Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation

Zafi Sherhan Syed
Cardiff University
United Kingdom
syedmz@cardiff.ac.uk

Kirill Sidorov
Cardiff University
United Kingdom
sidorovk@cardiff.ac.uk

David Marshall
Cardiff University
United Kingdom
marshallad@cardiff.ac.uk

ABSTRACT

This paper addresses the AVEC 2017 – Depression Sub-Challenge, where the objective is to propose methods which can aid automated prediction of depression severity. In this paper, we specifically focus on biomarkers of psychomotor retardation, which are a key trait of depressive episodes, to propose three sets of methods.

We propose a novel set of temporal features (which we called "turbulence features") and show their effectiveness. We offer a novel methodology to target specific craniofacial movements indicative of psychomotor retardation and hence of depression. Further, we present a novel method for quantifying abnormalities of speech spectra of individuals with depression using Fisher vector encoding of spectral low level descriptors (LLDs).

So far, in the AVEC challenge on prediction of patient health questionnaire (PHQ) scores on the Test set, we achieve a root mean square error (RMSE) score of **6.34** and a mean absolute error (MAE) score of **5.30**, both of which are better than the best results on the AVEC test set as given in the baseline paper i.e. **6.97** and **5.66**, respectively. This suggests that our method is a viable proof of concept and may lead to fully automated objective depression screening protocols.

CCS CONCEPTS

• **Computing methodologies** → *Machine learning approaches*; • **Applied computing** → *Psychology*;

KEYWORDS

Social Signal Processing, Affective Computing, Depression Screening, Machine Learning, AVEC 2017

ACM Reference Format:
Zafi Sherhan Syed, Kirill Sidorov, and David Marshall. 2017. Depression Severity Prediction Based on Biomarkers of Psychomotor Retardation. In *Proceedings of AVEC'17, Mountain View, CA, USA, October 23, 2017,* 7 pages. https://doi.org/10.1145/3133944.3133947

1 INTRODUCTION

Depression is a serious mental illness, which according to the World Health Organisation (WHO) affects more than 300 million individuals worldwide and is the leading cause of disability [38]. At its worst, depression can trigger thoughts of suicide and is directly blamed for around 800,000 deaths every year. Individuals who suffer from depression often face unemployment due to their inability to work and are susceptible to alchohol abuse [18]. Furthermore, long-term depression also increases the risk of dementia and Alzheimer's disease [16, 23].

Psychologists have convincingly shown that depressed individuals differ from non-depressed control groups with regard to objectively quantified gross motor activity, body movements, speech, and motor reaction time [33].

According to the "The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)" [3], psychomotor symptoms form core features of depression. Psychomotor retardation is one such symptom which can impair social functioning of individuals who suffer from this illness [39]. Psychomotor retardation affects not only the emotional behaviour of individuals along with related motor processes but also causes musclular contractions which slows down physical movement [26, 32]. Therefore, most work on automated recognition of depression [15], implicitly or explicitly is based on recognising psychomotor retardation.

This paper addresses the Depression Sub-Challenge (DSC), which is part of the Audio-Visual Emotion recognition Challenge 2017 (AVEC 2017) [30]. In the AVEC 2017 – DSC, participants are required to assess the depression severity of the interviewed subject, where the target depression severity is based on the 8 point patient health questionnaire (PHQ-8) scores recorded prior to interview sessions.

The broader goal of the AVEC 2017 workshop is to compare the relative merits of the approaches for audio-visual emotion recognition and severity of depression estimation. Therefore in line with goals of the workshop, our paper contributes to the workshop by honing in on the symptoms of psychomotor retardation to craft features which can be used to predict depression severity scores.

The contribution of this paper towards automated screening of depression are as follows:

(1) We propose a novel approach which measures turbulence in speech feature trajectories, which we call "turbulence features", and show that they are effective in predicting depression severity.
(2) We detail a novel methodology to measure specific craniofacial movements over multiple temporal resolutions which are targetted to measure retardation of facial muscles during a depressive episode. This is a departure from a tradition of naively modelling facial movements.

AVEC'17, October 23, 2017, Mountain View, CA, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-5502-5/17/10…$15.00
https://doi.org/10.1145/3133944.3133947

(3) Further, we present a novel method for quantifying abnormalities of speech spectra of individuals with depression using Fisher vector encoding of spectral low level descriptors (LLDs).
(4) Finally, by proposing a set of pipelines which are targeted to capture symptoms of psychomotor retardation rather than naively learn models from a dataset through black-box machine learning, we believe that our inherently intuitive approach will aid the larger research community working in the field of depression recognition in seeking to build more objective measures for depression rather than existing self-assessment forms.

