Interspeech 2018
2-6 September 2018, Hyderabad

Depression Detection from Short Utterances via Diverse Smartphones in Natural Environmental Conditions

Zhaocheng Huang[1], *Julien Epps*[1], *Dale Joachim*[2], *Michael Chen*[2]

[1]School of Electrical Engineering and Telecommunications, UNSW Sydney, Australia
[2]Sonde Health, Boston MA, USA

zhaocheng.huang@unsw.edu.au, j.epps@unsw.edu.au

Abstract

Depression is a leading cause of disease burden worldwide, however there is an unmet need for screening and diagnostic measures that can be widely deployed in real-world environments. Voice-based diagnostic methods are convenient, non-invasive to elicit, and can be collected and processed in near real-time using modern smartphones, smart speakers, and other devices. Studies in voice-based depression detection to date have primarily focused on laboratory-collected voice samples, which are not representative of typical user environments or devices. This paper conducts the first investigation of voice-based depression assessment techniques on real-world data from 887 speakers, recorded using a variety of different smartphones. Evaluations on 16 hours of speech show that conservative segment selection strategies using highly thresholded voice activity detection, coupled with tailored normalization approaches are effective for mitigating smartphone channel variability and background environmental noise. Together, these strategies can achieve F1 scores comparable with or better than those from a combination of clean recordings, a single recording environment and long utterances. The scalability of speech elicitation via smartphone allows detailed models dependent on gender, smartphone manufacturer and/or elicitation task. Interestingly, results herein suggest that normalization based on these criteria may be more effective than tailored models for detecting depressed speech.

Index Terms: Depression detection, mobile devices, environmental noise, elicitation tasks, normalization.

1. Introduction

Depression is a common and costly condition, affecting 10-15% of the population [1]. Despite the difficulty of objectively measuring depression severity, there remains an unmet need to detect depression in a range of settings [2]. An objective, passive, ubiquitous device for capturing behavioral and cognitive information conveniently and continuously would be a compelling tool for research and clinical practice [3, 4]. Over 50% of US adults own a smartphone, and many suffer from significant mental health conditions [5]. One highly promising method is to elicit speech [6] and automatically screen depression via smartphone [7], and this approach may be most effective if it is applicable to short utterances that can be easily and regularly elicited.

Despite the potential, research into smartphone-based analysis of emotion and mental state is still at an early stage. Variations between smartphone microphones, audio acquisition, and pre-processing represent a challenge to unwanted variability, both in general for feature extraction [8] and specifically for detection of depression [7]. Further, there are challenges of environmental noise likely to be encountered in the context of everyday use of smartphone-based speech acquisition. Noisy assessment of speech has been considered in emotion recognition [9, 10] and depression prediction [11, 12] contexts, but further work is needed to understand how best to select reliable data for analysis.

In this paper, we investigate the effect of speech segment selection and smartphone variability mitigation on depression classification from speech collected under realistic conditions, with a particular focus on short utterances.

2. Related Work

In their overview of robust *Automatic Speech Recognition* (ASR), Li *et al.* [13] suggest a series of different categories of approaches, including robust features, feature compensation, feature normalization and model-space methods. Feature compensation approaches such as spectral subtraction or Wiener filtering have shown some success in robust ASR [14], however they depend on how effectively the noise statistics can be estimated for each utterance. To date normalization may be one of the most successful of the available approaches for paralinguistic applications, in part due to the lack of data available for the other approaches from different typical operating conditions, and in part due to the different sources of variability that can be mitigated this way (e.g. speaker, phonetic, etc). Common normalization approaches include mean and mean-variance normalization and histogram equalization.

In paralinguistic applications, noise has proven a difficult challenge. The application of non-negative matrix factorization to emotion recognition gave a small improvement [9] and robust Damped Oscillator Cepstral Coefficient features performed well compared with *Mel-Frequency Cepstral Coefficients* (MFCCs) for depression detection tasks [11], yet in all cases, performance was significantly affected by noise. However it is not necessary to analyze the entire speech recording, motivating the investigation of conservative *Voice Activity Detection* (VAD) methods. For example, in speaker recognition, the choice of voice activity detector has been shown to provide significant improvements [15].

