from acoustic features for depressed female subjects were higher than for male subjects. Using temporal features, we found that the response time and average syllable duration were longer in depressed subjects, while the interaction involvement and articulation rate wesre higher in control subjects.

1 Introduction

Changes in affective state are a normal characteristic of human beings. However, when these changes increase in intensity, last longer, and a person's functioning drops, a clinical depression line might be crossed. Unlike emotion, which is short term, mood is a long term affective state and, therefore, clinical depression is a mood disorder that may last for weeks, months, even years, vary in severity, and could result in unbearable pain if appropriate treatment is not received. The World Health Organization lists depression as the fourth most significant cause of suffering and disability world wide and predicts it to be the leading cause in 2020 (Mathers, Boerma, and Fat 2008). For example, the Australian Survey of Mental Health and Well-Being (1997) reported that 6.3% of the population will suffer clinical depression in any one year, noting that this percentage does not include people who choose not to get professional help. Also, statistically, six million working days are lost each year to depression and ten million antidepressant prescriptions are written every year. Unfortunately, more than 180 Australians take their lives in depression related suicide each month (Australian Bureau of Statistics 2008). Even though people of all ages suffer from depression, Australia has one of the highest youth depression related suicide rate (Prendergast 2006).Fortunately, this can be prevented if depressed subjects seek help from professionals, and if health professionals could be provided with suitable objective technology for detection and diagnosing depression (Prendergast 2006).

Depression has no dedicated laboratory tests or procedures for diagnosis. Rather, it is diagnosed as part of a complete mental health evaluation. It depends on symptoms self-report and professional observation and evaluation (Albrecht 2006). However, professionals' evaluations vary depending on their expertise and the diagnostic methods used (e.g. Diagnostic and Statistical Manual of Mental Disorders (DSM-IV,2008), Hamilton Rating Scale for Depression, Montgomery-Asberg Depression Rating Scale, etc.). Currently, there is no objective method to diagnose depression.

While automatic affective state recognition has become an active research area in the past decade, methods for mood disorder detection, such as depression, are still in their infancy. Our goal here is to investigate the voice features that give the best result for recognising depression, which may ultimately lead to an objective affective sensing system that supports clinicians in their diagnosis of clinical depression.

2 Background

Clinical depression is a serious illness. (Albrecht 2006) defines it as a medical condition that affects and changes a person's thoughts, mood, and behaviour as well as the physical health. Early studies investigating vocal affect of depression found that depressed subjects have a lower dynamic range of the fundamental frequency than normal subjects (Ozdas et al. 2000). Moreover, the study found that "the fluctuation of fundamental frequency along with verbal content was more emphasized for healthy controls and subdued for depressed subjects" (Ozdas et al. 2000). Also, depressives have a slower rate of speech and relatively monotone delivery when compared with normal speaking patterns (Moore et al. 2004; 2008). The latter also confirmed a lacking in significant expression as previously found by (Darby, Simmons, and Berger 1984) when they described the triad in depressive speech of reduced stress, monopitch, and monoloudness.

Research on the vocal indicators in depressed subjects found an increase in pause time and a decrease in speech rate in depressives (Ellgring and Scherer 1996). (Zlochower and Cohn 1996) measured the vocal timing in clinically depressed mothers in response to their 4-months-old infants and concluded that depressed mothers had longer and more variable duration of silence. They found that the response delay increases with the severity level of depression. Therefore, we will analyse not only the voice feature but also the response time and duration of speech. Applying these previous findings of the characteristics of depressed speech with the acoustic features for spontaneous depressed speech is an under researched area. We investigate whether certain features can give better depression recognition. Recently, (Cummins et al. 2011) investigated depressed speech from read material and found that Mel-frequency cepstral coefficients (MFCC) and spectral centroid amplitudes were good discriminating features for speaker dependent and independent depression recognition.

In this paper, we look for a general characteristic for depressed spontaneous speech by examining, which acoustic features or feature groups can give better recognition, and whether these features give better results taking the subject's gender into account. We also examine duration and speech rate features to discriminate depressed speech.

