INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Glottal Source Features for Automatic Speech-based Depression Assessment

Olympia Simantiraki[1], *Paulos Charonyktakis*[2], *Anastasia Pampouchidou*[3], *Manolis Tsiknakis*[4], *Martin Cooke*[1]

[1]Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain
[2]Gnosis Data Analysis PC, Heraklion, Greece
[3]Le2i Laboratory, University of Burgundy, Le Creusot, France
[4]Technological Educational Institute of Greece and FORTH, Heraklion, Greece

olympia.simantiraki@ehu.eus, m.cooke@ikerbasque.org

Abstract

Depression is one of the most prominent mental disorders, with an increasing rate that makes it the fourth cause of disability worldwide. The field of automated depression assessment has emerged to aid clinicians in the form of a decision support system. Such a system could assist as a pre-screening tool, or even for monitoring high risk populations. Related work most commonly involves multimodal approaches, typically combining audio and visual signals to identify depression presence and/or severity. The current study explores categorical assessment of depression using audio features alone. Specifically, since depression-related vocal characteristics impact the glottal source signal, we examine Phase Distortion Deviation which has previously been applied to the recognition of voice qualities such as hoarseness, breathiness and creakiness, some of which are thought to be features of depressed speech. The proposed method uses as features DCT-coefficients of the Phase Distortion Deviation for each frequency band. An automated machine learning tool, Just Add Data, is used to classify speech samples. The method is evaluated on a benchmark dataset (AVEC2014), in two conditions: read-speech and spontaneous-speech. Our findings indicate that Phase Distortion Deviation is a promising audio-only feature for automated detection and assessment of depressed speech.

Index Terms: glottal source, Phase Distortion Deviation, binary classification, machine learning

1. Introduction

Depression is a mood disorder with a great societal cost [1], predicted to become the second most significant cause of disability worldwide by 2020 [2]. Early detection of the disorder is paramount. In a clinical setting, it is important to quantify depression in order to be able to track the progress of a depressed individual in a non-intrusive manner [3]. Speech-based assessment is one promising approach for the non-intrusive detection of depression [4]. The current paper describes an approach to automatic detection of depressed speech.

Depressed individuals are characterized by psychomotor retardation, manifested as a slowing of coordination, speech, and impaired articulation, leading to changes in the glottal source, vocal tract and prosodic features. Perceived depressed voice characteristics include monotony, hoarseness, breathiness and reduced speech rate. Here, we focus on the speech characteristics related to the glottal source signal for discriminating depressed from normal speech. Specifically, we use fundamental frequency (f_0) features in addition to measures based on the phase distortion [5] of the glottal source signal which characterizes glottal impulse variations. A major advantage of phase distortion is that it allows glottal source characteristics to be estimated from the speech signal without the need to estimate the glottal source signal itself. The statistic used in this study is the standard deviation of the phase distortion (PDD) that characterizes the shape of periodic pulses of the glottal source. PDD is associated with voice quality and has been used successfully as a quality assessment metric [6, 7, 8]; in [6] PDD was employed as an objective measure for quantifying spasmodic dysphonia, while in [8] the PDD spectrogram was introduced to distinguish highly intelligible speech (Lombard speech) from normal speech. Motivated by the correlation of PDD with the glottal source characteristics and its ability to quantify the degree of change of the glottal pulses, we propose PDD as a new metric for the detection and quantification of depressed speech.

Using an existing labelled corpus of depressed and normal speech in both read and spontaneous speaking styles (Section 3.3), we examine whether simple statistical properties of the f_0 and PDD feature distributions are able to separate depressed/normal speech (Section 4), and go on to employ a classification-based approach using a richer PDD feature set (Section 5). Classification between depressed and normal speech using the PDD and f_0 extracted features was performed with the fully-automated machine learning pipeline 'Just Add Data' [9] which produces a classification model given a training dataset and derives an estimate of its predictive performance.

2. Related Work

Research in depression detection has focused on two major domains for feature extraction, image and speech, employed individually or combined. Here we review previous studies using speech only. The third Audio/Visual Emotion Challenge [10] (AVEC-2013) incorporated a depression sub-challenge, aimed at predicting scores on Beck's Depression Inventory-II (BDI-II), leading to a regression problem for the continuous assessment of depression. The dataset consists of audio/visual recordings of volunteer participants, recorded while carrying out several powerpoint-guided tasks. The winning entry for the sub-challenge [11] exploited changes in correlations across formant frequencies and channels of the delta-mel-cepstrum using a Gaussian mixture model (GMM). Subsequent approaches using AVEC-2013 include Cummins et al. [12], who employed acoustic volume analysis combined with GMMs, Scherer et al. [13], who explore reduced vowel space as an indicator of distress, and the iVector-based approach of Lopez-Ottero et al. [14] applied to four depression severity classes.

