Characterising Depressed Speech for Classification

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

[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.m.f@gmail.com, roland.goecke@ieee.org, michael.wagner@canberra.edu.au, j.epps@unsw.edu.au, mjbreaks@gmail.com, g.parker@blackdog.org.au

Abstract

Depression is a serious psychiatric disorder that affects mood, thoughts, and the ability to function in everyday life. This paper investigates the characteristics of depressed speech for the purpose of automatic classification by analysing the effect of different speech features on the classification results. We analysed voiced, unvoiced and mixed speech in order to gain a better understanding of depressed speech and to bridge the gap between physiological and affective computing studies. This understanding may ultimately lead to an objective affective sensing system that supports clinicians in their diagnosis and monitoring of clinical depression. The characteristics of depressed speech were statistically analysed using ANOVA and linked to their classification results using GMM and SVM. Features were extracted and classified over speech utterances of 30 clinically depressed patients against 30 controls (both gender-matched) in a speaker-independent manner. Most feature classification results were consistent with their statistical characteristics, providing a link between physiological and affective computing studies. The classification results from low-level features were slightly better than the statistical functional features, which indicates a loss of information in the latter. We found that both mixed and unvoiced speech were as useful in detecting depression as voiced speech, if not better.

Index Terms: depression, speech characteristics, mood classification

1. Introduction

Depression is a serious mental health disorder that affects mood, thoughts, feelings, and the ability to function in everyday life. Some of its characteristics are prolonged feelings of extreme sadness, guilt and hopelessness, and thoughts of death. Major depression is the leading cause of disability and is the cause of more than two-thirds of suicides each year [1]. Therefore, recognising depression in primary care is a critical public health problem [2]. Effective depression treatment is limited by current assessment methods that rely almost exclusively on patient-reported or clinical judgments of symptom severity [3], risking a range of subjective biases. Affective sensing technology can play a major role in providing an objective assessment.

Recent research into potential bio-markers of central nervous system disorders, e.g. affective disorders, has explored subtle changes in speech characteristics as possible physiologically based indicators for diagnosis and treatment progress [3]. Depression patterns of speech have been analysed for many years, finding differences in the pitch, loudness, speaking rate, and articulation [3, 4, 5, 6]. Therefore, our goal here is to understand the *statistical characteristics* of depressed speech and their effect on *automatic depression detection*, and to reduce the gap between physiological and affective computing studies, which may ultimately lead to an objective affective sensing system that supports clinicians in their diagnosis and monitoring of clinical depression.

Unlike previous approaches, which relied either on machine learning or statistical measurement separately, in this paper, we perform a comparative study of the characteristics of speaker-independent depressed and non-depressed speech by analysing speech features statistically and link them to their effect on automatic recognition results. Although most previous emotion research examined changes on voiced speech only, we investigate voiced, unvoiced and mixed speech signals to identify which is more informative and useful in detecting depression. The remainder of the paper is structured as follows. Section 2 reviews related background literature of depressed speech in both psychology and affective computing. Section 3 describes the methodology, including the dataset, feature extraction and both statistical and classification methods. Section 4 presents the results. The conclusions are presented in Section 5.

2. Background

Psychology research of depressed speech has found several distinguishable prosody features. Formants are a widely used feature in the affect literature [7], being a significantly distinguishable feature for depression [8, 9]. Psychomotor retardation as a symptom of depression can lead to a tightening of the vocal tract, which tends to affect the formant frequencies [10]. Moreover, of the first three formants [8, 9], a noticeable decrease in the second formant frequency was shown for depressives compared to controls [8]. However, since formant features work best when used in speaker-dependent system, they will not be investigated in this work, which focuses on speaker-independent approaches. There is convincing evidence that sadness and depression are associated with a decrease in loudness [11], resulting in lower loudness for depressed people. Since the loudness is intimately related to sound intensity, both features will be investigated.

