WEIGHTED PAIRWISE GAUSSIAN LIKELIHOOD REGRESSION FOR DEPRESSION SCORE PREDICTION

Nicholas Cummins[1,2], Julien Epps[1,2], Vidhyasaharan Sethu[1], Jarek Krajewski [3,4]

[1]School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney Australia
[2] ATP Research Laboratory, National ICT Australia (NICTA), Australia
[3] Experimental Industrial Psychology, University of Wuppertal, Wuppertal, Germany
[4] Industrial Psychology, Rhenish University of Applied Sciences Cologne, Germany
n.p.cummins@unsw.edu.au, j.epps@unsw.edu.au

ABSTRACT

This paper presents a technique in which feature vectors are mapped onto ordinal ranges of clinical depression scores using weighted pairwise Gaussians. The position of a test vector with respect these partitions is used to perform depression score prediction. Results found on a set of spectral and formant based speech characteristics indicate the potential of this technique for performing depression score prediction. Key results on the AVEC 2013 development set indicate that the inclusion of weights and Bayesian adaptation improves system performance by 16.5% - 18.5% when compared to using an unweighted non-adapted system and fusion of multiply models corresponding to different feature spaces can offer up to 8% further improvement. Further, fusion consistently improves performance on both the AVEC 2013 development and test set, in contrast to conventional regressor fusion.

Index Terms— Depression, Score Level Prediction, Gaussian, Bayesian Adaptation, Fusion

1. INTRODUCTION

Clinical depression is the world's most commonly occurring mental illness [1]. Symptoms include, but are not limited to; reduced cognitive ability, continuous negative affect, increased fatigue and psychomotor retardation. This heterogeneous clinical profile introduces complexities when trying to assign a depressed individual into a clinical categorical grouping which relates to their level of depression [2]. As a result of these diagnostic complexities, research into the use of behavioural signals – such as speech – as a diagnostic aid for depression is gaining in popularity [3]–[5].

Currently clinicians use a variety of different depression assessment methods, such as the *Beck Depression Inventory* (BDI [6]) – a questionnaire in which the severity of 21 symptoms commonly observed in depression are self-rated then summed together to produce a score which relates to depression severity. The reliability of self-reported assessment such as the BDI is compromised by patient overfamiliarity and subjectivity when rating the severity of their symptoms [7].

Both the heterogeneity of depression symptoms [8] and issues relating to reliability mean that a severity score given through multi component depression assessments such as the BDI can be regarded as ordinal rather than numerical [9]. Often the severity score is used for ease of clinical categorization rather than as a reliable *numerical* severity measure. A patient who has taken the BDI is assigned to one of four clinical categories depending on their total score; *minimal* (0–13), *mild* (14–19), *moderate* (20–28), and *severe* (29–63).

The recent *Audio Visual Emotion Challenge* (AVEC 2013, [10]) required participants to predict a single BDI score using multimodal signal processing techniques from a given multimedia file. Given the ordinal nature of BDI scores, it seems reasonable to assume that direct regression on feature space coefficients would be unsuitable – there are no guarantees of correlation, in the parametric sense, between behaviour feature space coefficients and BDI scores.

A potential solution to increase the performance of the predicted depression scores is to assume that feature vectors relating to adjacent BDI scores do not differ greatly from each other, whilst feature vectors for scores separated by a considerable margin are distinct, and create a "staircase" of feature space probability density functions with each "step" of the staircase representing a different range of BDI scores [5]. The information from a series of estimates of how far from each "step" the test point lies can then be used as a the basis for regression, rather than raw features or model-based features like supervectors.

2. RELATION TO PRIOR WORK

This paper explores and extends the *Gaussian Staircase Regression* (GSR) paradigm proposed in [5], in combination with spectral and formant features, for performing depression score level prediction. In this paradigm, pairs of Gaussians are used to partition the feature space into regions corresponding to predefined BDI ranges. A test vector derived from likelihood ratios based on these pairwise partitions is then used to obtain a final score using a suitable regression model.