Despite the promise of mobile devices for emotion [16] and mental state assessment, to date relatively few studies have investigated the challenges related to environment and means of speech collection. Gideon *et al.* [17] found that Sadjadi and Hansen's VAD algorithm [18] provided some noise robustness when selecting speech collected from three different Samsung smartphones for analysis of bipolar disorder. In previous work by our group, different smartphones were found to have significantly different hardware and software characteristics

10.21437/Interspeech.2018-1743

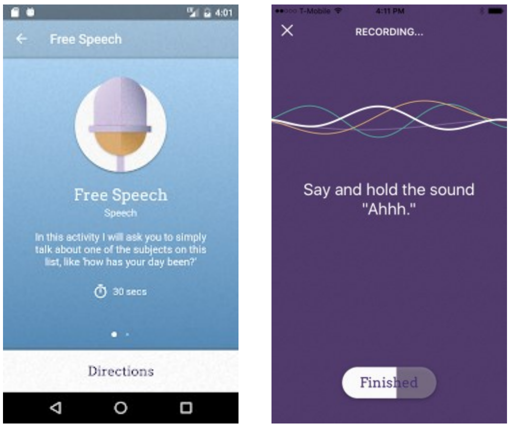
favoring manufacturer-specific mitigation methods, particularly for spectral features [7], while features such as F0 were more invariant across device types [8]. However some previous studies employing smartphones have considered only a single manufacturer or device model [19]. Further, although the limitations of small datasets for automatic depression assessment are acknowledged by many (e.g. [20]), with reasonable quantities of smartphone data, it is possible to investigate approaches that are specific to speech elicitation (task), smartphone manufacturer and/or gender.

When assessing mental state in everyday contexts, it is impractical to require speakers to complete a long protocol, and detection performance as a function of the amount of speech data available becomes a key consideration. Short speech segments can yield depression detection accuracies better than those of much longer utterances [21, 22]. Further, there is interest in designing protocols for eliciting speech that is most discriminative of depression (e.g. [6, 23]), and these aspects are also investigated herein.

3. Methods

3.1. Multi-platform Smartphone App-based Data Collection

Data were collected from an interactive app running on Android™ smartphones. The app (Figure 1), created by Sonde Health, was designed to elicit and collect speech samples along with questionnaires, including the *Patient Health Questionnaire* (PHQ-9), a validated instrument that measures risk for depression.



In this activity I will ask you to simply talk about one of the subjects on this list, like 'how has your day been?'

Say and hold the sound "Ahhh."

Figure 1: *Partial screenshots of the Sonde Health smartphone speech elicitation app.*

Speech (sampled at 16kHz), alongside device metadata and questionnaire data were collected from a general population sample in the United States under a human subject protocol reviewed and approved by an Institutional Review Board. All data were encrypted on-device and transmitted to a secure cloud storage. Participants completed several voice tasks on their personal smartphones in uncontrolled natural environments, including free speech, read speech, and elicited tasks (e.g. the sustained vowel "ahh"; diadochokinetic repetition). For example, participants were instructed to repeat a sentence from the Harvard Sentence database on the screen, or to freely respond for up to 30 seconds on a generic topic such as "What is the weather like outside?".

A subset of the collected data, which we refer to as the SH2 dataset, was used for experiments in this study. SH2 contains around 16 hours of speech for 887 participants (436 female and 450 male). The 5937 total audio files comprise six tasks (i.e. sustained vowel, diadochokinetic, free speech, rainbow passage, cognitive load and sentence), completed by 498 to 810 participants. The SH2 dataset also includes a wide variety of mobile device and smartphone manufacturers (28 in total), as shown in Figure 2.

