Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Speech reference intervals: an assessment of feasibility in depression symptom severity prediction

Lauren L. White[1,2], *Ewan Carr*[1], *Judith Dineley*[1], *Catarina Botelho*[3], *Pauline Conde*[1], *Faith Matcham*[4], *Carolin Oetzmann*[1], *Amos A. Folarin*[1,5], *George Fairs*[2], *Agnes Norbury*[2], *Stefano Goria*[2], *Srinivasan Vairavan*[6], *Til Wykes*[1,5], *Richard J.B. Dobson*[1, 7], *Vaibhav A. Narayan*[8], *Matthew Hotopf*[1,5], *Alberto Abad*[3,9], *Isabel Trancoso*[3,9], *Nicholas Cummins*[1,2], *The RADAR-CNS Consortium*[10]

[1]Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK
[2]Thymia, London, UK
[3]INESC-ID, Portugal
[4]School of Psychology, University of Sussex, Falmer, UK
[5]South London and Maudsley NHS Foundation Trust, London, UK
[6]Janssen Research and Development LLC, Titusville, NJ, United States
[7]Institute of Health Informatics, University College London, UK
[8]Davos Alzheimer's Collaborative
[9]Instituto Superior Técnico, University of Lisbon, Portugal
[10]www.radar-cns.org
lauren.white@kcl.ac.uk, nick.cummins@kcl.ac.uk

Abstract

Major Depressive Disorder (MDD) is a prevalent mental disorder. Combining speech features and machine learning has promise for predicting MDD, but interpretability is crucial for clinical applications. Reference intervals (RIs) represent a typical range for a speech feature in a population. RIs could increase interpretability and help clinicians identify deviations from norms. They could also replace conventional speech features in machine learning models. However, no work has yet assessed the feasibility of speech RIs in MDD. We generated and compared RIs from three reference datasets varying in size, elicitation prompt, and health information. We then calculated deviations from each RI set for people with MDD to compare performance on a depression symptom severity prediction task. Our RI-based models trained with demographic data performed similarly to each other and equivalent models using conventional features or demographics only, demonstrating the value of RI-derived features.

Index Terms: reference intervals, interpretability, speech biomarkers, depression.

1. Introduction

Major Depressive Disorder (MDD) is a common mental disorder that impacts emotional regulation, cognitive functioning and neurophysiological processes, which can lead to alterations in speech and language [1, 2]. Early identification and treatment of MDD is associated with better health outcomes [3]. Remote relapse monitoring could enable early detection of symptom change, allowing clinicians to more easily identify a need for treatment adjustments, thereby improving care [4].

Speech is a complex process requiring cognitive functions and coordination of the respiratory, laryngeal, and articulatory muscles [5]. This complexity makes it sensitive to changes in health [6]. Speech could contribute to the early identification of changes in depression symptom severity [1]. As the presentation of MDD varies widely between individuals [7], we cannot assume that depression affects every individual's speech similarly. This variability makes it difficult to distill salient changes into an interpretable and standardised format for research and clinical settings.

A novel approach to characterise health-related changes in speech is using reference intervals (RIs) [8, 9]. RIs, inspired by clinical laboratory science [10], define a typical 'healthy' range for speech features, which could then be used to assess an individual's speech features against an equivalent population [9]. Previous works have demonstrated that deviation from RIs can be used to classify Alzheimer's or Parkinson's Disease [8]. However, the benefits of RI-derived speech features have yet to be explored in a clinical MDD population.

The choice of reference data is a key consideration when using RI-derived features; RIs should be calculated to match a target population regarding sociodemographic characteristics (e.g. age, sex) [11]. A potential limitation of using datasets from a population sample is that the RIs could be influenced by the presence of diseased subpopulations [12]. Datasets which exclude diseased subpopulations may ensure ranges of features that are more representative of the healthy population and, thus, more reliable. Large data sources are preferred for RI generation [10], with a recommended minimum of 400 individuals per reference group partition [12]. Speech is also subject to channel effects and other sources of measurement variability [13]. It is therefore essential to assess the suitability of different reference datasets in speech RI generation.