Spectral and formant speech features have shown consistently strong performance when used as objective marker of depression [4], [5], [11]–[15]. This consistency is due to depression altering speech motor control, in particular vocal tract and articulatory coordination [5], [14]. As results published in AVEC 2013 Depression Recognition dataset [10] confirm the suitability of spectral and formant features for predicting a speaker's level of depression [4], [5], [14], we will use similar features in this paper. Although the *Vocal Tract Correlation* (VTC) features used in [5] are interesting, in this work we specifically wanted to understand the benefits of both GSR and our proposed system on a more conventional set of speech features.

An opportunity offered by expanding the GSR technique is feature fusion. Successful regression fusion needs an independent set of predictors to avoid the multicollinearity problem, however, as all learned models are attempting to perform the same prediction task, prediction scores from multiple back-ends are inherently correlated [16]. Therefore if GSR likelihood values are less correlated with BDI scores than predictor outputs, advantages may be found in fusion in the likelihood domain. Fortunately, fusion in

the likelihood domain is well-studied in the speech *classification* literature, e.g. [17]–[19].

The aim of this paper is to understand the benefits of GSR (Section 3), our proposed weighted GSR system (Section 4) and both likelihood fusion and predictor fusion (Section 5) when predicting a speaker's level of depression. In achieving this aim we present a series of prediction tests (Section 7) performed under the AVEC 2013 test conditions [10].

3. GAUSSIAN STAIRCASE REGRESSION

The GSR approach uses an ensemble of pairs of Gaussians to divide features vectors along the BDI scale in a staircase fashion [5]. The set of Gaussians from the lower part of the scale, $\{G_i^L\}_{i=1,\cdots,N}$, are referred to as the *Low Gaussian Staircase* and the set of Gaussians from the higher part of the scale, $\{G_i^H\}_{i=1,\cdots,N}$, are referred to as the *High Gaussian Staircase,* where N is the total number of Gaussian pairs (Figure 1). At each 'step' the test statistic is the *log mean likelihood ratio* (LMLR) of the low Gaussian G_i^L to the high Gaussian G_i^H. Given a set of vectors; $X = \{\boldsymbol{x}_1 \cdots \boldsymbol{x}_T\}$ the LMLR for the *i*-th step is given by:

$$LMLR_i = log\left\{\frac{\frac{1}{T}\sum_{t=1}^{T} P(x_T|G_i^L)}{\frac{1}{T}\sum_{t=1}^{T} P(x_T|G_i^H)}\right\} \quad (1)$$

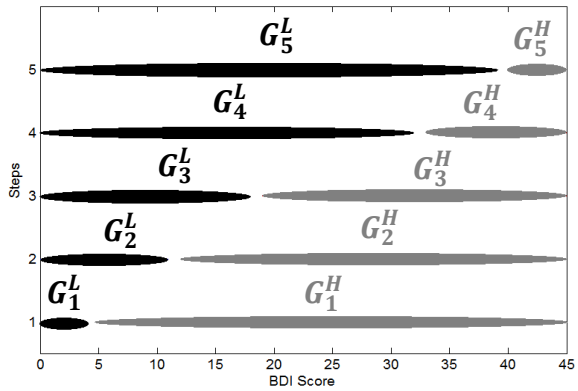

Figure 1: A five step Gaussian Staircase. The black partitions create the Low Gaussian Staircase and the grey partitions create the High Gaussian Staircase. Figure adapted from [5].

The training of the GSR system involves three steps (Figure 2). First a pair of low and high *Background Gaussian Staircases* (BGS) are created, using specific predefined partitions of training dataset, by training a pair of Gaussians per step.