Jitter and shimmer voice features were analysed, finding higher jitter in depression caused by the irregularity of the vocal fold vibrations [11]. On the other hand, shimmer is lower for depressed subjects [12]. Like the jitter feature, harmonic-to-noise (HNR) values are higher for depressed people, due to the patterns of air flow during speech production differing between depressed and control subjects [13]. Vocal source energy is also a distinguishable feature for depression, resulting in lower energy in the glottal pulses for depressed patients. The excessive tension or lack of coordination in the laryngeal musculature under affect disorders results in an alternation of the glottal flow waveform [14]. Finally, the pitch feature, which has been widely investigated in the literature, shows a lower range of fundamental frequency (F0) in depressed people [3, 4, 5, 6], which increases after treatment [15]. The lower range of F0, indicates a monotone speech [16], and its low variance indicates a lack of normal expression in depressives [9].

Recently, the automatic detection of depression using computer artificial intelligence techniques has been investigated [17, 18, 19, 13, 20]. While psychological investigations are concerned with the overall patterns of speech using statistical functionals of speech features, affective computing classification can be based on frame-by-frame low-level features extracted from speech. The automatic classification from the low-level features results was significant for several features, such as the first 3 formant features gave good classification results in [18], as did energy and loudness [19]. F0 classification results were not as good as expected [18, 19], except in a speaker-dependent context (after treatment) [17]. HNR, jitter and shimmer features gave moderate results in [19], though more investigation is needed. Non-linear features have been investigated recently to detect depression [13, 20]. These features are based on the Teager energy operator (TEO), which measure the number of harmonics produced from the non-linear air flow in the vocal tract. Originally, TEO based features were used to detect stressed speech [21], but they outperform linear features on detecting depression as well [13, 20]. Multi-dimensional speech features were not used in this work, for the purpose of characterising depressed speech and equal comparison.

Not only is there very little work on the automatic detection of depression from speech in the literature, previous researchers applied different methods and features, which make the comparison of results even harder. Most depression studies focus the feature extraction on voiced speech only [18, 13, 20]. Most emotion recognition studies use either low-level or utterance-level features. While low-level features are extracted frame-by-frame at small intervals (typically $10 - 20$ms), the utterance-level features are statistical functionals computed over the entire utterance. For example, [18, 19] used low-level features to recognise depression, while [8, 9] used statistical functional features. Moreover, some studies investigated speaker-dependent depression classification [17], while others used speaker-independent [18, 19, 13]. These diverse methods and features are applicable to speech emotion recognition studies in general [22, 23], which makes it difficult to compare emotional studies. To reduce comparison variability, there is a need for a unified method using the same dataset, classifier and measurements to identify the strongest features and the most suitable speech type (voiced/unvoiced/mixed) for depression detection.

In this work, we perform a comparative study of using linear and non-linear speech features on voiced, unvoiced and mixed speech signals, then compare depression classification results using those features in low-level form with their statistical functionals from spontaneous speech. This work aims to gain a better understanding of depressed speech characteristics, by a unified comparison that will reduce the gap between physiological and affective computing studies.