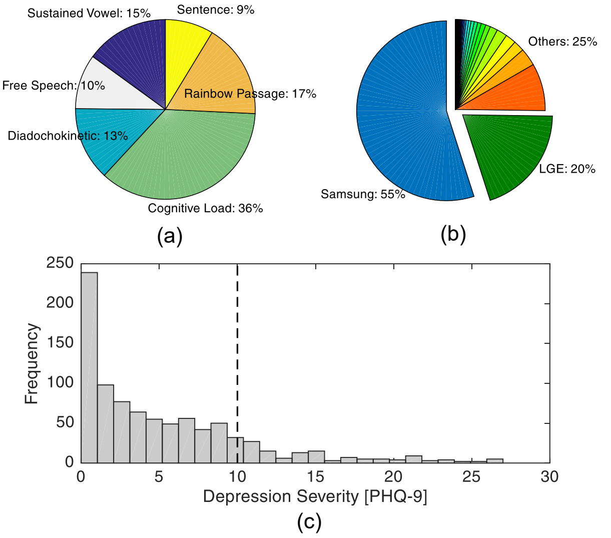


Figure 2: *Statistics of (a) task and (b) manufacturer proportions for all files, and (c) the histogram of PHQ-9 scores for the 887 speakers.*

Furthermore, an advantage of this corpus is that, unlike some synthetic datasets where noise was artificially added (e.g. [11]), all recordings in this corpus contained at least some background noise from real-world environments. Although it is not straightforward to estimate the recording SNR, subjective estimates suggest that most recordings fall within about 5 to 20 dB SNR. Typical noise types include office, babble (e.g. conversation or background TV), restaurant noise, etc.

3.2. Proposed System

3.2.1. *Adopted Voice Activity Detection Approaches*

While there have been extensive investigations into VAD within the speech communication community [24, 25], it remains difficult to designate a single ideal VAD for all conditions. We compared 7 VAD approaches, spanning a range of different criteria: VOICEBOX [24] is based on a statistical likelihood ratio test; KARMA [26], which has been effectively used in state-of-the-art depression prediction systems [27], is based on the smoothness of formant tracks; openSMILE [28] produces a voicing probability associated with subharmonic summation; MFCC-based VAD offers filter-based decisions; Sadjadi's VAD [18] linearly combines four voicing measures in both time domain (i.e. harmonicity, clarity, and prediction gain) and frequency domain (i.e. periodicity and perceptual spectral flux); the summation of the *Residual Harmonics* (SRH)-based VAD [29] employs the harmonic information from residual signals, which was found to work well in noisy conditions; and COVAREP [25] further combines the probabilistic outputs from the MFCC-based, Sadjadi and SRH-based VADs. VOICEBOX and KARMA only offer binary decisions, while the others produce frame-level probabilistic decisions, which allow different thresholds to be applied.

3.2.2. *Proposed Task/Manufacturer/Gender Normalization*

Similarly to the important role of VAD in handling noise, normalization is among the first choices to deal with variability of different kinds. For instance, in speaker verification, normalization is helpful for mitigating mismatch between different handsets and training-testing scores [30]. Given that the SH2 dataset consists of different tasks, manufacturers, and genders, it is reasonable to guess that these introduce unwanted characteristics that undermine depression detection, e.g. gender differences [31]. Among the normalization methods, some widely used candidates are mean norm, standard deviation norm, mean-variance norm, and histogram equalization. Based upon these choices, we investigated normalization specific to task, manufacturer and gender. The aforementioned four normalization methods (i.e. mean norm, etc.) were implemented, and the best performing normalization method was selected for task, manufacturer and gender respectively.

Given a set of training data $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n, \dots, \boldsymbol{x}_N\}$, where N is the total number of files, we defined data subsets specific to a particular task ($1 \le t \le T$), gender ($1 \le g \le G$), or manufacturer ($1 \le m \le M$), as $\boldsymbol{X}_t = \{\boldsymbol{x}_1^t, \dots, \boldsymbol{x}_n^t, \dots, \boldsymbol{x}_{N_t}^t\}$, $\boldsymbol{X}_m = \{\boldsymbol{x}_1^m, \dots, \boldsymbol{x}_n^m, \dots, \boldsymbol{x}_{N_m}^m\}$, and $\boldsymbol{X}_g = \{\boldsymbol{x}_1^g, \dots, \boldsymbol{x}_n^g, \dots, \boldsymbol{x}_{N_g}^g\}$, where N_t, N_m, and N_g are the total number of $g/m/t$-specific files. T=6, M=3 and G=2 are the numbers of tasks, manufacturer groups, and gender. The normalized features are then:

