LONGITUDINAL MODELING OF DEPRESSION SHIFTS USING SPEECH AND LANGUAGE

Paula Andrea Pérez-Toro[1,2,3★] *Judith Dineley*[3] *Agnieszka Kaczkowska*[3] *Pauline Conde*[3]
Yuezhou Zhang[3] *Faith Matcham*[3,4] *Sara Siddi*[5] *Josep Maria Haro*[5] *Stuart Bruce*[6]
Til Wykes[3,7] *Raquel Bailón*[8,9] *Srinivasan Vairavan*[10] *Richard J.B. Dobson*[3,11]
Andreas Maier[1] *Elmar Nöth*[1] *Juan Rafael Orozco-Arroyave*[1,2] *Vaibhav A. Narayan*[12]
Matthew Hotopf[3,7] *Nicholas Cummins*[3★] *The RADAR-CNS Consortium*[13]

[1]Pattern Recognition Lab. Friedrich-Alexander Universität, Erlangen-Nürnberg, Erlangen, Germany; [2]GITA Lab, Facultad de Ingeniería. Universidad de Antioquia, Medellín, Colombia; [3]Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK; [4]School of Psychology, University of Sussex, Falmer, UK; [5]Parc Sanitari Sant Joan de Déu, Fundació Sant Joan de Déu, CIBERSAM, Barcelona, Spain; [6]RADAR-CNS Patient Advisory Board, King's College London, UK
[7]NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK
[8]Biomedical Signal Interpretation and Computational Simulation (BSICoS) group, Aragon Institute for Engineering Research, University of Zaragoza, Zaragoza, Spain; [9]Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Spain; [10]Janssen Research and Development LLC, Titusville, NJ, United States; [11]Institute of Health Informatics, University College London, London, UK; [12]Davos Alzheimer's Collaborative; [13]www.radar-cns.org
[★]corresponding authors: `paula.andrea.perez@fau.de`, `nick.cummins@kcl.ac.uk`

ABSTRACT

Speech analysis can provide a potential non-invasive and objective means of assessing and monitoring an individual's mental health. Most studies to date have focused on cross-sectional analysis and have not explored the benefits of speech analysis as a longitudinal monitoring tool that can assist in the management of chronic conditions such as major depressive disorder (MDD). Objectively monitoring for shifts in depression symptom severity levels over time presents a notable challenge, which we address through an automated approach using longitudinal English and Spanish speech samples collected from a clinical population. We employ time–frequency representations and linguistic embeddings to enhance the early recognition of alterations in depression levels in individuals with MDD. We investigate the suitability of using siamese-based training for modeling these changes, intending to enable personalized and adaptive interventions.

Index Terms— Depression, Speech Analysis, Language Analysis, Longitudinal Assessment, Contrastive Training

1. INTRODUCTION

Major Depressive Disorder (MDD) is among the world's most common mental health issues. According to an Organisation for Economic Co-operation and Development (OECD) report released in 2018, approximately 21 million people (4.5% prevalence) were living with a depressive disorder across European Region (EU) countries in 2016 [1]. MDD is characterized by persistent sadness, loss of interest, and disruptions in sleep and appetite [2]. Monitoring changes in MDD symptom severity is crucial for recognizing unique triggers and early signs of potential relapse. However, the lack of objective assessment tools means that processes are typically based on a clinician's judgment of symptoms retrospectively reported by the patient, introducing bias on both sides [3].

Clinicians often observe someone's voices and speech patterns as criteria to assess their symptom severity; however, these observations are subjective. Speech affected by depression is often described clinically as having reduced verbal activity, shorter utterances, slower speech rate, and increased pauses [4, 5]. The predictive power of individual speech features has also been linked to changes in depression symptom severity; e. g. [6, 7, 8, 9]. Most recent speech and depression research is focused on developing machine learning models that use multivariate feature spaces to detect the presence or absence of depression in speech; e. g. [10, 11, 12, 13, 14]. Such works highlight the promise of using speech to monitor changes in MDD symptom severity.

Herein, we investigate the suitability of speech and language analyses for classifying shifts in depression symptom severity level. We utilize longitudinal speech recordings of 187 English speakers and 53 Spanish speakers, all with a clinical history of recurrent MDD. First, we select a pair of recordings from the same speaker, each associated with different depression scores; one indicating a lower level of depression and the other, a higher level. We then label the direction of the depression score shift, which can either be from low to high or from high to low.