The rest of the paper is organised as follows. In Section 2, we briefly discuss the dataset. This is followed by a discussion of challenges faced when developing methods for automated depression screening. We detail the methodology in section 4. Results of our experimentation are discussed in Section 5, followed by conclusions in subsequent section.

2 DATASET

In this work, we use the dataset provided for AVEC 2017 – DSC is part of the larger Distress Analysis Interview Corpus - Wizard of Oz (DAIC-WOZ) [17], which is a collection of dyadic interviews of individuals conducted by a virtual agent, called "Ellie". Participants of the interview process were asked fill in the 8-point Patient Health Questionnaire (PHQ), which is a self-assessment report form for depression screening [24]. The objective of the challenge is to propose methods which can predict the mental state of individuals based on continuous PHQ scores. For further details of the dataset, the reader is referred to [30].

3 CHALLENGES

Automated detection of depression is an active area of research, which faces a number of challenges.

The primary challenge is the lack of publically available datasets, which has traditionally restricted research in this field. AVEC challenges [30, 35, 36] on depression recognition have recently brought two sets of datasets in the form of AVEC 2013/2014 and AVEC 2016/2017 to the public sphere, and one can find a number of publications based on these datasets. Nevertheless, we posit that there are enough samples in these datasets to holistically represent a complex cognitive impairment such as depression.

The second challenge arises from the field of psychology itself. In most cases, labels for depression severity scores are based on self assessment forms, which essentially rely on individuals to honestly report on the questionnaires. This may not always be true. In fact, for the AVEC 2017 dataset, we note that certain participants have a PHQ score of zero – which may show that they have excellent mental health – but in the transcripts these participants go on to discuss their battles with depression and post traumatic stress disorder in the past. Indeed, Gorwood *et al* [16] argue that past episodes of depression may still cause psychomotor retardation (see also references in [16]). In fact, the duration and frequency of depressive episodes may increase the severity of psychomotor retardation. Therefore, individuals who no longer have depression and report such on the self-report forms may still have impaired social signals which they may not realise themselves.

To hone in on the point of potentially noisy labels in the dataset, consider the case of participant with ID 464, who has a PHQ score of 0. From the interview transcript, one finds this individual saying, "*I know how it's like to be depressed . . . how does depression feel like . . . like a bird in a cage . . . a fish who can't swim in water . . . a bird without wings . . . like you're limited*". When pressed by Ellie (the virtual avatar) the participant finally concedes, "*I could say today, you know, earlier when I was just by myself I felt a little depressed*", even though the PHQ score for this individual is 0.

Nevertheless, we believe that research towards automated screening of depression from the social signal processing (SSP) community can prove to be very useful if methods to screen for depression are interpretable or carry meaningful intuition such that it can be provided as a feedback to research community working in the field of psychology. Any new development from the psychology community can then feedback to the SSP community and so on. We are especially motivated by the works of [5, 6], where the authors emphatically argue in favour of interpretability of features as well as datasets and the machine learning algorithms used in SSP.

4 METHODOLOGY

4.1 Pre-processing for Speech Features

The audio files provided as part of the dataset contain dyadic communication between Ellie and the participants. As a first step we segment portions of the speech recording which consist of speech only from the participants. To this end, we use "start" and "stop" time stamps available as part of the dataset. We note, however, that these time stamps are not accurate and in some cases, there exists alignment errors of up to 4 seconds. Nevertheless, we continue to use the time stamps provided with the transcript under the assumption that there exist only a few such errors. The next step is to combine segments of speech file into a single speech file, following which we use the COVAREP toolbox [8] to compute a set of 73 features which include prosodic, voice quality, and spectral features. Reader is referred to [30] for further details.

4.2 Turbulence in Speech Patterns

Given that psychomotor retardation leads to uniqueness in an individual's speech pattern, therefore it must manifest itself as turbulence or lack-there-of in speech feature trajectories. Fundamentally, we hypothesise the LLDs of speech of individuals with and without depression are different, and if quantified may provide an insight into depression severity. We call these features as "turbulence features".

However, given that the nature of the dataset is such that it includes only non-scripted i.e. free speech, the task of recognising turbulent patterns in speech is complicated. Inspired by the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [13], we devise the following methodology to capture the hypothesised turbulence, and later demonstrate its effectiveness for the task of depression screening.