The fourth Audio/Visual Emotion Challenge [15] (AVEC-2014) also featured a depression sub-challenge for both read

http://dx.doi.org/10.21437/Interspeech.2017-1251

and spontaneous speech styles. Approaches using AVEC-2014 include that of Mitra and Shriberg [16], who employed spectral, articulatory, phonetic, and prosodic features. The binary classification method of Statak et al. [17] assigned BDI-II scores of [0-9] to 'none-low depression' and [19-63] to 'moderate-high depression', and used Geneva Minimalistic Acoustic Parameter Set features [18] for estimation of arousal level, achieving a classification accuracy of 82.5%. Pampouchidou et al. [19] proposed binary categorical assessment, by considering BDI-II scores [0-13] as not-depressed, and [14-63] as depressed. Using the covarep toolbox [20] for extracting speech-based DCT features, they achieved an F1-score of 0.64 for gender-based depression classification.

3. Methods

The proposed method for assessing depression involves extraction of the phase distortion deviation, feature compression, and automated classification.

3.1. Phase distortion deviation

Two of the perceived voice qualities that might be expected to characterize depressed speech are hoarseness and breathiness. Such voice qualities are largely associated with voice production and the characteristics of the glottal source [21]. Recently, it has been shown that the characteristics of the glottal source can be estimated from the phase component of the speech signal introducing a novel feature highly correlated with voice quality. This feature is the phase distortion described in [7]. That study suggests that PDD could capture the characteristics of depressed speech, a hypothesis we test in the current study.

To estimate the PDD, speech is first decomposed into time-varying harmonic components, that is the time varying amplitudes, frequencies and phases. The adaptive Harmonic model described in [22] is used to extract the instantaneous phases from the speech signal. Then, in order to measure only the phase distortion corresponding to the glottal source signal, the minimum phase component related to the influence of the vocal tract is removed from the instantaneous phase. Finally, PDD is estimated using the method of [23] for the computation of standard deviation for circular data, as described in [20, 6]. Phase distortion is estimated at a frame rate of 10 ms, and the PDD is computed in a window of 3 periods, enabling it to track variations in the ongoing speech signal.

Figure 1 illustrates the PDD metric for samples of both normal and depressed speech, uttered by a healthy female speaker and a depressed female speaker respectively. It is evident that the normal speech has lower PDD values than those of the depressed speech, at least for frequencies in the range 0-5 kHz. This difference is most likely to be due to a reduced or absent harmonic structure during voiced segments of depressed speech, linked to the presence of hoarseness and breathiness. This observation supports our assumption that a depressed individual's voice shares certain features of hoarse and breathy voice. In hoarse voice, escaping air produces irregular vibrations of the vocal cords, while in breathy voice the aperiodic component is generated by air passing through the glottis. PDD may describe the resulting noisiness with a glottal locus.

We further observe that PDD differences between depressed and normal speech appear mainly during the voiced segments. Therefore, solely voiced areas are used in the current analysis. Voicing and f_0 are estimated using a robust voicing detection method, Summation of Residual Harmonics

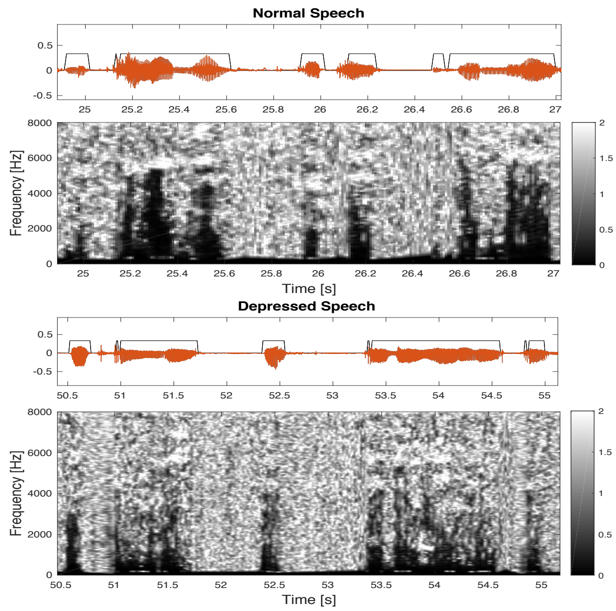

Figure 1: *Spectro-temporal plots of phase distortion deviation for normal and depressed speech. Regions of low distortion are mapped to darker shades of grey. Regions of voiced speech are indicated on the speech waveforms.*

[24], which uses the harmonic information of the residual signal for their estimation. Here f_0 values are limited to the range $70 - 500$ Hz.