We propose an approach for classifying depression level shifts that use contrastive learning and Siamese Neural Networks (SNNs). SNNs learn and compare representations of input pairs [15], making them ideal for detecting patterns in speech and language associated with depression level changes. Therefore, unlike conventional classification methods, this approach pairs of recordings to learn the the direction of the depression score shift. Our approach has the potential to enhance our understanding of depression dynamics and aid in early intervention by automatically detecting changes in depression severity using speech and language cues.

2. THE RADAR-MDD SPEECH CORPUS

Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD; [16]) is a longitudinal cohort study examining the utility of multi-parametric remote measurement technologies (RMT), including speech, to measure changes in symptoms and predict relapse in people with MDD. The full eligibility and exclusion criteria for RADAR-MDD are published fully in [16]. RADAR-MDD had three recruitment sites: London, United Kingdom; Amsterdam, The Netherlands; and Barcelona, Spain. In this work, we use data from the UK and Spain sites; ethical approval for these sites was obtained from the Camberwell St. Giles Research Ethics Committee (17/LO/1154) in London, from the Fundació Sant Joan de Deu Clinical Research Ethics Committee (CI: PIC-128-17) in Barcelona.

2.1. Patient Involvement

The experimental protocol was co-developed with a patient advisory board who shared their opinions on several user-facing aspects of the study, including the choice and frequency of survey measures, the usability of the study app, participant-facing documents, selection of optimal participation incentives, selection, and deployment of wearable device as well as the data analysis plan. The speech task and subsequent analysis have been discussed specifically with a Patient Advisory Board.

2.2. Speech collection

For full details on the preparation and organization of the speech data, the interested reader is referred to [7, 10]. In this work, we used the *free-response* data only to use language features alongside acoustic information. The total number of participants and information on the distribution of the audio files used in our analysis are presented in Table 1.

2.3. Data Availability

Due to the confidential nature of speech data, we are unable to make our data publicly available. Access to the data can be made through reasonable requests to the RADAR-CNS consortium and will be subject to local ethics clearances. Please email the senior author for details.

Table 1: Demographic and clinical information of the subjects for each dataset

	Spanish Dataset	English Dataset
	F/M	F/M
Gender	37 / 16	146 / 41
Age	52.8 (10.3) / 55.4 (12.1)	45.4 (15.7) / 49.1 (14.8)
Education	12.0 (4.2) / 13.62 (4.3)	17.4 (5.8) / 16.4 (3.5)
PHQ-8	13.1 (6.0) / 10.6 (6.3)	10.4 (6.2) / 10.9 (6.1)
PHQ-8 range	0–24 / 0–24	0–24 / 0–24
# of recordings per speaker	1.9 (1.8) / 1.9 (1.2)	2.8 (2.2) / 2.7 (2.2)

Values are expressed as mean (standard deviation). F: female. M: male. Age and education are given in years.

3. METHODS

3.1. Siamese Based Neural Network

Instead of learning to classify its inputs, the SNN learns how to differentiate between two inputs. It comprises two identical sub-networks that share weights among them [15]. Commonly, when training the network to differentiate between similar and dissimilar instances, we often provide one positive and one negative example at a time. Based on this methodology, we define a pair of recordings from the same speaker with two depression scores, one with a lower depression score and another with a higher one. The pair of recordings is selected so that the depression score differs by five points according to the eight-item Patient Health Questionnaire (PHQ-8) [17] scores; noting we choose five points as this is the range of the PHQ-8 severity intervals (0–4, 5–9, 10–14, 15–19, and 20–24) [17]. Then, we label the direction in which the depression score shifted. It could be from low to high or from high to low.

The first part of the network consists of creating a latent space or embedding based on the modality architecture (Figure 1; note for illustration purposes, the only architecture for the speech modeling part is shown.) This embedding will guide the network according to the direction label. For this, we use the Cosine Embedding loss (CS). These embeddings are then subtracted to get one embedding for the classification layer. Finally, the labels are passed through a Cross Entropy Loss (CE) and both CS and CE are then summed, similar to what is done in zero-shot learning [18]. Note that our objective is not primarily focused on identifying dissimilarities; instead, it revolves around detecting shifts in depression. This is why we opted not to employ the regular contrastive loss for the final outputs. Additionally, two different network architectures were considered for each modality.