Consider the pitch of an individual's speech (F0 feature). It has been computed at a frequency of 100 Hz using the COVAREP toolbox. Due to the free speech nature of the interview process there

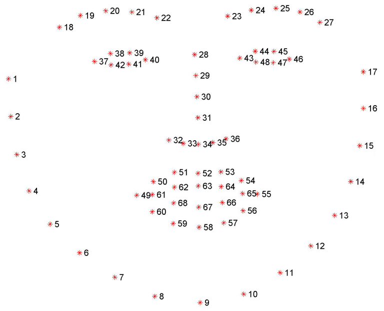

Figure 1: Reference for numbering of 68 point facial landmarks.

exists no prior knowledge where this turbulence may manifest. We therefore consider a multi-scale approach, by using a set of temporal windows of lengths $\{0.5, 2, 5, 10, 15\}$ seconds, with an overlap of $\{0.2, 1, 3, 5, 7\}$ seconds, respectively. Within each window, we compute the crest factor as the measure of turbulence. The crest factor measures the ratio between the max value of the signal and its root mean square (RMS) value. Therefore, if there indeed exist any irregularities in the pitch within any window, we are likely to capture them. Finally, the crest factor values from multiple windows at each scale are pooled using the following descriptive statistics: the 10th, 25th, 50th, and 75th percentile, the mean with 5% trimming, and the range.

In addition to the pitch, we apply the above multi-scale procedure to standard AVEC 2017 – DSC features, which are also one-dimensional speech LLDs. These features include: normalised amplitude quotient (NAQ), quasi open quotient (QOQ), the difference in amplitude of the first two harmonics of the differentiated glottal source spectrum (H1H2), parabolic spectral parameter (PSP), maxima dispersion quotient (MDQ), spectral tilt/slope of wavelet responses (PS), and shape parameter of the Liljencrants-Fantmodel of the glottal pulse dynamics (Rd). Reader is referred to [30] and references therein for details of these features.

4.3 Craniofacial Movement

To quantify psychomotor retardation, our objective is to craft visual features that are capable of representing muscular tightening. Thus we hypothesise that if an individual has depression then their head movements as well as facial muscle movements will be impaired compared to those who do not have depression. In order to track these movements, we use 68-point 3D facial landmarks of the participant available as part the dataset.

Similar to the previous work on this subject [2, 10], we compute velocity and acceleration contours from facial landmarks. However, unlike [2, 10] where contours are computed for many combinations of facial landmarks, we specifically target three types of movements: (1) head movement, (2) mouth movement (both horizontal and vertical) and (3) eyelid movement.

4.3.1 Head movement. For determining head movement we use a set of landmarks in the nose region i.e. the landmarks which are not expected to be affected due to any non-rigid movement of the face. These landmarks include $\{P_{30}, P_{31}, P_{33}, P_{34}, P_{35}\}$, as illustrated in Figure 1.

In order to quantify head movement, we first compute the 3D Euclidean distance between contiguous frames, which encodes the change in (x,y,z) coordinates of facial landmarks. This procedure, for all frames, provides a vector representing velocity of head movement for the individual. Similarly, applying the second order difference operation of the velocity contour provides the acceleration contour. It is important to mention here that we only use landmarks for which the tracking was successful, in order to avoid unintentional contamination of data.

4.3.2 Mouth openings. We also compute velocity and acceleration contours for mouth movement, in particular vertical and horizontal. These essentially measure the deformations of mouth as an individual speaks to Ellie. We hypothesise that the nature of these movements is indicative of depression.

For horizontal movement, we compute pairwise distances for every frame between the points P_{50} and P_{54}, P_{60} and P_{56} and P_{49} and P_{55} i.e. mouth corner regions, representing them as $||P_{50}P_{54}||$, $||P_{60}P_{56}||$, and $||P_{49}P_{44}||$ respectively (see Figure 1). Velocity and acceleration contours are then computed by applying 1st and 2nd order difference operator. Finally, the average value of each contour is taken between three pairs of landmarks as the horizontal mouth movement velocity and acceleration for the individual.

For vertical movement, we first compute pairwise distances $||P_{62}P_{68}||$, $||P_{63}P_{67}||$, and $||P_{64}P_{66}||$, again mouth corner regions, for every frame and follow similar procedure used for horizontal mouth movement to produce vertical mouth movement velocity and acceleration for the individual.