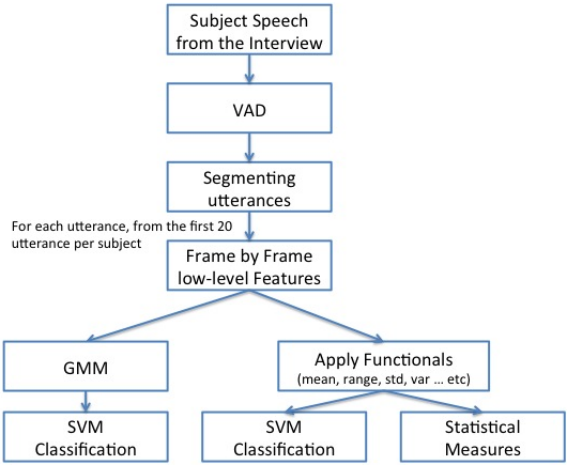

Figure 1: *System Structure and Steps*

3. Method

3.1. Data Collection and Preparation

The Black Dog Institute, which is a clinical research facility in Sydney offering specialist expertise in depression and bipolar disorder, is collecting data from patients who have been diagnosed with depression and healthy controls, who have no history of mental illness. To date, data from over 40 depressed subjects plus over 40 controls (both females and males) have been collected. For the experimental validation, we used a subset of 30 depressed subjects and 30 controls, with equal gender balance. We acknowledge that the amount of data used here is relatively small, but this is a common problem [15, 9]. As the Black Dog Institute continues to collect more data, future studies will be able to report on a larger dataset.

Interviewing the subjects is one part of the experimental paradigm, where specific open questions asking to describe events that had aroused significant emotions. Examples of 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?" The interview was manually labelled to extract pure subject speech, where the total duration was 290min. To automatically find utterances in the extracted subject spontaneous speech, a voice activity detector (VAD) was used. If the voice activity lasted for over 1.5s (as an average for the utterance duration), it was deemed to be an utterance. In our previous study [24], we investigated the effect of analysing shorter speech durations and found that the start of the utterances was more useful for a classification task. Therefore, only the first 20 utterances per subject were selected in this study to ensure a balanced comparison, giving on average 30s per subject.

3.2. Feature Extraction

Voice features can be divided into two categories: acoustic and linguistic. Acoustic features can be categorised into low-level descriptors (LLD) and statistical functionals. Several software tools are available for extracting sound features. In this work,

Table 1: *SVM Classification Results from low-level and functional features for voiced, unvoiced and mixed speech signal*

Speech Feature		Mixed Speech		Voiced Speech		Unvoiced Speech	
		GMM	Functionls	GMM	Functionls	GMM	Functionls
Linear	F0	-	-	63	67	-	-
	Intensity	60	**75**	62	70	65	68
	Loudness	67	68	68	67	68	67
	Voice Probability	75	60	67	60	57	68
	Voice Quality	68	70	**73**	70	70	67
	Jitter	72	62	65	57	68	60
	Shimmer	55	68	60	70	57	-
	HNR	-	63	67	65	-	68
	Log Energy	75	67	65	67	72	**72**
	RMS Energy	70	67	70	72	70	70
TEO Non-Linear	Log Teager Energy	78	70	72	67	**80**	70
	RMS Teager Energy	**80**	**75**	**73**	**73**	70	63
	Teager-Pitch	-	-	62	50	-	-
Average		69	68	67	66	68	67

we used the open-source software "openSMILE" [25] to extract several linear LLD features (see Table 1) and Matlab VoiceBox to extract TEO features [26]. The low-level descriptors were extracted for each utterance with a frame size of 25ms at a shift of 10ms and using a Hamming window. Using the F0 value of each frame, voiced and unvoiced frames were identified and separated. In the analysis, mixed speech used both voiced and unvoiced frames for analysis. Then, we applied several statistical functionals to the LLD for each utterance, including mean, minimum, maximum, range, variance and standard deviation. We chose the most common single-dimensional features in the literature from both the physiological and affective computing fields, for the purpose of statistical characterisation of depressed speech and a fair comparison.

3.3. Classification

In the low-level features, a Gaussian Mixture Model (GMM) with 7 mixture components was created for each utterance. In this context, the GMM serves as dimensionality reduction, as well as a hybrid classification method [27]. The Hidden Markov Model Toolkit (HTK) was used to implement a single-state HMM to train the GMMs. In this work, diagonal covariance matrices were used, and the number of mixtures was fixed to ensure consistency in the comparison. This approach was beneficial to get the same number of values of the extracted features that to be fed to the Support Vector Machine (SVM) regardless of the duration of the subject's utterance. The means, variances and weights of the 7 mixtures of GMM made up the supervector that was fed to the SVM classifier.

To test the effect of the speech characteristics of the voice features on the classification results, an SVM was used on the statistical functional measured on the low-level features.