3 Methodology

3.1 Data Collection

For the experimental validation, we use data collected in an ongoing study at the Black Dog Institute, a clinical research facility in Sydney, Australia, offering specialist expertise in depression and bipolar disorder. Subjects include healthy controls as well as patients who have been diagnosed with pure depression, i.e. those who have no other mental disorders or medical conditions. Control subjects are carefully selected to have no history of mental illness and age and gender match the depressed subjects. The experimental paradigm contains several parts, including an interview with the subjects (McIntyre et al. 2009). The interview is conducted by asking specific questions (in 8 question groups), where the subjects are asked to describe events that had aroused significant emotions. In this paper, the interview part with all 8 question groups is used for analysing spontaneous depressive speech. We also compare the expression of positive and negative emotions by analysing two related questions from the interview: "Can you recall some recent *good news* you had and how did that make you feel?" and "Can you recall news of *bad or negative nature* and how did you feel about it?" For simplicity, these two questions will be referred to as "Good News" and "Bad News", resp. We assume that those questions elicit the emotions, even though the answers were not validated for certain emotions.

Table 1: Duration (mins) of Depressed and Control speech

	All gender		Male Only		Female Only	
Questions	Depressed	Control	Depressed	Control	Depressed	Control
All 8 questions	183.15	107.68	71.00	62.59	112.15	45.09
"Good News"	17.04978333	10.40	6.89	5.01	10.16	5.39
"Bad News"	23.90	16.20	9.26	11.01	14.65	5.19

3.2 Participants

To date, data from over 40 depressed subjects with over 40 age-matched 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 interviewed by trained researchers following the DSM-IV diagnostic rules. Participants who met the criteria for depression were selected.

In this paper, a subset of 30 depressed subjects and 30 controls were analysed, with equal gender balance. Only native English speaking participants were selected in this research, to reduce the variability that might occur from different accents. For depressed subjects, the level of depression was a selection criterion, with a mean of 19 points of the diagnoses using DSM-IV (range 14-26 points, where 11-15 points refer to a "Moderate" level, 16-20 points to a "Severe" level, and ≥ 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 (Ozdas et al. 2000; Moore et al. 2008). As we continue to collect more data, future studies will be able to report on a larger dataset.

3.3 Data Preparation

The interview part was manually labelled to separate questions and speakers. Within the questions, the speakers were manually labelled with a focus on the lag between the interview asking the question and the participant answering it to measure the response time for depressed and non-depressed subjects. In addition, the duration of the overlap between speakers was labelled to measure the involvement style. The duration of subjects' laughs was labelled for further investigation, as well as the duration and the number of the interviewer's interactions to elicit speech from the participants. Out of a total 513min of interviews (313min for depressed and 200min for controls), the duration of "pure" subject speech (without silence) for depressed and control subjects for all 8 questions, the "Good News" question and the "Bad News" question used in this paper is shown in Table 1.

3.4 Feature Extraction

Speech features can be acquired from both uttered words (linguistic) and acoustic cues (para-linguistic). However, linguistic features including word choices, sentence structure etc. are beyond the scope of this study. We would also like to generalise the findings to other languages in the future.

Acoustic Features. In general, the more relevant features to recognise affect are considered to be duration, MFCC, energy and pitch variation (Batliner and Huber 2007). This view has been supported by a study that found the most relevant acoustic features to emotions are duration and energy, while all other features are of medium relevance (Schuller et al. 2007). A study on communication difficulties found that among duration, energy, and F0 features, duration was the feature that contributed most to classification (Batliner et al. 2003). Using all acoustic features together gives better results than single features (Schuller et al. 2007).

Several software tools are available for extracting and extracting sound features. In this work, we used the publicly available open-source software "openSMILE" (Eyben, Wöllmer, and Schuller 2010) to extract several low level voice features and functional features from the subject speech labelled intervals (Table 2). The frame size is set to 25ms at a shift of 10ms and using a Hamming window.