4.3.3 Eyelid movements. We measure eyelid movement as a correlate of blinking rate, which according to [2, 12] can be used to identify individuals who have depression. Similar to the approach already used for other types of movements, we compute velocity and acceleration contours using pairwise distance $||P_{38}P_{42}||$ and $||P_{39}P_{41}||$ for the right eye and $||P_{44}P_{48}||$ and $||P_{45}P_{47}||$ for the left eye.

4.4 Multi-scale Fisher Vector Encoding of Speech Spectra

It has been reported that individuals with cognitive impairments such as autism and schizophrenia have abnormal characteristics to their speech spectra [5, 11, 25]. Similar observations also been reported for individuals with depression, as detailed in [7].

We hypothesise that speech spectra of individuals with depression, represented as spectral LLDs, in the AVEC 2017 dataset is characteristically different from those who do not have depression, and can be quantified. The free speech nature of interview process, however, does not permit a direct comparison of the speech spectra of individuals, therefore, we propose the following framework to test this hypothesis.

We start by creating a background model of the feature space of spectral features using a Gaussian Mixture Model (GMM). Next, for spectral features of every individual in the dataset, we compute both the mean and covariance vectors which represent the deviation of

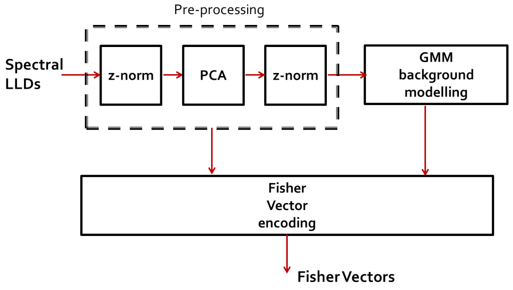

Figure 2: Block diagram for FV encoding of spectral LLDs.

these features from the background model. This process is known as Fisher Vector (FV) encoding.

FV encoding was originally proposed by [27] for use in object recognition tasks, but has recently become popular for a variety of applications in the field of social signal processing. The authors of [9, 19] use FV encoding of visual features for the task of depression recognition, meanwhile, the authors of [20] use it for emotion recognition, and [21, 22] use FV encoding for solutions to Interspeech Computational Paralinguistics (ComParE) challenges.

The overall layout of our framework is depicted in Figure 2: we start by concatenating spectral LLDs from each speech recording of the Training set into a single matrix and then build a background model for the spectral space using a Gaussian Mixture Model (GMM) [29].

However, in order to train the GMM efficiently, we perform the following pre-processing steps: all feature frames are z-score normalised i.e. made to have zero mean and unit standard deviation. Next, we use principal component analysis (PCA) to decorrelate the feature space. We retain dimensions such that they match the number of clusters of the GMM. In case the feature set has dimensions less than or equal to the number of clusters for GMM, we perform PCA over the dimensions of feature set, i.e. dimensionality reduction does not take place. We apply a second z-score normalisation on the output of PCA before using the resultant features to fit the GMM. Finally, we compute Fisher Vectors in order to describe the deviation of each participant's spectra from the background model.

It is also important to mention here that since the GMM is built using pre-processed features, the combination of z-norm + PCA + z-norm needs to be applied to any new features which need to be encoded as Fisher Vectors, for example features of individuals from the Development and the Test set.

In this work, we use `VLFeat` [37], a opensource library, for both, estimating the means, covariance matrices and priors of the GMM, and implementing FV encoding. Initial experimentation had showed that Perronnin's improved FVs [28], which perform normalisation on the Fisher Vectors actually produce worse results in terms of RMSE compared to vanilla FV encoding, therefore, we proceeded to using vanilla FV encoding. It is probable that the non-linear normalisation squashes the dynamic range of Fisher Vectors adversely affecting their discriminative ability.

As representations of speech spectra, we use Mel Frequency Cepstral Coefficients (MFCCs), which is a standard representations for spectra in speech processing. The LLDs, along with their velocity and acceleration contours were computed using the `openSmile` toolkit version 2.3.0 [14]. Similar to the multi-scale approaches discussed so far, we use multiscale FV encoding of spectral LLDs, over windows of $\{0.5, 5, 10\}$ seconds, with overlap of $\{0.2, 3, 5\}$ seconds, respectively. The FVs computed over each of these time scales are pooled into a single FV by applying the following descriptive statistics element-wise: mean, max, median, variance, crest factor (CF) and range.