3.2. Feature compression

The PDD is defined in the interval $[0, +\infty)$ and for each time instant (here, 10 ms) there are 512 frequency bins up to the Nyquist frequency (8 kHz); see Figure 1. In order to use the extracted PDD feature in a automated classification system to detect the presence or absence of depression, feature dimensionality has to be reduced. First, each PDD spectral slice is reduced to 12 Mel-Cepstral coefficients [25]. Second, we apply a discrete cosine transform along the time dimension to each Mel-Cepstral-transformed frequency feature, truncating at 20 coefficients. As a result, each speech sample is represented by a fixed-size parameter vector.

3.3. Datasets

The proposed method was evaluated on the Northwind (read speech) and Freeform (spontaneous speech) tasks of AVEC-2014. For the challenge, the dataset was organized in three partitions of 100 recordings each (training, development, testing) for both tasks. Labels of the testing partition were withheld for challenge purposes, so in the present study only the 200 recordings from the development and train sets were used. The dataset was annotated with a single label per recording, corresponding to speakers' scores on BDI-II, which according to its standardized cutoffs can be interpreted as minimal depression for a score of [0-13], mild [14-19], moderate [20-28], or severe [29-63]. The current binary classification approach uses two near-balanced subsets of not-depressed (BDI=[0-13], n=104), and depressed (BDI=[14-63], n=96) speech samples. Three of the Freeform recordings were excluded due to lack of a speech signal; all recordings were downsampled to 16 kHz.

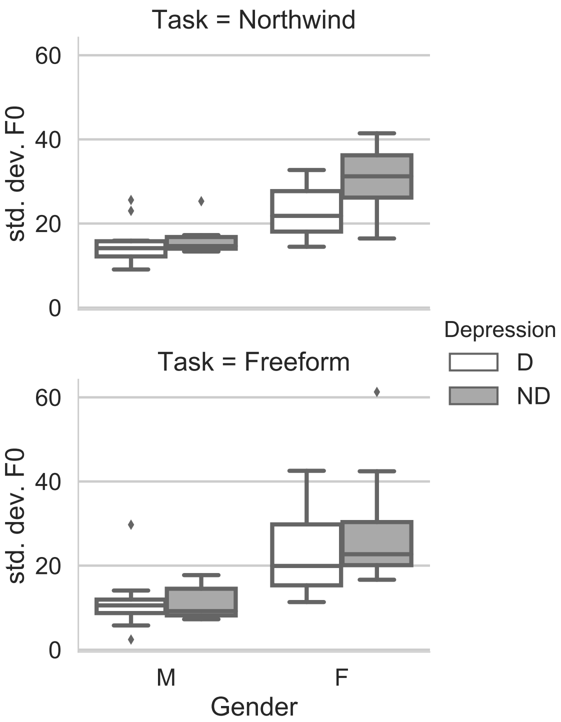

Figure 2: *Standard deviation of f_0 as a function of speaker gender, task and depression status. Data are presented as boxplots, with outliers indicated by diamonds.*

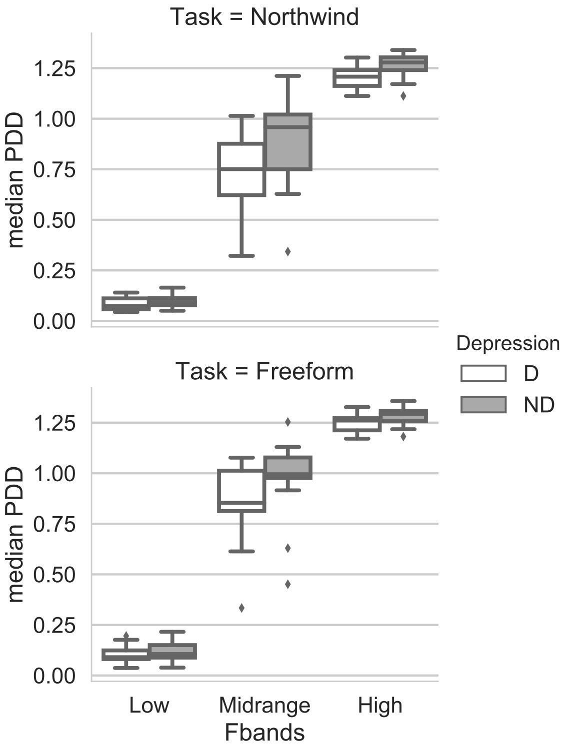

Figure 3: *Median PDD as a function of task, frequency band and depression status for female speakers.*