Table 2: Weighted Average Recall and F1 measures (in %) for acoustic feature classification from all 8 questions

Feature Group	All gender				Male Only				Female Only			
Feature	WAR	$F1_D$	$F1_C$	avg.F1	WAR	$F1_D$	$F1_C$	avg.F1	WAR	$F1_D$	$F1_C$	avg.F1
Pitch												
F0	37	0	54	27	49	36	57	46	29	0	45	22
F0 row	68	78	41	59	65	74	47	61	69	80	33	57
Voice prob.	58	66	47	56	37	40	33	37	65	73	49	61
MFCC												
MFCC	67	72	60	66	57	60	54	57	67	74	53	64
MFCC,Δ,$\Delta\Delta$	67	76	48	62	46	61	10	36	72	80	53	67
Energy												
root mean	70	80	36	58	62	72	40	56	77	86	33	60
log energy	69	74	61	68	67	68	65	67	71	79	57	68
Intensity												
Loudness	65	78	13	45	56	71	13	42	71	83	0	42
Intensity	70	80	36	58	62	72	40	56	77	86	33	60
Formants												
3 Formants	63	77	0	39	53	69	0	35	71	83	0	42
voice quality												
Jitter	64	78	6	42	53	68	12	40	71	83	0	42
Shimmer	68	80	24	52	59	69	38	53	77	86	33	60
voice quality	71	81	40	60	69	77	50	64	76	85	44	65
HNR	60	70	40	55	64	70	55	63	62	69	50	60
Average	64	71	36	53	57	64	37	51	68	75	35	55

Statistical Measures. Duration features were extracted from the manually labelled intervals for further statistical analysis. Regarding speaking rate, a Praat (Boersma and Weenink 2009) script by (De Jong and Wempe 2009) was used to calculate the speech and articulation rates as well as the pause rate. When measuring the speech rate, pauses are included in the duration time of the utterance, while the articulation rate excludes pauses (Goldman-Eisler 1968).

3.5 Classification and Evaluation

The spontaneous speech was classified in a binary speaker-independent scenario (i.e. depressed/non-depressed). An initial classification of gender (from voice) can increase the accuracy of the overall affect classification, while misclassifying gender may give the opposite results (Vogt and Andre 2006). Therefore, manual gender classification was used in ours study to test the overall accuracy. Besides investigating the effect of prior gender classification, a particular comparison between expressing positive and negative emotions between depressed and control subjects was examined.

Hidden Markov Models (HMM) are a suitable and widely used way of capturing the temporal information in the acoustic features extracted from speech. Following the approach of many para-linguistic speech classification studies, the Hidden Markov Model Toolkit (HTK) was used to implement a HMM using one state to train a Gaussian Mixture Model (GMM) with 16 mixtures and 10 iterations. The choice of the number of mixtures was fixed to ensure consistency in the comparison, knowing that some features benefit more from more detailed modelling. To mitigate the effect of the limited amount of data, a 30-fold leave-one-out cross-validation was used. That is, 29 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 and prevent contaminating the results (Schuller et al. 2011). This was done for both cohorts, resp.

In order to measure the performance of the system, several statistical methods could be calculated, such as accuracy, precision, recall, F1 measure (the harmonic mean of recall and precision), Kappa, and confusion matrix (Schuller et al. 2011). Another way is by graphing the results either using Detecting Error Trade-Off (DET) or Receiver Operating Characteristic (ROC) curves. In this paper, the weighted average recall (WAR) and F1 measures were computed and weighted using the duration in Table 1.

4 Results

4.1 Acoustic Features

We evaluated the depression recognition rate with and without initial (manual) gender separation to establish the influence of gender. Table 2 shows the WAR and F1 measures for each acoustic feature analysed here w.r.t. gender.

In general, recognising depression in female subjects was better in most features with a WAR mean of 68%, while males had a WAR mean of only 57%, with a mixed gender result of a WAR mean of 64%. This result confirms previous conclusions of gender differences (Nolen-Hoeksema 1987) that depressed women may be more likely to be detected than depressed men. This might be related to the fact that women are more likely to amplify their moods, while men are more likely to engage in distracting behaviours that dampen their mood when depressed (Nolen-Hoeksema 1987). The energy and intensity feature groups were the best features for mixed gender depression classification, in line with (Batliner and Huber 2007). The energy features in particular were the best features for male depression classification (cf. (Darby, Simmons, and Berger 1984)). While most features were good for female depression classification, shimmer alone and root mean square energy were the best ones, and fundamental frequency (F0) and Harmonic-to-Noise Ratio (HNR) were the worst resp.