4.5 Multivariate Regression for Prediction of PHQ Scores

For regression towards the PHQ scores, we use two different methods, on the basis of their qualities. The first is partial least squares regression, which we use when we expect our features to have collinearity and the second is support vector regression, which we use when we expect that our features will require projection into a hyperspace in order to achieve separability.

4.5.1 Partial Least Squares Regression. We use partial least squares regression (PLSR) [31] to develop models to predict PHQ scores of participants. The motivation for doing so is based on the understanding that when features (especially functionals) computed at multiple time-scales are concatenated, it is likely that the resultant feature vector is highly correlated or even collinear. PLSR is especially suited for such cases, as it creates new set of features aka "components" which are linear combinations of the original features. These components are created while considering their effect on the output PHQ scores. The PLSR only has one tuning parameter i.e. the number of components, which we optimise based on the RMSE achieved on the Development set.

4.5.2 Support Vector Regression. We use support vector regressor (SVR), to build models based on Fisher vectors which can predict PHQ scores for individuals. We train SVRs with a radial basis function (RBF) kernel. We utilise Matlab wrappers for ε-SVR available as part of the libSVM [4] library. The cost parameter C is searched between $2^{\{-10:10\}}$, *Epsilon* between $2^{\{-5:5\}}$ and the width of the RBF kernel *Gamma* between $2^{\{-16:4\}}$, with a step of 2. Amongst these, we select the parameters which yield the largest absolute Pearson correlation values on the Development set.

5 EXPERIMENTATION AND ANALYSIS

In this section we discuss experimentation on three different sets of features i.e. the so-called turbulence features, facial movement features, and finally Fisher vector features.

5.1 Turbulence Features

We compute turbulence features over five different time-scales, and use five descriptive statistics to summarise their values. Therefore, the resultant feature has 25 dimensions. In order to build a model for prediction of depression severity, we use PLSR. We vary the number of components between 4 and 8, and optimise for the objective of achieving the smallest RMSE on the Development set. There is, however, an interesting question which exists with the use of these features i.e. should unvoiced frames be removed from the features or should they be retained. On one hand it makes sense to remove them because they will contain information not related

Table 1: Performance of "turbulence features" while keeping unvoiced frames as "0".

Feat	Train			Dev			Comp
	MAE	*RMSE*	*Corr*	*MAE*	*RMSE*	*Corr*	
F0	4.23	5.20	0.29	4.81	5.95	0.43	4
NAQ	4.14	5.06	0.37	5.03	6.16	0.37	8
QOQ	4.24	5.23	0.27	6.2	8.29	0.04	4
H1H2	4.11	5.16	0.32	5.27	6.47	0.13	8
PSP	4.39	5.25	0.26	6.52	10.40	0.04	4
MDQ	4.04	5.05	0.37	5.21	6.43	0.21	4
PS	4.23	5.12	0.34	5.19	6.33	0.30	5
Rd	4.41	5.34	0.19	5.33	6.43	0.20	7

Table 2: Performance of "turbulence features" after removing unvoiced frames.

Feat	Train			Dev			Comp
	MAE	*RMSE*	*Corr*	*MAE*	*RMSE*	*Corr*	
F0	4.40	5.21	0.29	5.58	6.75	0.01	4
NAQ	4.31	5.34	0.19	5.33	6.40	0.27	7
QOQ	3.74	4.79	0.47	5.66	6.73	0.02	7
H1H2	4.07	5.14	0.33	5.38	6.54	0.12	5
PSP	4.00	4.88	0.44	6.22	8.22	0.10	7
MDQ	4.22	5.17	0.31	5.45	6.49	0.20	7
PS	3.93	4.79	0.48	5.55	6.95	0.10	4
Rd	4.06	4.93	0.42	5.49	6.64	0.03	4

to speech, but on the other hand, keeping unvoiced frames may provide information about the rhythm of an individual's speech.

We empirically test two approaches, with results summarised in Tables 1 and 2. In Table 1, instead of removing unvoiced frames, we simply change their value to zero, meanwhile in Table 2, we remove those frames altogether. An inspection of these two tables suggests that it is almost always beneficial to retain unvoiced frames if their value is changed to zero. The biggest beneficiary are features derived from F0, which comfortably beats the best RMSE and MAE scores for the Development set, as given in the baseline paper.