4. Analysis of PDD and f_0-based measures

To explore the discriminatory power of the proposed features in identifying depressed speech as a function of task, gender and frequency region, analyses of variance were performed on PDD and f_0 features. All speech material from each speaker was combined in order to estimate the PDD and f_0 measures. Ten subjects were excluded from the analysis due to either (i) possessing speech samples variously labelled as both depressed and normal, or (ii) where visual inspection of f_0 histograms suggested a failure of robust estimation (typically 'pitch' halving or doubling errors).

In all, 7 measures were analysed: mean, standard deviation, minimum and maximum of f_0, and mean, median and standard deviation of PDD. In the case of the f_0 features a robust estimation process was used to exclude outliers (defined as values lying outside 1.5 times the inter-quartile range). For the PDD measures, information was combined into 3 frequency bands, spaced quasi-logarithmically: low (0-750 Hz), midrange (750-2500 Hz) and high (2500-8000 Hz).

Mixed-effects analyses of variance with a between-subjects factor of depression status (D=depressed, ND=not depressed) and a within-subject factor of Task (Northwind, Freeform) for f_0- and PDD-based measures were carried out separately for male and female speakers. For the PDD measures an additional within-subject factor of frequency band was examined. The only statistically-significant outcomes involving the depression status factor are depicted in Figures 2 and 3.

Of the f_0-based measures, the standard deviation of f_0 showed an influence related to depression status, but only for female talkers in the Northwind read speech task (Figure 2) [$F(1, 29) = 5.56$, $p < .05$, $\eta^2 = 0.10$], although a similar tendency is apparent for the Freeform task. For the PDD-based measures, only the median and mean showed a statistically-significant difference dependent on depression status, and again only for the female talkers [$F(1, 29) = 4.85$, $p < .05$, $\eta^2 = 0.07$]. Differences were limited to the mid and high frequency bands (Figure 3) and were apparent in both tasks.

5. Automated classification

The performance of the proposed features in detecting depressed speech was tested using the Just Add Data (JAD) classification tool, v0.57 [9]. JAD performs multiple feature selection, training of different classification models and tuning of their hyper-parameters. Then, it automatically selects the best model and generates unbiased estimates of the mean performance and 95% confidence interval, using stratified K-fold cross-validation. Since the model was produced with the best configuration, its cross-validated performance is optimistic (biased) [26]. JAD removes the bias using a bootstrap method before returning the final performance estimate [27]. Note that the final performance estimates are slightly conservative. JAD was used without performing feature selection to maximize predictive performance. The tool trains several basic and advanced multivariate machine learning and statistical classification models: for classification problems, it uses Support Vector Machine

Table 1: *Classifier performance (AUC) for each dataset for three sets of features/combinations. Comp. PDD indicates compressed PDD features; PDD stats comprises mean, median and standard deviation of PDD; f_0 stats consists of mean, standard deviation, min and max f_0. The best male and female feature set results are highlighted in bold.* 95% *Confidence Intervals are provided in parentheses.*

Dataset	**comp. PDD**	**comp. PDD + f_0 stats**	**PDD stats & + f_0 stats**
Northwind	$0.68(0.53, 0.82)$	$0.67(0.52, 0.80)$	$0.75(0.60, 0.86)$
male only	$0.77(0.57, 0.90)$	$0.76(0.54, 0.89)$	$0.78(0.62, 0.90)$
female only	$0.75(0.57, 0.85)$	$0.75(0.59, 0.85)$	$0.76(0.60, 0.90)$
Freeform	$0.71(0.57, 0.84)$	$0.71(0.57, 0.83)$	$0.63(0.44, 0.73)$
male only	$0.72(0.48, 0.88)$	$0.72(0.48, 0.87)$	$0.75(0.51, 0.90)$
female only	$0.74(0.48, 0.84)$	$0.74(0.50, 0.84)$	$0.76(0.52, 0.85)$
Both datasets	$0.67(0.57, 0.73)$	$0.66(0.56, 0.72)$	$0.66(0.56, 0.73)$
male only	$\mathbf{0.87(0.77, 0.94)}$	$\mathbf{0.87(0.77, 0.94)}$	$\mathbf{0.88(0.78, 0.95)}$
female only	$\mathbf{0.79(0.75, 0.84)}$	$\mathbf{0.79(0.75, 0.84)}$	$\mathbf{0.76(0.67, 0.83)}$