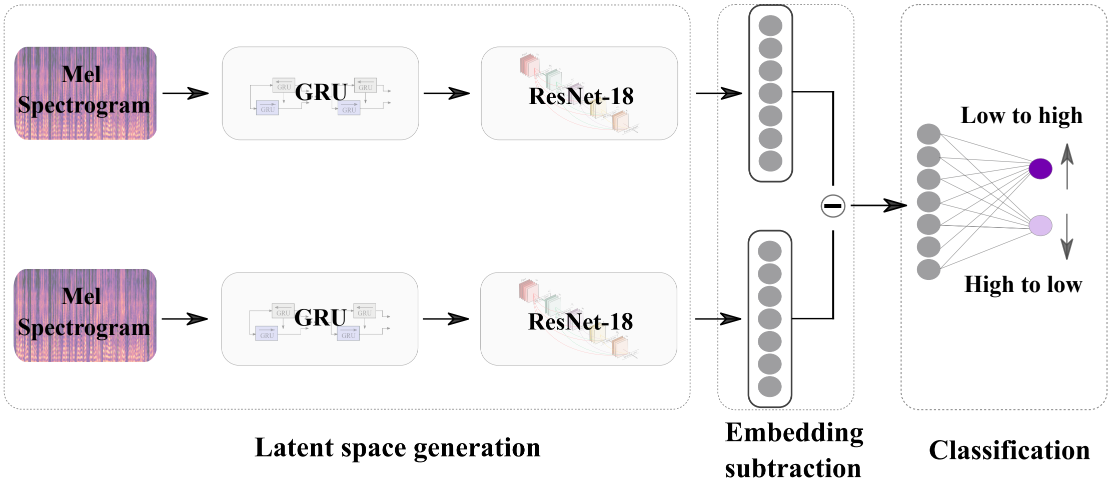

Fig. 1: Siamese architecture network applied in this study. The upper and lower networks correspond to a pair of recordings from the same speaker. (1) Latent Space Generation: Initially, the network extracts relevant information from the input features to create a latent representation. (2) Embedding Subtraction: Next, the embeddings obtained from the pair are subtracted from each other, resulting in a single combined representation. (3) Classification: Finally, a linear layer consisting of two neurons is employed to discriminate into either an increase or decrease in depression scores.

3.1.1. Speech Modeling

The overall architecture in Figure 1 aims to model different aspects related to articulation and prosody information by combining log-Mel spectrograms, Gated Recurrent Units (GRU), and Convolutional Neural Networks (CNN)[1]. We considered the use of Mel spectrograms since this time-frequency representation can help identify specific speech characteristics or patterns associated with depression, such as changes in pitch, speech rate, or the presence of specific acoustic markers such as pauses or emotional prosody [19]. To generate the log-Mel spectrogram, we used a window of 45 ms and a hop size of 10 ms. Sequences of 5 s (500 frames) were taken, and the number of Mel filters was set to 128. The spectrograms were normalized according to the training set using a z-score. A unit normalization layer is applied to the input, followed by a GRU composed of 128 hidden states and used to model the temporal dynamics. For the CNN part, a pre-trained ResNet–18 on *ImageNet-1K* [20] was fine-tuned and then modified to receive 1-channel input and to output 512 units passed that are then passed through a GELU activation. Finally, after subtracting the two embeddings, a classification layer is applied.

3.1.2. Language Modeling

In the text processing phase, we automatically generated transcriptions using *Whisper*, an open-source Automatic Speech Recognizer (ASR) [21]. We used the large model version for both languages. Furthermore, we consider the use of word embeddings, specifically a well-known pretrained model called Robustly Optimized Bidirectional Encoder Representations from Transformers–*BERT* Pretraining Approach (RoBERTa) [22] RoBERTa uses the same concept as BERT, but omits the Next Sentence Prediction (NSP) component and employs larger batch sizes. It incorporates multiple attention mechanisms called “heads” which operate concurrently. Furthermore, this approach enables the model to capture a broader range of word relationships through multi-head attention. RoBERTa follows a transfer learning paradigm, where it begins by pretraining on an unsupervised task known as Masked Language Modeling (MLM). In this task, the model learns to predict missing (masked) words within sentences. It also introduces dynamic masking for MLM, where masked tokens change during training epochs. We use RoBERTa–base pretrained models on English and a multi-lingual corpus for Spanish[2], respectively. The average pooling from the last layer (768 units) is taken as the embedding. A similar procedure is followed here. However, we directly subtract the pair embeddings to be passed to a classification layer.