5.2 Craniofacial Movement Features

Table 3 shows Pearson's correlation values for best performing visual features for each category defined in Section 4.3. While the table shows top results from each category only, we report that mouth movements had the highest correlation values at multiple temporal resolutions, which is intuitive given that we directly measure mouth activity and the dataset has been designed such that participants speak to the virtual agent. Surprisingly, head movements did not yield good results. It is likely due to the fact that participants were specifically asked to look into the camera, and were consciously controlling their head movement.

5.3 Fisher Vector Features

In Table 4, we provide a summary of the best performing pooling methods on the Training and Development sets in terms of RMSE and MAE as well as the absolute Pearson correlation coefficients, whilst selecting a cut-off p-value of 0.05. We note that all pooling methods are able to achieve virtually similar performance in terms of MAE and RMSE, when one has options to choose any particular temporal resolution for FV encoding along with the number of clusters for the GMM.

Table 3: Pearson correlation with PHQ values for various visual features.

Feature(s)	Corr	p-value
Mouth movement Vel. (vert.)	-0.323	< 0.05
Mouth movement Acc. (horiz)	0.317	< 0.05
Eyelid movement Acc.	0.297	< 0.05
Head movement Acc.	-0.246	< 0.05

Table 4: Summary of results for various pooling methods for multi-scale FV encoding.

Pool .	Train			Dev			Res.	GMM
	MAE	*RMSE*	*Corr*	*MAE*	*RMSE*	*Corr*		
Mean	3.05	3.38	0.90	5.53	6.50	0.43	3	16
Max	4.81	5.66	0.60	5.67	6.52	0.33	3	16
Med.	4.81	5.66	0.54	5.65	6.52	0.32	2	24
CF	4.81	5.66	0.71	5.66	6.52	0.34	1	32
Range	4.78	5.47	0.59	5.60	6.42	0.37	3	16
Var	4.81	5.66	0.70	5.65	6.52	0.41	2	16

There are however subtle differences. Firstly, we note that the absolute Pearson correlation value of 0.43 on the Development set is achieved through Mean pooling of FVs, when using *Res*3 i.e. a window of 10 seconds and using a 24 cluster GMM. Mean pooling also has the smallest MAE and RMSE on the Development set compared to other pooling methods. Another important observation is that most pooling methods perform well at *Res*3 i.e. a temporal resolution of 10 seconds, while the crest factor (CF) stands out as the only pooling method which works best at a temporal resolution of 500 ms. We believe that this is due to the nature of the crest factor, which essentially measures turbulence, and at larger time-scale, micro-level description is not as fruitful for the task of predicting labels.

While the objective of the AVEC 2017 – DSC is to achieve the smallest possible RMSE, we believe that there may be cases where one may want to use measure depression severity using a parameter which may not match PHQ scores in terms of its dynamic range (therefore have poor RMSE), but closely matches the PHQ labels through correlation. For example, consider Table 5, where we summarise possible trade-offs between the choice of smaller RMSE or a higher absolute Pearson correlation.

5.4 Results on the Test Set and Work in Progress

We train an SVR using features from mean pooling of Fisher Vectors (see section 5.3), which provided us an RMSE of 6.50 and MAE of 5.50 on the Development set, as well as the highest correlation

Table 5: Trade-off between minimising RMSE and maximising correlation.

Pooling	Train			Dev		
	MAE	*RMSE*	*Corr*	*MAE*	*RMSE*	*Corr*
Max (best RMSE)	4.81	5.66	0.60	5.67	6.52	0.33
Max (best corr)	4.45	5.62	0.89	5.45	6.93	0.42
Mean (best RMSE)	3.05	3.38	0.90	5.53	6.50	0.43
Mean (best corr)	3.67	3.87	0.92	5.50	6.55	0.46

value of 0.43. On the Test set, this model achieves an RMSE of **6.42** with an MAE of **5.42**. These results are better than best results for the Test set as provided in the baseline paper i.e. RMSE of 6.97 and MAE of 5.66.

However, we achieve slightly better results i.e. RMSE equal to **6.34** and MAE equal to **5.30** on the Test set when we use turbulence features computed for F0 followed by PLSR (see section 5.1). On the Development set too, this model performed better than the SVR model.