This study aimed to explore the feasibility of using RI deviations to predict depression symptom severity. Our specific objectives were to (1) assess whether RI-derived features achieve comparable performance to the conventional, underlying speech features and (2) evaluate how the choice of reference dataset impacts prediction performance.

10.21437/Interspeech.2025-1438

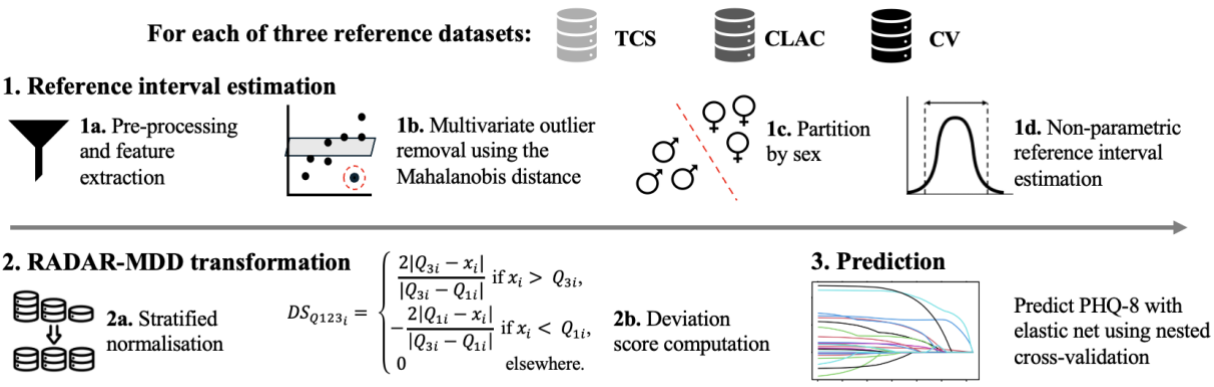

Figure 1: *Key methodological steps in developing (1) reference intervals and (2) deviation scores; and (3) the use of these deviation scores as features in a task predicting depression symptom severity as measured by PHQ-8.*

2. Method

Forming RI-derived features involved (1) generating the RI from a reference dataset and (2) calculating deviation scores for the clinical depression dataset from the derived references (Figure 1). This section describes our reference and clinical datasets, overviews the key methodological steps in forming the deviation features, and outlines our prediction experiments.

2.1 Reference datasets

As part of our feasibility assessment, we formed three sets of RIs using three reference corpora (Table 1).

Thymia Cross-Sectional Dataset (TCS): Comprises audio recordings collected remotely through a web browser API [14]. We used TCS as it has rich health and sociodemographic metadata. Eligibility criteria to partake were to: be aged 18-100; speak English as a first language; and be resident in the UK or US. Participants recorded themselves completing a variety of elicitation tasks and provided sociodemographic data and information on various health outcomes. Participants who self-reported mental, physical, or neurological conditions, allergies or illnesses, who indicated a birth sex other than male or female, or who were missing sex or health data, were excluded. We used speech from recordings of Aesop's fable 'The North Wind and the Sun'.

Crowdsourced Language Assessment Corpus (CLAC): Contains remotely collected data from various speech tasks [15]. We used CLAC as it is a standard open-source reference dataset for health analysis [8, 9]. Participants provided information on gender (used in place of birth sex) and whether they were experiencing current illness or allergies at the time of recording. We excluded participants who reported experiencing a current illness or who did not indicate male or female gender. We used recordings of readings of 'The Rainbow Passage'.

Mozilla Common Voice (CV): A project crowdsourcing open-source speech datasets [16]. Participants are prompted to read a set of short sentences. Each recording goes through a validation process to which anyone can contribute by listening to the clips to accept or reject them based on quality (e.g. based on background noise or misreading). Participants can choose to provide age, accent, and gender (used in place of birth sex). We used CV 1.0 (henceforth, CV). We randomly selected one recording for each voice in the corpus; those who did not indicate gender were excluded.