The subjects were classified in a binary speaker-independent scenario (i.e. depressed/non-depressed) using an SVM, which can be considered as a state-of-the-art classifier for some applications since it provides good generalisation properties [28]. In order to increase the accuracy of the SVM, the cost and gamma parameters need to be optimised. In this paper we used LibSVM [29] to implement the classifier, with a wide range of grid search for the best parameters. To mitigate the effect of the limited amount of data, a leave-5-utterances-per-subject-out cross-validation was used, without any overlap between training and testing data.

The main objective was to correctly classify the subjects as depressed or control. In order to measure the classification results, utterance level classification were calculated first and then subject level classification. That is, if more than 50% of utterances for a given subject were classified as depressed class, the subject is classified as depressed, and vice versa. The performance of a system can be calculated using several statistical methods, such as recall or precision [28]. In this paper, the average recall (AR) was computed. Figure 1 summarises the general structure and the steps of our system.

3.4. Statistical Test

To characterise depressed speech, the extracted statistical functionals from each group (depressed and controls) were compared. A two-way analysis of variance (ANOVA) test was used for this purpose. The ANOVA test studies the effect of one or more qualitative variables (factors) on a quantitative outcome variable. In our case, the tests were two-way ANOVA tests, using the state of depression and subject as factors with each functional, with significance p=0.05. The sample size was 20 utterances for each of the 60 subjects.

4. Results

4.1. Classification Results

Table 1 shows the classification results from linear and TEO speech features comparing low-level features represented by the GMM and their statistical functionals, for voiced, unvoiced and mixed speech. Comparing low-level and the statistical functional features in general, the classification results from low-level were slightly better than those from the statistical functional features, but they were not statistically significant, which indicates a loss of information in the statistical measures. However, using statistical functionals with intensity and shimmer features gives statistically significant results compared with its low-level form. These findings signify the features that benefit from the statistical modeling rather than low-level modeling.

Most emotional speech studies use voiced speech for the analysis. In this work, we investigate whether unvoiced speech holds useful cues in detecting depression, as well as analysing voiced and mixed signal for comparison. As shown in Table 1, the unvoiced speech and mixed signal proved to be as useful as voiced speech, if not better. In general, using mixed signal performed best compared to voiced and unvoiced speech, especially when using low-level features. For example, intensity, log-energy and voicing probability performed best using mixed signal. Using voiced speech, root mean square (RMS) energy and voice quality features performed better than unvoiced and mixed signal. Even though F0 is restricted to voiced speech, its performance was low compared with most other features, which indicates that F0 works better for speaker-dependent comparison. Interestingly, unvoiced speech performs better than voiced speech using log-energy and log-Teager-Energy in both low-

Table 2: *Statistical Significant Results from functional features for voiced, unvoiced and mixed speech signal*

Speech Feature		Statistical Feature	Mixed Speech		Voiced Speech		Unvoiced Speech	
			DIR.	P val	DIR.	P val	DIR.	P val
Linear Features	F0	Mean	-	-	D > C	0.01	-	-
		Variance	-	-	D < C	0.04	-	-
		Std	-	-	D < C	0.04	-	-
	Intensity	Mean	D < C	0.01	D < C	0.01	D < C	0.01
		Range	D < C	0.01	D < C	0.01	D < C	0.01
		Variance	D < C	0.04	NS	NS	D < C	0.01
		Std	D < C	0.01	D < C	0.01	D < C	0.004
	Loudness	Mean	D < C	0.01	D < C	0.01	D < C	0.02
		Range	D < C	0.01	D < C	0.01	D < C	0.03
		Variance	D < C	0.01	D < C	0.004	D < C	0.01
		Std	D < C	0.01	D < C	0.005	D < C	0.03
	Voice Quality	Mean	D < C	0.02	D < C	0.02	D < C	0.04
	Log Energy	Mean	D < C	0.02	D < C	0.01	D < C	0.03
	RMS Energy	Mean	D < C	0.01	D < C	0.01	D < C	0.01
		Range	D < C	0.01	D < C	0.01	D < C	0.01
		Variance	D < C	0.01	D < C	0.01	D < C	0.01
		Std	D < C	0.01	D < C	0.01	D < C	0.01
Non-Linear Features	Log Teager Energy	Mean	D < C	0.01	D < C	0.003	D < C	0.01
	RMS Teager Energy	Mean	D < C	0.02	D < C	0.004	D < C	0.04
		Range	D < C	0.03	D < C	0.01	NS	NS
		Std	D < C	0.03	D < C	0.01	NS	NS