While these results are interesting, we believe futher experimentation is important, not only using datasets of larger size but also with individuals from different cultural backgrounds. Sharifa *et. al.* [1] discussed automated depression recognition using three different datasets with individuals from diverse cultural backgrounds (American, Australian and German). The authors report poor transferability of features from one dataset to another, even though on the same dataset the performance is quite acceptable. Our exploration on the suitability of the proposed approaches on the AVEC 2014 dataset [34], a dataset with participants from German background, is currently a work in progress.

6 CONCLUSION

In summary, we proposed a novel set of temporal features (which we called "turbulence features") and showed their effectiveness for depression screening. We detailed a novel methodology to target specific craniofacial movements which are indicative of psychomotor retardation and hence of depression. Further, we presented a novel method for quantifying abnormalities of speech spectra of individuals with depression using Fisher vector encoding of spectral low level descriptors (LLDs).

REFERENCES

[1] Sharifa Alghowinem, Roland Goecke, Julien Epps, Michael Wagner, and Jeffrey Cohn. 2016. Cross-Cultural Depression Recognition from Vocal Biomarkers. In *INTERSPEECH 2016*.

[2] S Alghowinem, R Goecke, M Wagner, J Epps, M Hyett, G Parker, and M Breakspear. 2016. Multimodal Depression Detection: Fusion Analysis of Paralinguistic, Head Pose and Eye Gaze Behaviors. *IEEE Trans. Affect. Comput.* 99 (2016), 1–14.

[3] American Psychiatric Association. 2013. *DSM-5*. 4–5 pages.

[4] Chih-Wei Hsu, Chih-Chung Chang and Chih-Jen Lin. 2008. A Practical Guide to Support Vector Classification. *BJU Int.* 101, 1 (2008), 1396–400.

[5] Alex S Cohen, Jessica E McGovern, Thomas J Dinzeo, and Michael A Covington. 2014. Speech Deficits in Serious mental Illness: A Cognitive Resource Issue? *Schizophr. Res.* 160, 0 (dec 2014), 173–179.

[6] Alex S Cohen, Tyler L Renshaw, Kyle R Mitchell, and Yunjung Kim. 2016. A psychometric investigation of "macroscopic" speech measures for clinical and psychological science. *Behav. Res. Methods* 48, 2 (2016), 475–486.

[7] Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F. Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* 71 (2015), 10–49.

[8] G Degottex, J Kane, T Drugman, T Raitio, and S Scherer. 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In *Acoust. Speech Signal Process.* 960–964.

[9] A Dhall and R Goecke. 2015. A temporally piece-wise fisher vector approach for depression analysis. In *ACII.* 255–259.

[10] Hamdi Dibeklioglu, Zakia Hammal, Ying Yang, and Jeffrey F Cohn. 2015. Multimodal Detection of Depression in Clinical Interviews. In *IMCI 2015.* 307–310.

[11] Joshua John Diehl and Rhea Paul. 2011. Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders. *Appl. Psycholinguist.* 34 (2011), 1–27.

[12] Dieter Ebert, Roland Albert, Gerhard Hammon, Bernhard Strasser, Albrecht May, and Antje Merz. 1996. Eye-blink rates and depression. Is the antidepressant effect of sleep deprivation mediated by the dopamine system? *Neuropsychopharmacology* 15, 4 (1996), 332–339.

[13] F Eyben, K R Scherer, B W Schuller, J Sundberg, E André, C Busso, L Y Devillers, J Epps, P Laukka, S S Narayanan, and K P Truong. 2016. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* 7, 2 (2016), 190–202.

[14] Florian Eyben, Felix Weninger, Florian Gross, and Bjorn Schuller. [n. d.]. Recent developments in openSMILE, the munich open-source multimedia feature extractor. In *ACM MM 2013.* 835–838.

[15] Jeffrey M Girard and Jeffrey F Cohn. 2014. Automated Audiovisual Depression Analysis. *Curr. Opin. Psychol.* 4 (2014).

[16] P Gorwood, S Richard-Devantoy, F Baylé, and M L Cléry-Melun. 2014. Psychomotor retardation is a scar of past depressive episodes, revealed by simple cognitive tests. *Eur. Neuropsychopharmacol.* 24, 10 (oct 2014), 1630–1640.

[17] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Albert "Skip" Rizzo, and Louis-Philippe Morency. 2014. The Distress Analysis Interview Corpus of Human and Computer Interviews. In *Int. Conf. Lang. Resour. Eval.* 3123–3128.