2.2 Clinical Population

We used speech from people with MDD from an international longitudinal observational cohort study, Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD) [17, 2] (Table 2). Participants were invited to complete two speech tasks: answering a question and reading one of three parts of 'The North Wind and the Sun'. We used only English speech data from the reading task completed by UK-based participants. Participants completed the 8-item Patient Health Questionnaire (PHQ-8), a measure of depression symptom severity [18] where higher scores indicate a more severe symptom presentation.

A patient advisory board (PAB) co-developed the study protocol with input on several study aspects, such as survey frequency, usability of the app used for data collection, documents, incentives, wearable devices, and data analysis. The speech tasks were discussed with the project's PAB.

2.3 Ethical approvals and licenses

Ethical approval for TCS was provided via the *Association of Research Managers and Administrators* service. CLAC and CV are distributed under Creative Commons licenses (CC BY-SA and CC0, respectively). Ethical approval for RADAR-MDD UK was provided by London, Camberwell St. Giles Research Ethics Committee (17/LO/1154). Access to the data can be made through reasonable requests to the senior author and will be subject to local ethics clearances.

Table 1: *Reference descriptives after exclusions and outlier removal. *CV age is organised by categorical age group.*

	Total N [% female]	Mean (SD)	
		Age	Recording length (s)
TCS	2,591 [49.5]	37.2 (12.7)	40.2 (11.8)
CLAC	892 [49.1]	35.7 (12.0)	12.8 (2.2)
CV	4,713 [19.3]	N/A*	2.9 (2.5)

Table 2: *RADAR-MDD descriptives based on 4,242 observations from n=272 (female=213) participants.*

	N observations	Age	Recording length (s)	PHQ-8
Mean	15.6	48.4	14.0	9.2
(SD)	(10.4)	(15.4)	(3.7)	(6.0)

Table 3: *Speech reference intervals (standardised) derived from reference datasets. The remaining 20 features included in our prediction model analysis are listed below. Mean (SD) indicates the unstandardised values for each dataset*.

		Study						
	Feature	TCS		CLAC		CV		RADAR-MDD
		Mean (SD)	Interval	Mean (SD)	Interval	Mean (SD)	Interval	Mean (SD)
Female	Speaking rate	3.5 (0.7)	-2.4, 1.5	3.8 (0.5)	-2.2, 1.8	4.1 (0.9)	-2.0, 2.0	3.9 (0.6)
	Mean F0	186.9 (29.3)	-2.0, 1.7	195.7 (27.9)	-2.1, 1.9	204.3 (36.6)	-1.9, 2.0	182.3 (27.8)
	CPP	9.1 (1.6)	-2.0, 1.9	9.4 (1.6)	-1.8, 2.2	8.7 (1.5)	-1.9, 2.1	7.9 (1.0)
	Mean F1	496.3 (56.1)	-1.6, 2.0	529.0 (64.0)	-1.6, 1.7	519.7 (74.2)	-1.6, 2.2	499.4 (52.8)
	Spectral gravity	478.3 (122.8)	-1.4, 2.6	566.1 (200.9)	-1.2, 2.0	515.0 (176.5)	-1.2, 2.2	440.4 (99.0)
Male	Speaking rate	3.2 (0.7)	-2.4, 1.7	3.7 (0.5)	-1.9, 1.9	4.0 (0.9)	-2.0, 2.0	3.7 (0.6)
	Mean F0	117.2 (28.8)	-1.1, 2.7	117.8 (29.4)	-1.1, 2.5	128.5 (32.1)	-1.2, 2.8	112.1 (20.2)
	CPP	8.2 (1.6)	-1.8, 2.1	8.7 (1.7)	-1.7, 2.2	8.2 (1.6)	-1.8, 2.1	6.8 (0.9)
	Mean F1	452.8 (68.7)	-1.2, 1.9	481.1 (71.6)	-1.3, 2.2	472.3 (79.6)	-1.5, 2.3	462.5 (71.1)
	Spectral gravity	438.7 (151.4)	-1.1, 1.9	531.5 (233.8)	-1.1, 2.3	461.0 (174.6)	-1.3, 2.5	358.8 (86.9)