NS: Not Significant, Std: Standard Deviation, DIR.: Direction of Effect, D: Depressed group, C: Control group

level and functional features, which implies that some features from unvoiced speech hold important information that should be considered while analysing depression.

Although most TEO based features used in emotional investigation used voiced speech only [13, 20, 21], we investigate unvoiced and mixed speech as well. The best depression recognition results in this study were achieved by RMS-Teager-Energy giving 80% accuracy using mixed speech and Log-Teager-Energy, which was also 80% accurate using unvoiced speech. This finding points to the suitability of TEO based features in detecting depression from voiced and unvoiced speech, which requires further study. In contrast, the best linear feature recognition rates were provided by mixed signal of both intensity, which gave 75% recognition rate using the functional, and voicing probability and log-energy which gave 75% recognition rate using the low-level features, which is in line with our previous finding using longer speech duration [19, 27].

4.2. Statistical Results

Table 2 shows only the statistically significant results from the ANOVA tests for the functional features, for each of voiced, unvoiced and mixed speech. Among the statistically significance features, only the F0 mean was higher in depressed than controls. Even though the F0 range is supposed to be lower in depressed patients as mentioned in the literature [3, 4, 5, 6], it was not significant in our comparison, which indicated that F0 range might be significant once compared using the same speaker [17]. As can be seen in Table 2, unvoiced speech gives the same separation as voiced and mixed speech, which indicates that it holds important information that should be considered while analysing speech for depression in particular and possibly emotions in general. In line with the literature, loudness [11], intensity, linear energy features (log and RMS) [14], and non-linear TEO energy (log and RMS) have lower mean in depressed patients. Those statistical significance tests explain the high recognition rates using the SVM classifier provided in Table 1. Higher jitter [11], lower shimmer [12], and higher harmonic-to-noise (HNR) [13] in depressives were expected; in our study, their changes were not statistically significant. Their performance were lower using SVM compared with other features as shown in Table 1.

5. Conclusions

To gain a better understanding of depressed speech, we investigated depressed speech characteristics and their utility in the classification of depression. This understanding may ultimately lead to an objective affective sensing system that supports clinicians in their diagnosis and monitoring of clinical depression. Several low-level features and statistical functionals from linear and TEO non-linear speech features were extracted from depressed and non-depressed speech utterances. In general, the classification results from low-level features were slightly better than the statistical functionals but their difference was not statistically significant, which indicates the loss of information in the latter. However, intensity and shimmer features benefitted more from their statistical functionals, which signals that they benefit from statistical modeling rather than low-level modeling. Although most emotional speech studies use voiced speech for the analysis, we found that unvoiced speech holds useful cues in detecting depression as well, which implies that unvoiced speech should be considered while analysing depression in particular and possibly emotions in general. Generally, using mixed signal performed best compared to voiced and unvoiced speech, indicating that mixed speech is rich in emotional cues and useful for detecting depression. Interestingly, unvoiced speech performs better than voiced speech using log-Energy and log-Teager-Energy in both low-level and functional features. In general, TEO non-linear energy speech features outperformed (log 78% and RMS 80%) linear speech features, which points to the suitability of TEO based features in detecting depression, where more investigation is required. On the other hand, best linear feature recognition rates were achieved for mixed speech with intensity, voicing probability and log-energy features (75%). Finally, most speech features classification results are consistent with their statistical characteristics, which conforms the results of both physiological and affective computing studies. Future work will investigate multi-dimensional speech features, including TEO based features.