The second step of the GSR training is to adapt the means of the BGS to form a *file-specific* GSR for each file in the training data set. This adaptation attempts to account for differing speech characteristics between files in the training set which otherwise might limit overall accuracy. Given a set of training vectors; $\boldsymbol{X} = \{\boldsymbol{x}_1 \cdots \boldsymbol{x}_T\}$, the mean of each Gaussian is updated one frame at a time

$$\widehat{\boldsymbol{\mu}}_i^c = (1 - \alpha_m^c)\boldsymbol{\mu}_i^c + \alpha_m^c \boldsymbol{x}_t \quad (2)$$

where $\boldsymbol{\mu}$ denotes the mean being adapted, $\widehat{\boldsymbol{\mu}}$ the updated mean, $c \in \{L, H\}$ denotes the low and high GSRs, i =1,…,N and α_m^c the staircase specific adaptation coefficient for the *m*-th file. Note: as the adaption methodology is not presented in [5] we tested a series of non-Bayesian adaptation methods, and (2) gave the most consistent performance.

Each α_m^c is set on a *per-speaker* basis:

$$\alpha_m^c = \frac{n}{0.5 + n} \quad (3)$$

where n is the total number of frames the *speaker* of the *m*-th file has contributed to staircase c [5]. Note, for some subjects the AVEC 2013 dataset contains multiple files with their speech at different levels of depression therefore, staircase c could contain multiple files from the same subject.

The final training step is to calculate test statistics for each file in the training set using each file's adapted models, and use the resulting test statistics and corresponding BDI score to train a regression model. To predict the BDI score for a given test file, the same mean adaptation procedure is used before calculating the mean test statistic and predicting the BDI using the regression model:

$$\tilde{Y} = f(\boldsymbol{V}, \boldsymbol{\beta}) \quad (4)$$

where $\boldsymbol{V} = [LMLR_1 \ldots . LMLR_N]^T$ is a vector of LMLR's obtained from each step, $\boldsymbol{\beta}$ the estimated regression parameters and $\tilde{Y}$ the predicted BDI scores.

4. WEIGHTED GAUSSIAN STAIRCASE REGRESSION

This paper expands on the GSR approach by the inclusion of weighting parameters for the individual Gaussians and the use of speaker specific parameter adaptations similar to GMM-MAP adaptation [20]. The inclusion of weights in particular may be significant since the BDI scores present in the AVEC 2013 dataset – training and development partitions – are skewed toward lower BDI scores [4].

When forming the low and high BGS for the proposed *Weighted Gaussian Staircase Regression* (WGSR) a weight parameter, ω_i^c, is combined with each Gaussian which is proportional to the percentage of the overall data used to train that Gaussian compared to the overall amount of data that went into training the set – low or high - that the Gaussian is an element of. The *i*-th WGSR test statistic is the *weighted log mean likelihood ratio* (wLMLR) of the low Gaussian G_i^L to the high Gaussian G_i^H:

$$wLMLR_i = log\left\{\frac{\widehat{\omega}_i^L \frac{1}{T}\sum_{t=1}^{T} P(\boldsymbol{x}_t|G_i^L)}{\widehat{\omega}_i^H \frac{1}{T}\sum_{t=1}^{T} P(\boldsymbol{x}_t|G_i^H)}\right\} \quad (5)$$

where ω_i^c denotes the *i*-th weight parameter.

When forming the individual WGSR for each file, the means and weights of each set of Gaussians are updated to better match the target file by an amount dependent on the posterior probability of step i. Given an individual BGS Low or High Staircase and a set of adaptation vectors $\boldsymbol{X} = \{\boldsymbol{x}_1 \cdots \boldsymbol{x}_T\}$, the posterior probability of each step given each frame in $\boldsymbol{X}$ is calculated by:

$$P(G_i^c|\boldsymbol{x}_t) = \frac{\omega_i^c \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_i^c, \boldsymbol{\Sigma}_i^c)}{\sum_{j=1}^{N} \omega_j^c \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_j^c, \boldsymbol{\Sigma}_i^c)} \quad (6)$$

where G_i^c represents an individual Gaussian and $\boldsymbol{\mu}_i^c$, $\boldsymbol{\Sigma}_i^c$ and ω_i^c denote that Gaussian's mean vector, covariance matrix and weighting coefficient respectively.