$$\tilde{\boldsymbol{x}}_n^t = \frac{\boldsymbol{x}_n^t - \boldsymbol{\mu}^t}{\boldsymbol{\sigma}^t}, \forall \boldsymbol{x}_n^t \in \boldsymbol{X}_t, 1 \le t \le T \qquad (1)$$

$$\tilde{\boldsymbol{x}}_n^m = \boldsymbol{x}_n^m - \boldsymbol{\mu}^m, \forall \boldsymbol{x}_n^m \in \boldsymbol{X}_g, 1 \le m \le M \qquad (2)$$

$$\tilde{\boldsymbol{x}}_n^g = \frac{\boldsymbol{x}_n^g - \boldsymbol{\mu}^g}{\boldsymbol{\sigma}^g}, \forall \boldsymbol{x}_n^g \in \boldsymbol{X}_g, 1 \le g \le G \qquad (3)$$

where $\boldsymbol{\mu}^t$, $\boldsymbol{\mu}^m$, $\boldsymbol{\mu}^g$, $\boldsymbol{\sigma}^t$, $\boldsymbol{\sigma}^m$, $\boldsymbol{\sigma}^g$ are parameters learnt from the training data, and applied to normalize the test data.

3.3. Experimental Settings

The SH2 Corpus was divided into training (4641 files for 695 speakers) and test partitions (1296 files for 192 speakers). A threshold of 10 was used to separate healthy (PHQ-9<10) and depressed (PHQ-9≥10) speakers, which is recommended in [32]. As a result, there are 122 and 35 depressed speakers in the training and test data respectively. It is worth noting that most studies leave a gap in the PHQ-9 scores when defining healthy and depressed speakers, e.g. [7].

The 38-dimensional IS2010 *Low-Level Descriptors* (LLDs) [33], which primarily consist of spectral features that are found to be informative for depression classification [34] were extracted from speech, e.g. MFCCs, pitch, loudness, jitter and shimmer. Frame-level VAD decisions were then used to select only voice-active frames from the extracted LLDs before calculating file-level functionals (i.e. global statistics). For functionals, the mean, standard deviation, 20%, 50%, 80% percentiles, range of 20-80% percentiles, skewness and kurtosis were chosen, due to being widely used and robust [35].

Linear *Support Vector Machine*(SVM) [36] with parameter sweeps of C values from 10^{-5} to 10 in a log space was trained in a 3-fold cross validation scheme within the training data and the best parameter was adopted for testing on the test data. During training, C was weighted inversely proportional to class frequencies to handle imbalanced training data for the healthy and depressed classes.

The primary evaluation measure adopted was F1 score (for *depressed* speakers), which combines precision and recall of depression detection, satisfying the needs of a real screening method, recognizing that depression is the focus. Conventional measures such as accuracies often do not reflect the real ability of a system to recognize depressed speech, since healthy speakers are a much larger proportion of the dataset. Note that C was optimized for F1 on the training data. The F1 scores and accuracies were calculated per-speaker, and the fusion of task-specific decisions was also investigated. 3-fold cross validation was conducted to find the best fusion model within the training data, and the best model (which was re-trained on the whole training data using the parameter C that performed the best across 3 folds) was used to generate the final decision on the test data.

4. Results

4.1. Effect of Segment Selection

To examine the effect of segment selection, different threshold settings were applied on VAD decisions, as shown in Figure 3. For openSMILE, the thresholds of {0.5, 0.6, 0.7, 0.75, 0.8} were used (beyond 0.8 leads to a very large % dropped). Thresholds ranging from 0.5 to 0.99 were trialed for the MFCC-based, Sadjadi, SRH and COVAREP VADs. In this series of experiments, all training features were centered and scaled to unit variance, and the normalization coefficients were used to normalize the test data.