Moreover, the recognition results for MFCC features were slightly better with the inclusion of the first (Δ) and second ($\Delta\Delta$) order derivatives than using MFCC features by themselves, but not in a statistically significant wat. This result is consistent with (Cummins et al. 2011), which analysed data from the read sentences of the Black Dog Institute data set, while our analysis here is on the spontaneous speech data, as well as in line with (Low et al. 2009), where there was

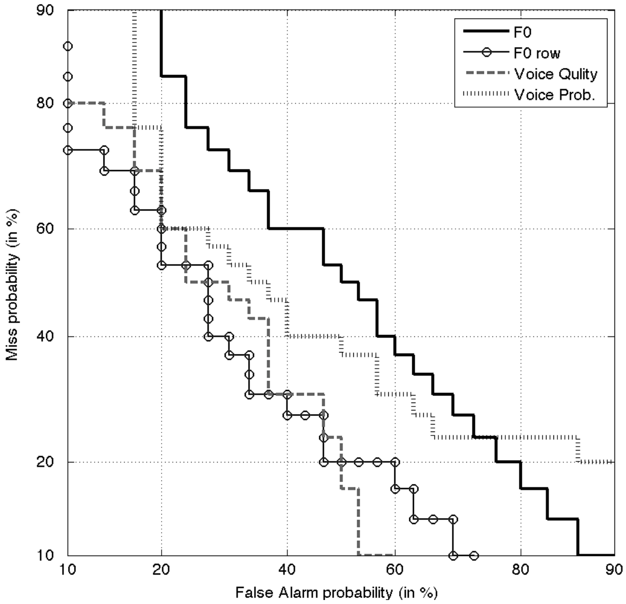

Figure 1: DET curves for the pitch features group classification results using all 8 questions for both genders combined

only a 3% increase in the accuracy of depression classification. Using raw F0, i.e. without thresholding (i.e. forcing to 0) in unvoiced segments, gives better classification results than using thresholded F0. The reason lies in the fact that the HMM gives better results for continuous data streams (Yu and Young 2011). As an example for the results of our system, Figure 1 shows DET curves for each *Pitch* feature group using all questions for both genders combined.

While acknowledging the potential impact of the huge reduction of training data from using all 8 questions to using only one question, we investigated the differences in expressing positive and negative emotions between depressed and control subjects. This was done by evaluating the "Good News" and "Bad News" questions from the spontaneous speech data. While the subjects' answers were not validated for a certain emotion, we assume that the questions elicited positive and negative emotions, resp.

For the "Good News" question (positive emotion), recognising depression was almost as accurate as when using all 8 questions, both with and without prior gender separation (Table 3). Getting good recognition rates from such a small dataset indicates the clearly noticeable differences in expressing positive emotions in depressives and controls.

On the other hand, analysing the "Bad News" question (negative emotion), gives worse recognition rates than using all 8 questions or the positive question (Table 4). This indicates that both groups express negative emotions in the same or a similar manner. Linking this finding with the previous one, we conclude that positive emotions are expressed less in depressed subjects at all times, that negative emotions dominate in depressives (Ekman 1994; Ekman and Fridlund 1987) and, hence, negative emotional speech has less discriminatory power than positive emotional speech.

Table 3: Weighted Average Recall and F1 measures (in %) for acoustic feature classification from "Good News"

Feature Group	All gender				Male Only				Female Only			
Feature	WAR	$F1_D$	$F1_C$	avg.F1	WAR	$F1_D$	$F1_C$	avg.F1	WAR	$F1_D$	$F1_C$	avg.F1
Pitch												
F0	38	0	55	27	58	73	0	37	35	0	51	26
F0 row	38	0	55	27	42	0	59	30	35	0	51	26
Voice prob.	59	64	51	58	45	41	48	44	58	62	52	57
MFCC												
MFCC	61	67	52	60	58	65	48	57	64	71	54	62
MFCC,Δ,ΔΔ	71	79	51	65	61	72	37	55	69	74	62	68
Energy												
root mean	68	78	39	59	67	75	51	63	75	84	42	63
log energy	71	76	63	70	72	77	64	70	73	78	66	72
Intensity												
Loudness	66	78	18	48	66	77	33	55	68	80	13	46
Intensity	69	79	43	61	67	75	51	63	75	84	42	63
Formants												
3 Formants	62	77	0	38	39	39	39	39	65	79	0	40
voice quality												
Jitter	62	77	0	38	60	73	22	47	65	79	0	40
Shimmer	67	79	24	51	54	70	0	35	70	81	24	52
voice quality	69	78	45	61	67	75	51	63	68	78	42	60
HNR	38	0	55	27	42	0	59	30	35	0	51	26
Average	60	59	39	49	57	58	40	49	61	61	39	50