Note. Other features included in the prediction modelling: (i) Fluency Measures: articulation rate, phonation ratio, pause rate, mean pause duration, (ii) Respiration Measures: mean intensity, intensity range; (iii) Phonation Measures: F0 SD, harmonic to noise ratio, spectral slope, spectral tilt; (iv) Articulatory Measures: F1 SD, mean F1 Bandwidth, F1 Bandwidth SD, mean F2, F2 SD, mean F2 Bandwidth, F2 Bandwidth SD; and (v) Acoustic Measures: mean spectral deviation; mean spectral skewness; mean spectral kurtosis.

2.4 Pre-processing and feature extraction

All speech data were converted into mono Waveform Audio File Format (wav) files, with 16kHz sampling frequency and 16-bit resolution. All speech features were extracted using Parselmouth, which runs Praat in Python [19].

2.5 Reference interval generation and deviation scores

Twenty-five features were used for RI development and depression symptom severity prediction (Table 3). We selected five for presentation: speaking rate, mean fundamental frequency (F0), cepstral peak prominence (CPP), mean first formant (F1), and spectral gravity. These five were selected for their relevance for depression applications [1], representing fluency characteristics of speech and the speech production subsystems of respiration, phonation, and articulation [5].

RIs were developed using the methodology outlined in [8, 9] (Figure 1). First, outliers were removed based on Mahalanobis distance [20] to the population mean with a cutoff of three times the standard deviation from the mean. Data were then partitioned by sex, including only male and female participants to maintain a minimum sample size of n>400 per partition. Each feature within each partition was normalised using Z-score standardisation, and RIs were determined with the non-parametric approach.

To calculate deviation scores, we used the deviation score DS_{Q123} [8] based on the median (Q2), first (Q1) and third (Q3) quartiles for each feature (Figure 1). The score is 0 when a feature is within a reference range, negative when below the range, and positive when above the range. RIs and deviation scores were calculated separately for partitions, but partitions were recombined before prediction. To compare the stability of the RIs we calculated the percentage of overlap between the ranges for each of our five presentation features.

For feature extraction, reference interval generation, deviation score calculations and overlap analysis code, see `https://github.com/LaurenLWhite/IS25`.

2.6 Prediction models

Nine elastic net models were developed to predict depression symptom severity, as measured by PHQ-8, using four feature combinations: (1) A baseline model including age, sex, and years of education only; (2) conventional speech features; (3) deviation from TCS RIs (D-TCS), (4) deviation from CLAC RIs (D-CLAC); and (5) deviation from CV RIs (D-CV). Each model (2-5) was considered with and without age, sex/gender, and education, variables that are known biases in the RADAR-MDD dataset [17, 2] and affect speech [21, 22].

Elastic net is widely used in speech-depression based literature [23] and provides an initial basis to evaluate the utility of RIs for depression prediction. Models were developed using scikit-learn [24]. Hyperparameters were tuned within inner loops using a grid search. The hyperparameters tuned were alpha [0.1, 1.0, 10.0, 100] and L1 ratio [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. Models were evaluated using speaker-stratified nested cross-validation (10 inner, 10 outer folds) to enhance generalisability [25]. Models were assessed in terms of root mean square error (RMSE), mean absolute error (MAE) and the coefficient of determination (R^2).

Table 4: *Percentage overlap of reference interval (RI) ranges between reference datasets.*

	Feature	TCS-CLAC	TCS-CV	CLAC-CV	Mean
Female	Speaking rate	89.3	85.3	95.7	90.1
	Mean F0	93.6	89.0	91.8	91.5
	CPP	87.3	92.9	94.2	91.5
	Mean F1	90.9	91.3	84.5	88.9
	Spectral gravity	78.9	95.8	82.4	85.7
	Mean % overlap	88.0	90.9	89.7	89.5
Male	Speaking rate	83.0	81.4	94.2	86.2
	Mean F0	97.0	96.8	93.9	95.9
	CPP	97.3	98.2	98.6	98.0
	Mean F1	91.1	82.5	90.6	88.1
	Spectral gravity	88.6	80.4	88.7	85.9
	Mean % overlap	91.4	87.9	93.2	90.8