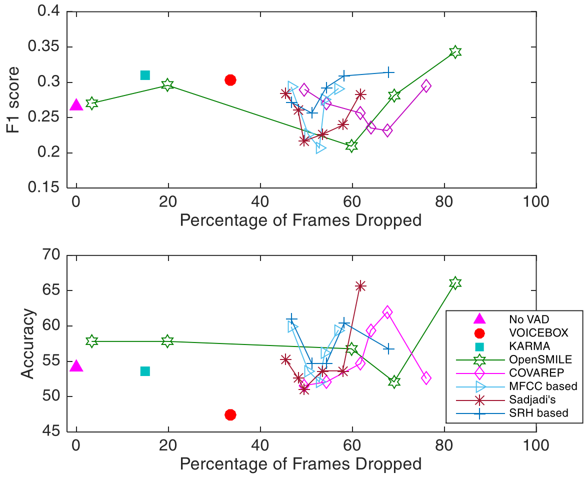


Figure 3: *F1 (depressed) scores and accuracies using various VAD approaches. The VADs result in 3% to 82% of all frames being dropped.*

In general, VAD-based systems can outperform the same system without VAD in terms of F1 score. Interestingly, as the threshold increases, all systems have different extents of drops in F1 and accuracy before achieving the best performance.

The best F1 score (0.343) and accuracy (66.1%) were obtained for the most conservative approach (openSMILE voiceProb>0.8), which interestingly retained only 18% of all recorded frames. This is perhaps not surprising given the existence of background noise in a large number of recordings. The openSMILE voiceProb>0.8 VAD setting was used in all subsequent experiments.

4.2. Mitigation of Smartphone Task, Manufacturer and Gender Variability

This experiment aims to examine whether modelling or normalization that is specific to certain tasks, smartphone manufacturer groups, and genders (referred to as *X*-specific normalization/modelling) can help reduce potential variability in data collected in realistic environments.

For *X*-specific normalization (Section 3.2.1), *X*-specific subsets of the training and test data were normalized, based on normalization coefficients learnt from *X*-specific subsets of the training data. Afterwards, *X*-specific SVM models were trained. Further, using the same model, predictions were generated from the subset of the test data that are specific to each task, manufacturer, gender, and their F1 scores were calculated. For *X*-specific modelling, differing from Section 4.1 where a single global model was trained, we trained one model on each subset during training (e.g. male). Accordingly, there were six models for tasks, three models for manufacturers, and two models for genders. These predictions were concatenated to calculate F1 scores for task, manufacturer, and gender.

Table 1: *F1 (depressed) scores for X-specific modelling and normalization. The baseline F1 score for no normalization was 0.290.*

		Model		**Normalization**	
Gender	Male	.322	.321	.396	.462
	Female		.323		.327
Manufacturer	Samsung	.349	.317	.360	.394
	LGE		.250		.133
	Others		.421		.400
Task	Sentence	.260	.304	.333	.412
	CL		.184		.237
	Free Speech		.196		.333
	Vowel		.267		.283
	Passage		.321		.321
	Diadochokinetic		.267		.262

Comparisons between *X*-specific normalization and modelling can be seen in Table 1. In general, *X*-specific modelling was outperformed by normalization, which further improves F1 scores to 0.360 for manufacturer and 0.396 for gender. However, it is worth noting that less training data were used in *X*-specific modelling, whilst the proposed normalization can mitigate variability well while maintaining the whole data for training. The configuration of gender-specific normalization was used in the following experiments.

4.3. Classification of Short Utterances

Apart from noise and variability, there remains a question as to what duration of speech recording is required for efficient depression detection in practice, as a guide to elicitation design. Accordingly, based on the pre-trained SVM model with gender-specific normalization in Table 1, we calculated F1 scores and accuracies on subsets of test data specific to each task, as shown in Figure 4.

These results suggest that longer speech recordings do not necessarily yield better performance, e.g. "diadochokinetic" vs "free speech". One possibility is that longer utterances have more non-speech and hence more background noise. Moreover, it was found that "diadochokinetic" and "sentence" are efficient tasks for eliciting healthy and depressed speech that are systematically distinguishable.

An important note is all the above individual results were outperformed by fusion of the individual task results, suggesting the benefits of fusing across multiple tasks. Majority voting yielded F1 0.396 and accuracy 69.8%, while score fusion using logistic regression yielded the best result of F1 0.422 (depressed), F1 0.823 (healthy) and accuracy 72.9%.