Table 4: Weighted Average Recall and F1 measures (in %) for acoustic feature classification from "Bad News"

Feature Group	All gender				Male Only				Female Only			
Feature	WAR	$F1_D$	$F1_C$	avg.F1	WAR	$F1_D$	$F1_C$	avg.F1	WAR	$F1_D$	$F1_C$	avg.F1
Pitch												
F0	60	75	0	37	53	21	67	44	36	31	39	35
F0 row	40	0	58	29	54	0	70	35	29	12	41	27
Voice prob.	44	42	46	44	36	40	31	36	23	28	19	23
MFCC												
MFCC	53	60	43	52	53	51	55	53	63	74	32	53
MFCC,Δ,ΔΔ	54	62	41	52	43	60	0	30	39	44	31	38
Energy												
root mean	64	76	27	52	50	63	22	43	72	83	20	52
log energy	63	71	49	60	59	60	59	59	63	76	16	46
Intensity												
Loudness	64	76	27	52	54	65	32	48	67	80	18	49
Intensity	61	75	6	41	49	64	13	38	74	85	0	42
Formants												
3 Formants	59	66	48	57	46	63	0	31	50	61	29	45
voice quality												
Jitter	41	17	54	36	57	13	72	42	74	85	0	42
Shimmer	55	64	39	52	50	19	63	41	36	24	45	34
voice quality	43	59	5	32	54	65	32	48	59	74	0	37
HNR	40	0	58	29	54	0	70	35	26	0	41	21
Average	53	53	36	45	51	42	42	42	51	54	24	39

4.2 Statistical Analysis of Temportal Features

As mentioned earlier, manually labelled speaker turns were used to extract duration for statistical analyses of temporal features, namely length of subject speech, length of interviewer (RA) speech, number of RA interactions (turns), time to first response, total length of response, length of subject laughing, length of both subject and interviewer laughing, length of overlapping speech (Table 5). All statistical significance tests were one-tailed T-tests for two samples assuming unequal variances and p=0.05. Table 5 shows the probability value and the direction of effect (DIR.) to indicate which group (depressed -D- or control -C-) has a stronger effect.

- *First response time* – The duration of the silence after asking a question until an acknowledgement indicated by any sounds or words that are not the actual answer for the question (e.g. "ahhh","hmm", "well", etc.) was longer in depressed subjects, especially in depressed males, in line with (Zlochower and Cohn 1996; Ellgring and Scherer 1996). While we measured the first response time manually, it could be measured automatically using speaker

Table 5: T-test for speech duration for all 8 Questions

Feature	All gender P value	DIR.	Male Only P value	DIR.	Female Only P value	DIR.
Subject Speech	0.10	D>C	0.50	D>C	0.05025	D>C
RA Speech	0.28	D>C	0.17	D>C	0.48	C>D
# RA Interaction	0.001	D>C	0.12	D>C	0.00152	D>C
First Response	0.01	D>C	0.03	D>C	0.11	D>C
Total Response	0.49	D>C	0.26	D>C	0.33	C>D
Subject Laughing	0.01	C>D	0.07	C>D	0.01316	C>D
Both Laughing	0.00001	C>D	0.005	C>D	0.0001	C>D
Overlap Speech	0.000001	C>D	0.002	C>D	0.00003	C>D

diarization techniques.

- *Total response time* – Differences in the lag between asking the question and the actual answer were not statistically significant between the two groups.
- *Duration of the RA speech* – No statistically significant difference between two groups, but the *Number of RA interactions* was higher for depressives to encourage them to speak more, which may be the reason for having longer speech duration for depressed subjects (*Subject Speech*).
- *Laughs duration* – Measured to indicate a positive reaction. Depressives laughed less, especially females.
- *Duration of overlapping speech* – The involvement by depressed subjects was less than by the controls.

The T-test for the "Good News" question alone showed comparable results to when all 8 questions were used. In contrast, the "Bad News" T-test showed that there were almost no statistically significant differences between depressed and controls, which supports the finding in Sec. 4.1. There were two differences between using all 8 questions and the "Good News" question:

- Longer total response time in depressed subjects in the "Good News" question.
- A significant difference in laugh duration between depressed and control subjects when using all questions, but not when only using the "Good News" question.