The means and weights are then updated using:

$$\widehat{\boldsymbol{\mu}}_i^c = (1 - \alpha_m^c)\boldsymbol{\mu}_i^c + \alpha_m^c \left[\frac{1}{\zeta_i}\sum_{t=1}^{T} P(G_i^c|\boldsymbol{x}_t)\boldsymbol{x}_t\right] \quad (7)$$

$$\widehat{\omega}_i^c = \left[(1 - \alpha_m^c)\omega_i^c + \alpha_m^c \frac{\zeta_i}{T}\right]\varphi \quad (8)$$

where $\boldsymbol{\mu}$ and ω denote the mean and weight being adapted, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\omega}}$ the updated mean and weight, α_m^c is given by (3), φ is a scaling factor to ensure weighting components of each staircase sum to unity and ζ_i is the weight statistic given by:

$$\zeta_i = \frac{1}{T}\sum_{t=1}^{T} P(G_i^c|\boldsymbol{x}_t) \quad (9)$$

Note, whilst it is also possible to update the covariance matrices, this was not done for the results presented in this paper.

5. LIKELIHOOD AND PREDICTION FUSION

Three separate fusion methods are trialed in this paper. The first is fusion in the *likelihood domain* (LHD), in which the set of (w)LMLR's obtained from different feature spaces and a single regression model is trained to operate on the concatenated (w)LMLR vector (Figure 2).

We also trial conventional regression fusion, in which an overall prediction is formed through the weighted sum of different regressor outputs:

$$\tilde{Y} = \sum_{k=1}^{K} \delta_k f_k(\boldsymbol{V_k}, \boldsymbol{\beta_k}) \quad (10)$$

where $\tilde{Y}$ are the predicted BDI scores, f_k the prediction from the *k*-th regression model and δ_k the weight of that prediction to *Y*. Two methods of *regression* model fusion are trialed; *Basic Ensemble Method* (BEM) in which all predictions are given equal weight and *Linear Regression* (Lin Reg.) in which the δ's are estimated via least square analysis. For complete details of either method see [16].

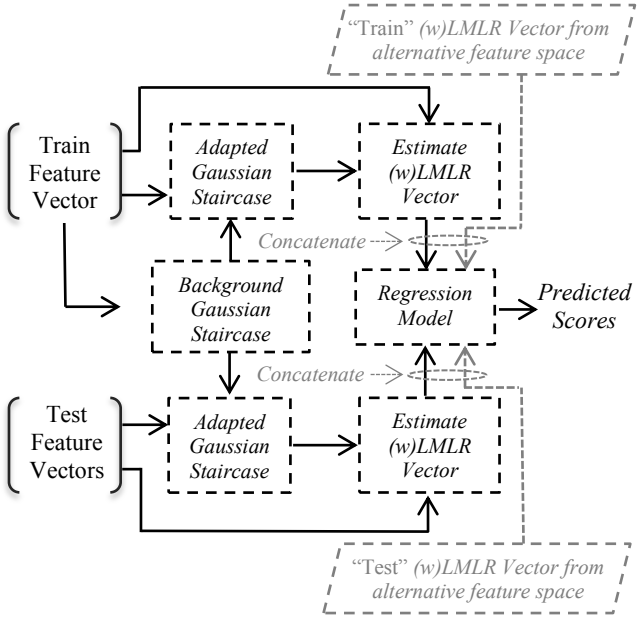

Figure 2: Block diagram showing the training and testing phases of a Gaussian Staircase Regression System. Also shown in grey is likelihood domain fusion, in which different sets of (w)LMLR's are concatenated together before regression.

6. EXPERIMENTAL SETTINGS

The AVEC 2013 dataset consists of 150 files with a mean file length of 14 min 52secs, divided into training, development and testing partitions each of 50 files (note: unless stated all results are reported on the development set). All files are labelled with a BDI score ranging from 0-45. All prediction results are reported in terms of the prediction *Root Mean Squared Error* (RMSE)**.**