4. EXPERIMENTS AND RESULTS

Following the proposed strategy in Section 3.1, we randomly composed pairs of recordings (from the same speaker) to simulate a range of shifts in depression severity levels, we specifically selecting pairs that exhibited a minimum 5-point difference in PHQ-8 scores.vIn the case of the Spanish corpus, we compiled 100 pairs, with 54 pairs reflecting an increase in PHQ-8 scores and 46 pairs indicating a decrease. In the English corpus, we obtained 524 pairs, with 260 pairs represent-

[1] he source code is available online `https://github.com/PauPerezT/DepShifts_SNNs`

[2] English: `https://huggingface.co/roberta-base`, Multilingual: `https://huggingface.co/xlm-roberta-base`

Table 2: Classification results for each modality and language considering speaker dependent

Modality	Language	AUC	Recall ↓	Recall ↑
Speech	EN	0.66 (0.02)	0.60 (0.03)	0.61 (0.06)
	ES	0.77 (0.05)	0.66 (0.11)	0.75 (0.07)
Language	EN	0.69 (0.03)	0.61 (0.04)	0.63 (0.03)
	ES	0.66 (0.06)	0.51 (0.32)	0.58 (0.31)

EN: English. ES: Spanish. ↓: high to low PHQ-8. ↑: low to high PHQ-8. Values are expressed as mean (standard deviation)

Table 3: Classification results for each modality and language considering speaker independent

Modality	Language	AUC	Recall ↓	Recall ↑
Speech	EN	0.65 (0.04)	0.51 (0.04)	0.72 (0.07)
	ES	0.70 (0.04)	0.60 (0.07)	0.64 (0.03)
Language	EN	0.71 (0.02)	0.58 (0.05)	0.74 (0.06)
	ES	0.69 (0.04)	0.39 (0.31)	0.71 (0.25)

EN: English. ES: Spanish. ↓: high to low PHQ-8. ↑: low to high PHQ-8. Values are expressed as mean (standard deviation)

ing a shift from low to high scores and 263 pairs indicating the reverse, from high to low. Two experiments were performed based on speaker-dependent and speaker-independent stratification strategies. The models were trained following a 4-fold cross-validation, where the reported results are based on the average of the folds. The performance of the classifiers was measured in terms of Recall and Area under the ROC Curve (AUC). An additional baseline experiment using eGeMAPs with a support vector machine was considered but later excluded due to unsatisfactory results.

For the speaker-dependent set-up (Table 2), the highest performance for the speech modality was achieved in the Spanish corpus, indicating moderate to good discrimination ability (AUC = 0.77). The English data tends to yield slightly higher AUC values for the language modality (AUC = 0.69). We observed that in most cases, the classification of an increase in the depression score was more accurately predicted.

The speaker-independent results exhibit a similar trend (see Table 3). The highest performance for the speech modality was in the Spanish data (AUC = 0.70), whilst for the linguistic modality it was in the English data (AUC = 0.71). Regarding the speech modality, the performance decreased by approximately 7% compared to the speaker-dependent experiment. However, the AUCs in the speaker-independent scenario are slightly higher for the language modality than those in the speaker-dependent scenario.

5. DISCUSSION AND CONCLUSIONS

This study explores the feasibility of automatically leveraging speech and language data to identify shifts in depression severity levels. The proposed methodology tested the capability of SNNs with two longitudinal corpora (English and Spanish). In addition, we considered both speaker-dependent and independent scenarios as speaker independence could help model the phenomenon itself (depression) and, in contrast, speaker-dependent analyses could allow the modeling of disease progression. The results showed that while there are differences in specific performance metrics between the speaker-dependent and speaker-independent scenarios, the general trends and patterns remain consistent. Speaker-dependent training tends to yield slightly better discrimination performance. The model depression-related shifts from low to high PHQ-8 scores according to the obtained recall values in different direction shifts. We will explore potential reasons for this in future work.

Regarding individual languages, we observed that English exhibits more consistent results in terms of standard deviation values. This may be attributed to the larger training dataset available for the proposed model, including pre-trained word embeddings. A limitation of this study is that the PHQ-8 relies on self-reported symptoms, potentially introducing social desirability bias, misinterpretation, and susceptibility to cultural or linguistic influences, leading to potential inaccuracies and biases. For future work, we will explore personalized training, where each speaker has a dedicated network. This personalized approach is relevant because depression shifts can be highly influenced by the patient's personality and self-awareness, and willingness to disclose symptoms. By considering this factor, we can gain deeper insights into the dynamics of depression over time. Personalized speech models demonstrate a higher resilience to confounders in the signal, ensuring more accurate and consistent performance customized to individual users' speech patterns [23].