6. Acknowledgement

This research was funded in part by the ARC Discovery Project grant DP130101094.

7. References

[1] U. D. of Health and H. Services, "Healthy people 2010: Understanding and improving health," 2000.

[2] S. Baik, B. J. Bowers, L. D. Oakley, and J. L. Susman, "The recognition of depression: The primary care clinicians perspective," *Annals Of Family Medicine*, vol. 3, no. 1, pp. 31–37, 2005.

[3] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, and D. 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. Nilsonne, "Speech characteristics as indicators of depressive illness," *Acta Psychiatrica Scandinavica*, vol. 77, no. 3, pp. 253–263, 1988.

[5] 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.

[6] 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.

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

[8] A. J. Flint, S. E. Black, I. Campbell-Taylor, G. F. Gailey, and C. Levinton, "Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression," *Journal of Psychiatric Research*, vol. 27, no. 3, pp. 309–319, Jul. 1993.

[9] E. Moore, M. Clements, J. Peifer, and L. Weisser, "Critical analysis of the impact of glottal features in the classification of clinical depression in speech." *IEEE Trans. on Bio-medical Eng.*, vol. 55, no. 1, pp. 96–107, Jan. 2008.

[10] 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 Trans. on bio-medical Eng.*, vol. 47, no. 7, pp. 829–37, Jul. 2000.

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

[12] A. Nunes, L. Coimbra, and A. 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.

[13] L. A. Low, N. C. Maddage, M. Lech, L. B. Sheeber, and N. B. Allen, "Detection of clinical depression in adolescents speech during family interactions." *IEEE Trans. on Biomedical Eng.*, vol. 58, no. 3, pp. 574–586, 2011.

[14] A. Ozdas, R. G. Shiavi, S. E. Silverman, M. K. Silverman, and D. M. Wilkes, "Investigation of vocal jitter and glottal flow spectrum as possible cues for depression and near-term suicidal risk." *IEEE Trans. on Biomedical Eng.*, vol. 51, no. 9, pp. 1530–1540, 2004.

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

[16] E. Moore, M. Clements, J. Peifer, and L. 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.

[17] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. 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, Sep. 2009.

[18] N. Cummins, J. Epps, M. Breakspear, and R. Goecke, "An Investigation of Depressed Speech Detection: Features and Normalization," in *Proc. Interspeech*, 2011, pp. 2997–3000.

[19] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "From Joyous to Clinically Depressed: Mood Detection Using Spontaneous Speech," in *Proc. FLAIRS-25*, 2012, accepted.

[20] K. E. B. Ooi, L.-S. A. Low, M. Lech, and N. Allen, "Early prediction of major depression in adolescents using glottal wave characteristics and teager energy parameters," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4613–4616.

[21] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Classification of speech under stress based on features derived from the nonlinear teager energy operator," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 549–552.

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

[23] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual, and spontaneous expressions." *IEEE Trans. on PAMI*, vol. 31, no. 1, pp. 39–58, 2007.

[24] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting Depression: A Comparison between Spontaneous and Read Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, May 2013.

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

[26] M. Brookes *et al.*, "Voicebox: Speech processing toolbox for matlab," *Software, available [Mar. 2011] from www. ee. ic. ac. uk/hp/staff/dmb/voicebox/voicebox. html*, 1997.

[27] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, G. T, M. Breakspear, and G. Parker, "A Comparative Study of Different Classifiers for Detecting Depression from Spontaneous Speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, May 2013.

[28] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. Feb, pp. 1062–1087, 2011.

[29] C. C. Chang and C. J. Lin, "Libsvm: a library for svm," *2006-03-04]. http: www. csic. ntu. edu. tw/rcjlin/papers/lib. svm*, pp. 1–30, 2001.