models (SVMs) [28] with linear, full polynomial and Gaussian kernels, Ridge Logistic Regression [29] models, and Random Forests [30] models. Given that AVEC 2014 includes more than one recordings per participant, we have to avoid the bias introduced if the same participant belongs to more than one fold. For that, the JAD was used with the aforementioned constraint.

Three different sets of data were analyzed using JAD: (i) the 7 PDD/f_0 measures described in section 4; (ii) the compressed PDD features (12 MCEP coefficients x 20 DCT coefficients) described in section 3.2; and (iii) a combination of the compressed PDD features and the 4 f_0 features (mean, s.d., min and max). For each set, we performed three analyses based on using the entire set and the gender as a feature, or for each gender separately.

Table 1 summarises the classifier performance in terms of the area under the receiver operating characteristic curve (AUC). Not surprisingly given the source-derived nature of the features, gender-specific training results in the best classification performance for each dataset/feature set combination. Of the three feature sets, compressed PDD features show a slight advantage over PDD statistics. The addition of f_0 statistics to the compressed PDD features is not beneficial in any condition.

Given that in most cases the confidence intervals do not contain the point AUC=0.5, we infer that AUC is statistically significantly different than 0.5 (random classifier). Furthermore, in the best performing training results, the confidence intervals are higher than AUC=0.75.

6. Discussion

The current study demonstrates that features derived from the glottal source are valuable for the automated detection of depression in both read and spontaneous speech. While direct comparison with previous work is precluded by differences in experimental setups, our proposed system based solely on compressed PDD features appears that comparable to earlier studies. PDD-based features represent a small subset of potential audio-domain features; incorporation of complementary features such as those associated with speech rate or rhythm can be expected to lead to gains in classification performance.

The best outcomes are seen for gender-specific training. We speculate that given a more extensive corpus it may be possible to learn gender-conditioning at the classifier level. On the other hand, the analyses of individual features highlighted gender differences, at least for these datasets. Depressed speech from female speakers exhibited a significantly compressed speech range compared to normal speech, but no such effect was seen for male speakers.

The key measure tested in the current study, the standard deviation of phase distortion, shows a frequency-dependent pattern that distinguishes depressed and non-depressed speech. While informal observations (e.g., of Figure 1) suggest that these differences are present in the range 0-5 kHz, our statistical analysis suggests that it is in the mid- and high-frequency regions where the largest differences are apparent. It is possible that the lack of harmonic clarity that is hypothesised to be a feature of hoarse, breathy and depressed speech, is mainly disruptive in the region above F1 where harmonic amplitudes are relatively low.

7. Conclusions

We present a method to identify speech samples labelled as depressed based on glottal source parameters. Specifically, a time-frequency representation of the standard deviation of phase distortion, compressed to a fixed-size parameter vector by compression in both time and frequency, leads to a classification performance of 79 and 87% for female and male speakers respectively for read and spontaneous speech tasks of the AVEC-2014 challenge benchmark datasets. Potential developments of the proposed method include the inclusion of complementary features both within the audio domain and from the visual modality cf. [31, 19].

8. Acknowledgements

The authors would like to thank Prof. Ioannis Tsamardinos and Dr. Maria Koutsogiannaki for their valuable support. Anastasia Pampouchidou was funded by the “Maria Zaousi” scholarship from the Greek State Scholarships Foundation (I.K.Y). Olympia Simantiraki was funded by the European Commission under the Marie Curie European Training Network ENRICH (675324).

9. References

[1] R. C. Kessler, "The Costs of Depression," *Psychiatric Clinics of North America*, vol. 35, no. 1, pp. 1 – 14, 2012.

[2] R. C. Kessler and E. J. Bromet, "The Epidemiology of Depression Across Cultures," *Annual Review of Public Health*, vol. 34, pp. 119–138, 2013.

[3] J. M. Girard and J. F. Cohn, "Automated Audiovisual Depression Analysis," *Current Opinion in Psychology*, vol. 4, pp. 75 – 79, 2015.

[4] N. Cummins, 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, 2015.