Four different feature spaces are used: *Mel Frequency Cepstral Coefficients* (MFCC's), *MFCC Shifted Delta Coefficients* (SDC), *Formants* (FMT) and *Spectral Centroid Frequencies* (SCF). MFCC's and FMT's were extracted using the openSMILE toolkit [21]. Thirteen MFCC's, including C0, were extracted and appended with Δ and ΔΔ coefficients. Eighteen formant features were used, being the first three formant frequency and bandwidths appended with Δ and ΔΔ coefficients. The SDC's were extracted with parameters *N-d-p-k* equal to 13-1-3-7 [22]. SCF is a measure of the average weighted frequency for the *k*-th sub-band, where the weights are the normalized energy [23]. Twenty SCF's were extracted and appended with Δ and ΔΔ coefficients using a mel-scale Gabor filterbank as per [23]. Only voiced frames were used in the modeling, determined using openSMILE's voicing probability function.

Unless otherwise stated, the BGS was formed using the entire AVEC 2013 training partition. As in [5], N = 5 Gaussians were estimated per staircase with the Low Class partitions at each step being 0-4, 0-11, 0-18, 0-32 and 0-39 and High Class partitions at each step being 5-45, 12-45, 26-45, 33-45 and 40-45. As per [5], full covariance matrices were used per division with 0.2 added to diagonal terms for improved regularization. The adaptation file for forming the individual speaker staircases was a read passage taken from the novel "Homo Faber" by Max Frisch, mean length 3 min 24 secs. Note that this passage was unavailable for two speakers in the test set, as in [5], [14] and an appropriate length of continuous read speech was used instead.

Unless stated, all regression equations were generated using the training partition. For GSR, as per [5], the mean of each file's LMLR's was used to train a linear regression model:

$$LMLR_{mean} = 1/N \sum_{i=1}^{N} LMLR_i \quad (11)$$

For the WGSR the mean and weighting parameters were updated using 1, 2, 5 or 10 adaptation iterations (iter) and all 5 test statistics were used to train a linear regression model. For Lin Reg fusion the weights were trained using the training set.

We compare the GSR and WGSR results with a set of "brute-forced" predictions (No-GSR). The No-GSR predictions were generated by training a regression system directly on a feature space comprised of 16 functions, [13], evaluated from either the MFCC, SDC, FMT and SCF features. The regressor was LIBSVM's *epsilon Support Vector Regressor* (ε-SVR) in combination with a linear kernel, and default hyper-parameter settings [24].

7. RESULTS

7.1. Feature Space - Development Set Results

The No-GSR systems, with the exception of the one using SDCs, outperform the AVEC 2013 challenge (Table 1). It is likely that the dimensionality of the SDCs made them unsuitable for this approach. The SCF results were surprising, offering one of the lowest development scores for this dataset in the literature. We speculate this result could be due in part to the robustness of the SCF feature to recording conditions [23]. SCFs have also be shown to be suitable for discriminating between the presence and absence of depression in speech [11].

The GSR *no adaptation* results are all round chance level (RMSE of 11.90 [10]) but the inclusion of speaker adaptation, using (2), improved the performance of all features except MFCC (Table 1). These results are consistent with those presented in [5], - Formant based VTC features gave an RMSE of 9.99 before adaption and 8.40 after. We speculate the lower RMSE's presented in [5] are due to the difference feature spaces, the VTC is specifically designed to exploit changes in articulatory coordination commonly seen in depression.

WGSR offers a significant improvement over GSR – again with the exception of SDC – (Table 1), with the lowest RMSE for each feature space well below chance level and the challenge baseline. We speculate that the difference in the pattern of results for SDC is due to its higher dimensionality. Preliminary results not given indicated the importance of using and updating the weights. Given the skewness of AVEC 2013 development set towards lower BDI scores [4], we speculate that including updating weights helps the system adjust for the data imbalance among the different steps.