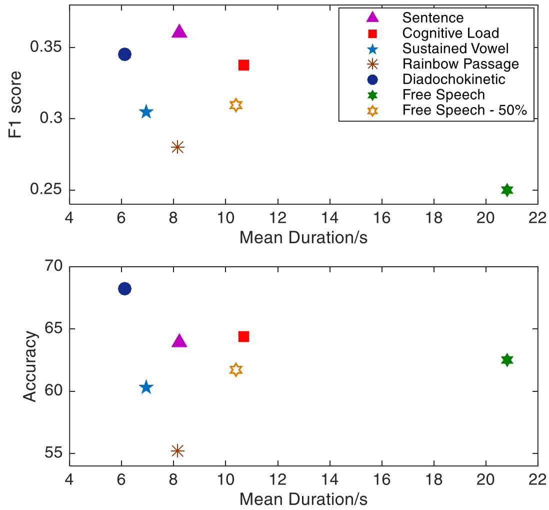


Figure 4: *F1 (depressed) scores and accuracies for the various types of elicited speech in SH2 dataset, using gender-specific normalization system. "Free Speech – 50%" refers to selection of the middle 50% of data from the "Free Speech" task before VAD.*

5. Conclusions

This research has investigated three essential practical aspects for the deployment of speech-based automated depression detection systems in natural environments: noise, smartphone/mobile device variability, and short utterances. To accomplish this, a wide spectrum of VAD approaches were tested, and normalization specific to tasks, smartphone manufacturers and genders was proposed. The combination of highly thresholded VAD, tailored normalization and fusion of task-specific scores yielded the highest performance (F1 0.422 (depressed), F1 0.823 (healthy) and accuracy 72.9%).

These results are comparable with or better than those of the AVEC 2016 audio baselines on the DAIC dataset [37], which has clean recordings, a single recording environment and long utterances. Apart from representing the first study of this kind for automatic depressed speech detection, two significant details should be noted: (i) the results were tested on a large number of speakers compared with previous studies; and (ii) the majority of previous studies include a gap between the PHQ scores when constructing the 'depressed' and 'non-depressed' classes, whilst no gap was used herein. Taken together, this paper gives an important perspective on depression detection under realistic conditions. For future work, more robust feature sets (e.g. vocal tract coordination features [27]) and machine learning models (e.g. Gaussian staircase regression [27] and neural network based methods), may create higher benchmarks.

6. Acknowledgements

This work was supported by Australian Research Council Linkage Project LP160101360. Julien Epps is also partly supported by Data61, CSIRO, Australia.

7. References

[1] Walker, J., K. Burke, M. Wanat, R. Fisher, J. Fielding, A. Mulick, S. Puntis, J. Sharpe, M. Degli Esposti, and E. Harriss, "The Prevalence of Depression in General Hospital Inpatients: A Systematic Review and Meta-Analysis of Interview Based Studies," *Psychological Medicine*, 2018.

[2] Cohn, J. F., N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal Assessment of Depression from Behavioral Signals," in *Handbook of Multi-Modal Multi-Sensor Interfaces*, D. Oviatt, S., Schuller, B., Cohen, P., and Sonntag, Ed. Morgan and Claypool, 2017, pp. 113–155.

[3] Insel, T. R., "Digital phenotyping: Technology for a new science of behavior," *JAMA - Journal of the American Medical Association*, vol. 318, no. 13, pp. 1215–1216, 2017.

[4] Cummins, N., S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, Jul. 2015.

[5] Ben-Zeev, D., E. A. Scherer, R. Wang, H. Xie, Andrew, and T. Campbell, "Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health," *Psychiatric Rehabilitation Journal*, vol. 38, no. 3, pp. 218–226, 2015.

[6] Stasak, B., J. Epps, and R. Goecke, "Elicitation design for acoustic depression classification: An investigation of articulation effort, linguistic complexity, and word affect," in *INTERSPEECH*, 2017, pp. 834–838.

[7] Stasak, B. and J. Epps, "Differential performance of automatic speech-based depression classification across smartphones," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2017, pp. 171–175.

[8] Grillo, E. U., J. N. Brosious, L. Staci, and S. Anand, "Influence of smartphones and software on acoustic voice measures," *International Journal of Telerehabilitation*, vol. 8, no. 2, pp. 9–14, 2016.

[9] Weninger, F., B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *Eurasip Journal on Advances in Signal Processing*, vol. 2011, 2011.