[5] G. Degottex, A. Roebel, and X. Rodet, "Function of phase-distortion for glottal model estimation," in *Proc.IEEE Int. Conf. on Acoustics, Speech, and Signal Process-ing (ICASSP)*, 2011, pp. 4608–4611.

[6] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou, "The importance of phase on voice quality assessment," in *INTERSPEECH*, 2014, pp. 1653–1657.

[7] G. Degottex and O. Nicolas, "Phase distortion statistics as a representation of the glottal source: Application to the classification of voice qualities," in *INTERSPEECH*, 2014.

[8] G. Degottex, E. Godoy, and Y. Stylianou, "Identifying tenseness of Lombard speech using phase distortion," in *Proc. The Listening Talker: An interdisciplinary workshop on natural and synthetic modification of speech in response to listening conditions*, 2012.

[9] "Just Add Data; Gnosis Data Analysis PC," www.gnosisda.gr.

[10] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge," in *Proceedings of the 3rd International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. ACM, 2013, pp. 3–10.

[11] J. R. Williamson, T. F. Quatieri, B. S. Helfer, R. Horwitz, B. Yu, and D. D. Mehta, "Vocal Biomarkers of Depression Based on Motor Incoordination," in *Proceedings of the 3rd International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC'13. ACM, 2013, pp. 41–48.

[12] N. Cummins, V. Sethu, J. Epps, and J. Krajewski, "Probabilistic Acoustic Volume Analysis for Speech Affected by Depression," in *INTERSPEECH*, 2014, pp. 1238–1242.

[13] S. Scherer, L. P. Morency, J. Gratch, and J. Pestian, "Reduced Vowel Space is a Robust Indicator of Psychological Distress: A Cross-Corpus Analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4789–4793.

[14] P. Lopez-Otero, L. Docio-Fernandez, and C. Garcia-Mateo, "Assessing Speaker Independence on a Speech-based Depression Level Estimation System," *Pattern Recognition Letters*, vol. 68, Part 2, pp. 343 – 350, 2015.

[15] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "AVEC 2014: 3D Dimensional Affect and Depression Recognition Challenge," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '14. ACM, 2014, pp. 3–10.

[16] V. Mitra and E. Shriberg, "Effects of feature type, learning algorithm and speaking style for depression detection from speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4774–4778.

[17] B. Stasak, J. Epps, N. Cummins, and R. Goecke, "An Investigation of Emotional Speech in Depression Classification," *INTERSPEECH*, pp. 485–489, 2016.

[18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andr, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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.

[19] A. Pampouchidou, O. Simantiraki, C.-M. Vazakopoulou, C. Chatzaki, M. Pediaditis, A. Maridaki, K. Marias, P. Simos, F. Yang, F. Meriaudeau, and M. Tsiknakis, "Facial Geometry and Speech Analysis for Depression Detection," in *39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2017*, Submitted.

[20] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - A Collaborative Voice Analysis Repository for Speech Technologies," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 960–964.

[21] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge* , ser. AVEC'14. ACM, 2014, pp. 65–72.

[22] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," in *IEEE Trans. on AudioSpeech and Lang. Proc.*, 2013, pp. 2085–2095.

[23] N. Fisher, "Statistical Analysis of Circular Data." Cambridge University Press, 1995.

[24] T. Drugman and A. Alwan, "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," in *INTERSPEECH*, 2011.

[25] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Acoustics, Speech, and Signal Processing, 1992. (ICASSP),IEEE International Conference*, 1992, pp. 137–140.

[26] I. Tsamardinos, A. Rakhshani, and V. Lagani, "Performance-Estimation Properties of Cross-Validation-Based Protocols with Simultaneous Hyper-Parameter Optimization," in *International Journal on Artificial Intelligence Tools*, 2015, pp. 1–30.

[27] B. B. Efron, "Bootstrap Confidence Intervals," in *Journal of the American Statistical Association* , 1987, pp. 171–185.

[28] B. Boser, I. Guyon, and V. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[29] A. Hoerl and R. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," in *Technometrics* , 1970, pp. 55–67.

[30] L. Breiman, "Random forests," in *Machine Learning* , 2001, pp. 5–32.

[31] A. Pampouchidou, O. Simantiraki, A. Fazlollahi, M. Pediaditis, D. Manousos, A. Roniotis, G. Giannakakis, F. Meriaudeau, P. Simos, K. Marias, F. Yang, and M. Tsiknakis, "Depression Assessment by Fusing High and Low Level Features from Audio, Video, and Text," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '16. ACM, 2016, pp. 27–34.