Table 1. No-GSR and GSR and WGSR RMSE on AVEC 2013 Development Set (AVEC 2013 Baseline 10.75 [10])

System	MFCC	SDC	FMT	SCF
No-GSR	9.96	11.92	10.66	8.70
GSR - No adapt.	11.17	11.74	11.57	11.22
GSR	11.60	9.56	10.52	10.35
WGSR - 1 iter	9.67	9.57	11.34	10.98
WGSR - 2 iter	9.41	10.89	11.13	10.48
WGSR - 5 iter	9.38	10.71	10.56	9.52
WGSR - 10 iter	9.33	10.91	9.67	9.37

7.2. No-GSR and WGSR Fusion - Development Set Results

A benefit of the WGSR system can be seen in the fusion results (Table 2). *None of the fusion methods* trialed on the No-GSR systems significantly lowered the SCF result, whilst *all fusion methods* trialed for WGSR consistently lowered the RMSE's obtained. These results are also consistent with those presented in [5] where fusion consistently improved their system performance.

Fusion of the WGSR feature space predictions gave the most consistent performance and reduced the RMSE's below the best No-GSR result. A combination of MFCC, FMT and SCF with BEM fusion gave our best development set result offering an 8% reduction in RMSE when compared to the best WGSR result and 23% reduction when compare to the best GSR no adaptation result. We speculate this combination of features covers a diverse variety of input spaces for depression prediction.

Note that all fusion results were generated using 10 adaptation iterations on the individual feature spaces; replacing SDC 10 iterations with the SDC 1 iteration did not improve performance.

Table 2. No-GSR and WGSR and Fusion RMSE's on AVEC 2013 Development Set (AVEC 2013 Baseline 10.75 [10])

Fusion System	ALL	MFCC + FMT+ SCF	MFCC+SCF
Feature Space *No-GSR*	8.99	9.00	8.67
BEM - *No-GSR*	9.22	9.31	8.71
Lin Reg *No-GSR*	10.54	10.59	10.15
LHD - *WGSR*	9.04	8.70	8.74
BEM -*WGSR*	8.68	**8.55**	8.60
Lin Reg-*WGSR*	8.75	8.66	8.70

7.3. Test Set Results

All test set RMSEs are well below the AVEC 2013 baseline and chance level RMSE of 11.51 (Table 3). Our best performance – 9.75 – was obtained using the BEM fusion of MFCC's, FMT' and SCF's. To the best of the author's knowledge this is the second best test set result in the literature (Table 3).

The results are consistent with the development set, confirming the suitability of WGSR for depression score prediction. The unfused WGSR results indicate that the technique is *less susceptible to overfitting* when compared to No-GSR SCF and challenge baseline. As in the development set results fusion improves system performance.

Note that for the No-GSR system both the training and development sets were used to train the SVR. For WGSR both the training and development partitions were used to train the BGS and regression model.

Table 3. Key results from literature and WGSR RMSE's on AVEC 2013 Test Set

System	RMSE
Baseline [10]	14.12
MFCC-based Supervector [4]	10.17
Canonical Correlation Analysis [25]	9.78
VTC fused with Phoneme Rate [5]	8.50
No-GSR (SCF)	10.96
WGSR - SCF (10 iter)	10.26
WGSR - MFCC (10 iter)	10.23
WGSR- BEM (ALL)	9.90
WGSR-LHD (MFCC+FMT+SCF)	9.76
WGSR-BEM (MFCC+FMT+SCF)	9.75

8. CONCLUSION

This paper presents WSGR a regression method, in which feature vectors are mapped, using weighted and individually adapted pairwise Gaussians, onto multiple partitions of a feature space before using the output of this mapping to perform depression score prediction. Results indicate that WSGR offers useful solution to dealing with two commonly associated with depression score prediction – ordinal clinical scales for predicting against and skewed datasets. WGSR offers several promising research directions including exploring the effect of changing step partitions, adaptation coefficient and the regularizing coefficient.

A key result presented is that fusion of WGSR systems consistently improved system performance which was not seen in the fusion of our No-GSR systems. The strong fusion results indicate the need to explore both fusion paradigms in more details, in particular weight fusion of likelihood ratios – a method commonly used in speaker verification tasks [18]. Future work will also include; an exploration into the benefits offered to depression analysis through VTC features and the use of both different probability distributions and classification paradigms to perform pairwise analysis.