[10] Schuller, B., D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," *Construction*, pp. 276–289, 2006.

[11] Mitra, V., A. Tsiartas, and E. Shriberg, "Noise and reverberation effects on depression detection from speech," in *IEEE ICASSP*, 2016, pp. 5795–5799.

[12] Low, L. A., N. C. Maddage, M. Lech, L. B. Sheeber, and N. 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.

[13] Li, J., L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 4. pp. 745–777, 2014.

[14] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms," 2002.

[15] Mak, M. W. and H. B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," *Computer Speech and Language*, vol. 28, no. 1, pp. 295–313, 2014.

[16] Marchi, E., F. Eyben, G. Hagerer, and B. Schuller, "Real-time tracking of speakers' emotions, states, and traits on mobile platforms," in *INTERSPEECH*, 2016, pp. 1182–1183.

[17] Gideon, J., E. M. Provost, and M. Mclnnis, "Mood State Prediction from Speech of Varying Acoustic Quality for Individuals with Bipolar Disorder," *Pediatr Neurol*, vol. 52, no. 6, pp. 566–584, 2016.

[18] Sadjadi, S. O. and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[19] Vásquez-Correa, J. C., T. Arias-Vergara, J. R. Orozco-Arroyave, J. F. Vargas-Bonilla, J. D. Arias-Londoño, and E. Nöth, "Automatic detection of Parkinson's disease from continuous speech recorded in non-controlled noise conditions," in *INTERSPEECH*, 2015, pp. 105–109.

[20] Or, F., J. Torous, and J.-P. Onnela, "High potential but limited evidence: Using voice data from smartphones to monitor and diagnose mood disorders.," *Psychiatric Rehabilitation Journal*, vol. 40, no. 3, pp. 320–324, 2017.

[21] Stasak, B., J. Epps, and N. Cummins, "Depression Prediction Via Acoustic Analysis of Formulaic Word Fillers," in *Australasian International Conference on Speech Science and Technology*, 2016, pp. 277–280.

[22] Alghowinem, S., R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," in *ICASSP*, 2013, pp. 7547–7551.

[23] Mundt, J. C., 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 (IVR) technology," *Journal of Neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[24] Sohn, J., "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[25] Drugman, T., Y. Stylianou, Y. Kida, *et al.*, "Voice Activity Detection: Merging Source and Filter-based Information," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 252–256, 2016.

[26] Mehta, D. D., D. Rudoy, and P. J. Wolfe, "Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking," *The Journal of the Acoustical Society of America*, vol. 135, no. 5, pp. 3128–3128, 2012.

[27] Williamson, J. R., T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal biomarkers of depression based on motor incoordination," in *Proceedings of the 4th ACM International Workshop on AVEC, ACM MM*, 2013, pp. 41–47.

[28] Eyben, F., F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 835–838.

[29] Drugman, T. and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *INTERSPEECH*, 2011, pp. 1973–1976.

[30] Bimbot, F., J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-chagnolleau, S. Meignier, T. Merlin, J. Ortega-garc, D. Petrovska-delacr, and D. A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP Journal on Applied Signal Processing*, pp. 430–451, 2004.

[31] Hönig, F., A. Batliner, E. Nöth, and S. Schnieder, "Automatic modelling of depressed speech: relevant features and relevance of gender.," in *INTERSPEECH*, 2014, pp. 1248–1252.

[32] Kroenke, K., R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[33] Schuller, B., S. Steidl, A. Batliner, F. Burkhardt, *et al.*, "The INTERSPEECH 2010 Paralinguistic Challenge," in *INTERSPEECH*, 2010, pp. 2794–2797.

[34] Cummins, N., J. Epps, V. Sethu, and J. Krajewski, "Weighted pairwise Gaussian likelihood regression for depression score prediction," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4779–4783.

[35] Eyben, F., K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[36] Fan, R.-E., K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A Library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[37] Valstar, M., J. Gratch, F. Ringeval, M. T. Torres, S. Scherer, and R. Cowie, "AVEC 2016 – Depression , Mood , and Emotion Recognition Workshop and Challenge," in *Proceedings of the 6th International Workshop on AVEC, ACM MM*, 2016, pp. 3–10.