4.3 Speech Rate Features

Speech rate, articulation rate and pause rate were extracted using Praat (De Jong and Wempe 2009). T-tests were applied to the results to indicate statistically significant differences between the two groups. All T-tests were one-tailed for two samples assuming unequal variances and p=0.05. Table 6 shows the probability value and the direction of effect (DIR.) to indicate which group (depressed -D- or control -C-) has a stronger effect.

- *Average syllable duration* (total duration-pauses duration / # syllables) was longer for the depressed group, especially females, which indicates that depressives speak slower than controls (in line with (Moore et al. 2004)).
- *Articulation rate* (#syllables / total duration-pauses duration) was lower in depressives, especially females (in line with (Pope et al. 1970)).
- *Speech rate* (#syllables / total duration) was not significantly different.

Table 6: T-test for speech rate features for all questions

Feature	All gender P value	DIR.	Male Only P value	DIR.	Female Only P value	DIR.
Speech Rate	0.284	C>D	0.271	D>C	0.099	C>D
Articulation Rate	0.056	C>D	0.013	D>C	0.003	C>D
Average Syllable Dur.	0.036	D>C	0.018	C>D	0.012	D>C
Pause Rate	0.093	C>D	0.149	C>D	0.164	C>D
Average Pause Dur.	0.201	C>D	0.065	C>D	0.424	C>D

- *Pause rate* (#pauses / total duration) was not significantly different.

For the purpose of comparing positive and negative emotions, speech rate features were extracted for the specific related questions. For positive emotions ("Good News" question), the results from the T-test were similar to the results using all questions. There were only two differences between using all questions and "Good News" question:

- In the "Good News" question, the *pause rate* was lower in the depressed group than when using all questions, especially in males, which indicates longer pauses.
- The *Average Pause Duration* was higher in depressives with and without prior gender separation than in controls in the "Good News" question, indicating longer pauses duration.

For the negative emotion comparison, there were no significant differences between depressed and controls, which supports the previously mentioned findings that both groups express negative emotions in a similar manner.

5 Conclusions

Our aim is to work towards an objective affective sensing system that supports clinicians in their diagnosis of clinical depression. To this end, we investigated which features are better for recognising depression from spontaneous speech and whether initial gender separation influences the recognition rate. This included both acoustic and temporal features. In general, we conclude that recognising depression from female subjects was better in most acoustic features than for male subjects. Log energy and shimmer features (individually) were the best for recognising depression in females, while loudness was the best feature for depression recognition in males. For mixed genders, MFCC, energy and intensity features gave better recognition rates.

We also investigated the difference in expressing positive and negative emotions in depressed and control subjects. We found that expressing positive emotions in spontaneous speech resulted in higher correct recognition for depressed and control subjects, while there were no statistically significant differences between the cohorts in spontaneous speech related to negative emotions. Furthermore, we found that using data from the positive emotion question resulted in recognition rates almost equal to those when using all questions. However, when using only the negative emotion question, the recognition rate dropped; implying that depressed and control subjects express negative emotions in a similar manner and that the differences between the two cohorts best express themselves in the positive emotion data.

Analysing duration, we found that the response time was longer in depressed, that the interaction involvement was higher in controls, and that controls laughed more often than the depressed. With the speech rate analysis, we found that the average syllable duration was longer in depressed, especially females, and that the articulation rate was lower in depressed females, which confirms previous results that depressed subjects speak more slowly than non-depressed.

6 Limitations and Further Work

This paper is a first in a series of investigations of depression cues from the Black Dog Institute dataset. Fusing face (Saragih and Goecke 2006), body, eye features with this current research will be a next step towards multi-modal system (McIntyre and Goecke 2007). A known limitation is the fairly small number of (depressed and control) subjects. As data collection is ongoing, we anticipate to report on a larger dataset in the future.

References

Albrecht, A. T. 2006. *100 Questions & Answers About Depression*. Jones & Bartlett Learning, 1 edition.

Australian Bureau of Statistics, A. 2008. *Causes of death 2006*. Number 3303.0.