[18] Juha Hämäläinen, Kari Poikolainen, Erkki Isometsä, Jaakko Kaprio, Martti Heikkinen, Sari Lindeman, and Hillevi Aro. 2005. Major depressive episode related to long unemployment and frequent alcohol intoxication. *Nord. J. Psychiatry* 59, 6 (2005), 486–91.

[19] Varun Jain, James L Crowley, Anind K Dey, and Augustin Lux. 2014. Depression Estimation Using Audiovisual Features and Fisher Vector Encoding. In *AVEC 2014 (AVEC '14).* ACM, New York, NY, USA, 87–91.

[20] Heysem Kaya, Furkan Gürpinar, Sadaf Afshar, and Albert Ali Salah. 2015. Contrasting and Combining Least Squares Based Learners for Emotion Recognition in the Wild. In *ACMI (ICMI '15).* ACM, New York, NY, USA, 459–466.

[21] Heysem Kaya and Alexey A. Karpov. 2016. Fusing Acoustic Feature Representations for Computational Paralinguistics Tasks. In *INTERSPEECH 2016.* 2046–2050.

[22] Heysem Kaya, Alexey A. Karpov, and Albert Ali Salah. 2015. Fisher Vectors with Cascaded Normalization for Paralinguistic Analysis. In *INTERSPEECH 2015.* 909–913.

[23] Lars Vedel Kessing. 2012. Depression and the risk for dementia. *Curr. Opin. Psychiatry* 25, 6 (2012), 457–461.

[24] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet B W Williams, Joyce T Berry, and Ali H Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disord.* 114, 1âĂŞ3 (2009), 163–173.

[25] S.H.R.E. Motlagh, H Moradi, and H Pouretemad. 2013. Using general sound descriptors for early autism detection. In *Control Conf. (ASCC), 2013.* 1–5.

[26] Paula M Niedenthal, Lawrence W Barsalou, Piotr Winkielman, Silvia Krauth-Gruber, and François Ric. 2005. Embodiment in Attitudes, Social Perception, and Emotion. *Personal. Soc. Psychol. Rev.* 9, 3 (2005), 184–211.

[27] Florent Perronnin and Christopher Dance. 2007. Fisher kernels on visual vocabularies for image categorization. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*

[28] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the Fisher kernel for large-scale image classification. In *Lect. Notes Comput. Sci.*, Vol. 6314. 143–156.

[29] Douglas A. Reynolds and Richard C. Rose. 1995. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. *IEEE Trans. Speech Audio Process.* 3, 1 (1995), 72–83.

[30] Fabien Ringeval, Bjorn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic. 2017. AVEC 2017 âĂŞ Real-life Depression, and Affect Recognition Workshop and Challenge. In *AVEC.*

[31] Roman Rosipal and Nicole Kr. 2006. Overview and Recent Advances in Partial Least Squares. *Subspace, Latent Struct. Featur. Sel. Saunders, C., al. (heidelb. Springer-Verlag, 2006)* 3940 (2006), 34–51.

[32] Didier Schrijvers, Wouter Hulstijn, and Bernard G C Sabbe. 2008. Psychomotor symptoms in depression: A diagnostic, pathophysiological and therapeutic tool. (2008), 20 pages.

[33] Christina Sobin and Harold A. Sackeim. 1997. Psychomotor symptoms of depression. *Am. J. Psychiatry* 154, 1 (1997), 4–17.

[34] Michel Valstar. 2014. Automatic Behaviour Understanding in Medicine. In *Proc. 2014 Work. Roadmapping Futur. Multimodal Interact. Res. Incl. Bus. Oppor. Challenges (RFMIR '14)*. ACM, New York, NY, USA, 57–60.

[35] Michel Valstar, Jonathan Gratch, Bjorn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Guiota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge. In *Int. Work. Audio/Visual Emot. Chall.*

[36] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic. 2013. AVEC 2013 - the continuous audio/visual emotion and depression recognition challange. In *Proc. 3rd Int. Audio/Visual Emot. Chall. Work.*

[37] A. Vedaldi and B. Fulkerson. 2008. {VLFeat}: An Open and Portable Library of Computer Vision Algorithms. (2008).

[38] World Health Organisation. 2017. Depression fact sheet. (2017). http://www.who.int/mediacentre/factsheets/fs369/en/

[39] Katherine S Young, Christine E Parsons, Alan Stein, and Morten L Kringelbach. 2015. Motion and emotion: depression reduces psychomotor performance and alters affective movements in caregiving interactions. *Front. Behav. Neurosci.* 9, February (2015), 26.