6. ACKNOWLEDGEMENTS

The RADAR-CNS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This Joint Undertaking receives support from the EU's Horizon 2020 research and innovation programme and EFPIA (www.imi.europa.eu). This communication reflects the views of the RADAR-CNS consortium, and neither IMI nor the EU and EFPIA are liable for any use that may be made of the information contained herein. The funding body has not been involved in the study's design, data collection or analysis, or data interpretation. We thank our colleagues within the RADAR-CNS consortium and all involved institutions for contributing to developing this protocol. We thank all RADAR-CNS patient advisory board members for contributing to the device selection procedures and invaluable advice throughout the study protocol design. This paper represents independent research funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR, or the Department of Health and Social Care.

7. REFERENCES

[1] OECD and European Union, “Promoting mental health in Europe: Why and how,” OECD iLibrary, 2018.

[2] A. J. Rush et al., “The Inventory of Depressive Symptomatology (IDS): clinician (IDS-C) and self-report (IDS-SR) ratings of depressive symptoms,” *International journal of methods in psychiatric research*, vol. 9, no. 2, pp. 45–59, 2000.

[3] J. E. Wells et al., “How accurate is recall of key symptoms of depression? A comparison of recall and longitudinal reports,” *Psychological medicine*, vol. 34, no. 6, pp. 1001–1011, 2004.

[4] N. Cummins et al., “A review of depression and suicide risk assessment using speech analysis,” *Speech communication*, vol. 71, pp. 10–49, 2015.

[5] D. M. Low et al., “Automated assessment of psychiatric disorders using speech: A systematic review,” *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[6] A. Abbas et al., “Remote digital measurement of facial and vocal markers of major depressive disorder severity and treatment response: a pilot study,” *Frontiers in digital health*, vol. 3, pp. 610006, 2021.

[7] N. Cummins et al., “Multilingual markers of depression in remotely collected speech samples: A preliminary analysis,” *Journal of Affective Disorders*, vol. 341, pp. 128–136, 2023.

[8] J. C. Mundt et al., “Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology,” *Journal of neurolinguistics*, vol. 20, no. 1, pp. 50–64, 2007.

[9] Y. Yang et al., “Detecting depression severity from vocal prosody,” *IEEE transactions on affective computing*, vol. 4, no. 2, pp. 142–150, 2012.

[10] E. L. Campbell et al., “Classifying depression symptom severity: Assessment of speech representations in personalized and generalized machine learning models.,” in *Proceedings INTERSPEECH 2023*, pp. 1738–1742.

[11] E. Rejaibi et al., “MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech,” *Biomedical Signal Processing and Control*, vol. 71, pp. 103107, 2022.

[12] L. Yang et al., “Feature augmenting networks for improving depression severity estimation from speech signals,” *IEEE Access*, vol. 8, pp. 24033–24045, 2020.

[13] Z. Zhao et al., “Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 423–434, 2019.

[14] P. A. Pérez-Toro et al., “Depression assessment in people with Parkinson’s disease: The combination of acoustic features and natural language processing,” *Speech Communication*, vol. 145, pp. 10–20, 2022.

[15] G. Koch et al., “Siamese neural networks for one-shot image recognition,” in *ICML deep learning workshop*. Lille, 2015, vol. 2.

[16] F. Matcham et al., “Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol,” *BMC psychiatry*, vol. 19, no. 1, pp. 1–11, 2019.

[17] K. Kroenke et al., “The PHQ-8 as a measure of current depression in the general population,” *Journal of affective disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[18] N. Wojke et al., “Deep cosine metric learning for person re-identification,” in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 748–756.

[19] A. M. Badshah et al., “Speech emotion recognition from spectrograms with deep convolutional neural network,” in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[20] K. He et al., “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[21] A. Radfor et al., “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[22] Z. Liu and otherso, “A Robustly Optimized BERT Pre-training Approach with Post-training,” in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 2021, pp. 1218–1227.

[23] J. R. Green et al., “Automatic speech recognition of disordered speech: Personalized models outperforming human listeners on short phrases.,” in *Interspeech*, 2021, pp. 4778–4782.