Batliner, A., and Huber, R. 2007. Speaker classification i. Berlin, Heidelberg: Springer-Verlag. chapter Speaker Characteristics and Emotion Classification, 138–151.

Batliner, A.; Fischer, K.; Huber, R.; Spilker, J.; and Nth, E. 2003. How to find trouble in communication. *Speech Communication* 40(1-2):117 – 143.

Boersma, P., and Weenink, D. 2009. Praat: doing phonetics by computer.

Cummins, N.; Epps, J.; Breakspear, M.; and Goecke, R. 2011. An Investigation of Depressed Speech Detection: Features and Normalization. In *Proc. Interspeech 2011*.

Darby, J. K.; Simmons, N.; and Berger, P. A. 1984. Speech and voice parameters of depression: a pilot study. *Journal of Communication Disorders* 17(2):75–85.

De Jong, N. H., and Wempe, T. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41(2):385–390.

Ekman, P., and Fridlund, A. J. 1987. Assesment Of Facial Behvior In Affective Disorders. In *Depression and Expressive Behavior. x*. Hillsdale, N.J: Lawrence Erlbaum. 37–56.

Ekman, P. 1994. Moods Emotions And Traits. In *P. Ekman & R. Davidson (Eds.) The Nature of Emotion: Fundamental Questions*. New York: Oxford University Press. 15–19.

Ellgring, H., and Scherer, K. R. 1996. Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior* 20(2):83–110.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. ACM Multimedia (MM'10)*, 1459–1462.

Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in spontaneous speech*. Academic Press.

Low, L.-S. A.; Maddage, N. C.; Lech, M.; and Allen, N. 2009. Mel frequency cepstral feature and Gaussian Mixtures for modeling clinical depression in adolescents. In *2009 IEEE Int. Conf. Cognitive Informatics*, 346–350.

Mathers, C.; Boerma, J.; and Fat, D. 2008. *The Global Burden of Disease: 2004 Update*. Geneva, Switzerland: WHO.

McIntyre, G., and Goecke, R. 2007. Towards Affective Sensing. In *HCII2007*, volume 3 of *LNCS 4552*, 411–420.

McIntyre, G.; Goecke, R.; Hyett, M.; Green, M.; and Breakspear, M. 2009. An Approach for Automatically Measuring Facial Activity in Depressed Subjects. In *ACII2009*, 223–230.

Moore, E.; Clements, M.; Peifer, J.; and Weisser, L. 2004. Comparing objective feature statistics of speech for classifying clinical depression. *Proc. 26th Ann. Conf. Eng. Med. Biol.* 1:17–20.

Moore, E.; Clements, M.; Peifer, J. W.; and Weisser, L. 2008. Critical analysis of the impact of glottal features in the classification of clinical depression in speech. *IEEE Transactions on Bio-medical Engineering* 55(1):96–107.

Nolen-Hoeksema, S. 1987. Sex differences in unipolar depression: Evidence and theory. *Psychol* (101):259–282.

Ozdas, A.; Shiavi, R.; Silverman, S.; Silverman, M.; and Wilkes, D. 2000. Analysis of fundamental frequency for near term suicidal risk assessment. *IEEE Conf. Systems, Man, Cybernetics* 1853–1858.

Pope, B.; Blass, T.; Siegman, A. W.; and Raher, J. 1970. Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology* 35(1):128–133.

Prendergast, M. 2006. *Understanding Depression*. VIC Australia: Penguin Group.

Saragih, J., and Goecke, R. 2006. Iterative Error Bound Minimisation for AAM Alignment. In *ICPR2006*, volume 2, 1192–1195.

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

Schuller, B.; Batliner, A.; Steidl, S.; and Seppi, D. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* 53(February):1062–1087.

Vogt, T., and Andre, E. 2006. Improving Automatic Emotion Recognition from Speech via Gender Differentiation. In *LREC2006*.

Yu, K., and Young, S. 2011. Continuous F0 Modeling for HMM Based Statistical Parametric Speech Synthesis. *IEEE Trans. Audio, Speech, and Lang. Proc.* 19(5):1071–1079.

Zlochower, A. J., and Cohn, J. F. 1996. Vocal timing in face-to-face interaction of clinically depressed and nondepressed mothers and their 4-month-old infants. *Infant Behavior and Development* 19(3):371 – 374.