9. ACKNOWLEDGEMENTS

This work was partly funded by ARC Discovery Project DP130101094 (Epps) and the German Research Foundation (KR3698/4-1) (Krajewski). NICTA is funded by the Australian Government as represented by the Department of Broadband, Communication and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. The authors wish to thank Kirsty Smith from the University of Nottingham for calculating the AVEC 2013 test set results.

10. REFERENCES

[1] C. D. Mathers, and D. Loncar, "Projections of Global Mortality and Burden of Disease from 2002 to 2030," *PLoS Med*, vol. 3, no. 11, pp. 2011–2030, 2006.

[2] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: a meta-analysis," *Lancet*, vol. 374, no. 9690, pp. 609–619, 2009.

[3] S. Scherer, G. Stratou, G. Lucus, M. Mahmoud, J. Boberg, J. Gratch, A. (Skip) Rizzo, L.-P. L. P. Morency, and G. Lucas, "Automatic Audiovisual Behavior Descriptors for Psychological Disorder Analysis," *Image Vis. Comput.*, vol. 32, no. 10, pp. 1–21, Oct. 2014.

[4] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of Depression by Behavioural Signals: A Multimodal Approach," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 11–20.

[5] 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 ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 41–48.

[6] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients," *J. Pers. Assess.*, vol. 67, no. 3, pp. 588–597, Dec. 1996.

[7] C. Cusin, H. Yang, A. Yeung, and M. Fava, "Rating Scales for Depression," in *Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health SE - 2*, L. Baer, and M. A. Blais, Eds. Humana Press, 2010, pp. 7–35.

[8] S. D. Østergaard, S. O. W. Jensen, and P. Bech, "The heterogeneity of the depressive syndrome: when numbers get serious," *Acta Psychiatr. Scand.*, vol. 124, no. 6, pp. 495–496, Dec. 2011.

[9] D. Faries, J. Herrera, J. Rayamajhi, D. DeBrota, M. Demitrack, and W. Z. Potter, "The responsiveness of the Hamilton Depression Rating Scale," *J. Psychiatr. Res.*, vol. 34, no. 1, pp. 3–10, Jan. 2000.

[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 ACM International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 3–10.

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

[12] N. Cummins, J. Epps, V. Sethu, M. Breakspear, and R. Goecke, "Modeling Spectral Variability for the Classification of Depressed Speech," in *Proceedings of Interspeech*, 2013, pp. 857–861.

[13] N. Cummins, J. Epps, and E. Ambikairajah, "Spectro-Temporal Analysis of Speech Affected by Depression and Psychomotor Retardation," in *Proceedings of ICASSP*, 2013, pp. 7542–7546.

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

[15] B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, and B. Yu, "Classification of depression state based on articulatory precision," in *Proceedings of Interspeech*, 2013, pp. 2172–2176.

[16] C. Merz, and M. Pazzani, "A Principal Components Approach to Combining Regression Estimates," *Mach. Learn.*, vol. 36, no. 1–2, pp. 9–32, 1999.

[17] K. Kinnunen, and H. Li, "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors," *Elsevier speech Commun.*, vol. 52, no. 1, pp. 12–40, 2009.

[18] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 7, pp. 2072–2084, Sep. 2007.

[19] N. Poh, and S. Bengio, "Why do multi-stream, multi-band and multi-modal approaches work on biometric user authentication tasks?," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004, vol. 5, pp. 893–896.

[20] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digit. Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[21] F. Eyben, M. Wollmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the International Conference on Multimedia*, 2010, pp. 1459–1462.

[22] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. Deller R., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *Proc. Int. Conf, Spok. Lang. Proc*, pp. 88–92, 2002.

[23] J. M. K. Kua, T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition," *Proc. Odyssey Speak. Lang. Recognit. Work. (Brno, Czech Repulic)*, pp. 34–39, 2010.

[24] C.-C. Chang, and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, p. 27, 2011.

[25] H. Kaya, F. Eyben, and A. A. Salah, "CCA based Feature Selection with Application to Continuous Depression Recognition from Acoustic Speech Features," in *Proceedings of ICASSP*, 2014, pp. 3757–3761.