3. Results and Discussion

3.1 Reference interval generation and deviation scores

Table 3 describes RIs of five example features for the reference datasets, alongside the mean and standard deviation of the unstandardised data for all corpora. Overlaps in RIs varied across datasets, features, and sex/gender (Table 4). The average overlap across datasets also differed between males and females: the overlap between TCS and CV was greatest for females, but lowest for males. Conversely, the CLAC and CV RI overlap was greatest for males yet was the lowest for females.

Regarding individual features, RI overlap for F1 and spectral gravity was similar for males and females. However, RIs for F0 and CPP were less variable (i.e. displayed more overlap) for males than females, whereas speaking rate RIs were less variable for females than males. These results emphasise the importance of partitioning the data [9], and suggest that, despite all RIs being formed from n>400 samples, potentially larger data sources are needed to provide more stable intervals for speech applications.

A key feasibility issue related to the consistency of elicitation tasks in the different reference datasets. Due to short readings in CV, many samples did not contain pauses, meaning we could not calculate valid deviation scores for pause rate, pause duration, or phonation ratio. This presents a limitation of CV as a reference dataset for MDD as pause differences are often observed in depression [1].

3.2 Prediction models

As in [8], comparing the RI-derived features with conventional features did not result in information loss (Table 5). We also observed similar performance across D-TCS, D-CLAC, and D-CV. Thus, there does not seem to be any added benefit of using data from a strictly healthy population (TCS). These results highlight the usefulness of RI representations.

Interestingly, all models performed similarly, but the baseline model outperformed speech-only models, highlighting the importance of the demographic predictors in the clinical dataset. These results support concerns in the literature that while speech-based methods are promising, model performance in the speech-health literature may present over-optimistic results due, in part, to the presence of confounds [26, 27].

Table 5: *RMSE, R^2, and MAE from nine elastic net models predicting PHQ-8 using 25 speech features with or without demographic data (sex/gender, age, and education).*

		Model	RMSE	R^2	MAE
With demographics?	Yes	(1) Baseline	5.97	0.01	4.93
		(2) Conventional	5.83	0.05	4.77
		(3) D-TCS	5.87	0.04	4.82
		(4) D-CLAC	5.86	0.05	4.81
		(5) D-CV	5.85	0.05	4.81
	No	(2) Conventional	5.98	0.01	4.98
		(3) D-TCS	5.98	0.01	4.97
		(4) D-CLAC	5.97	0.01	4.96
		(5) D-CV	5.98	0.01	4.98

3.3 Limitations and future work

Our work had several limitations. We could not form valid RIs for all features using CV due to the short readings. Additionally, for some speech features, deviation from RIs may only be clinically meaningful when the deviation is in a specific direction [8]. Future work should consider a literature-informed approach to the upper and lower bounds of RIs [8] and develop RIs partitioned by age *and* sex [8, 9]. This may improve the validity of RI-derived features and give insights into the suitability of different datasets for RI development. We also focused on knowledge-derived features; future work could combine RIs with features derived from foundational models.

The RADAR-MDD dataset contains repeated, longitudinal measures. Work from [28] highlighted that repeated observation of individuals can improve precision when predicting mental health symptom severity with a reference-based approach; future work could incorporate repeated observations to develop person-specific RIs [10]. This work focused on analysis of a reading task. Future work could explore RI-derived features from other speech tasks that may also contain suitable MDD related biomarkers [1].

Finally, our analysis took a complete case approach in removing all missing data and used basic prediction models; missing data imputation and more sophisticated prediction algorithms may present more insightful results.

4. Conclusion

This work explored the feasibility of using RIs to predict depression symptom severity. We observed comparable performance using RI-derived features versus conventional speech features (Objective 1). Given the greater interpretability provided by RI-derived features, RIs may provide a valuable tool for advancing speech as a depression biomarker. Moreover, RIs represent a more convenient means of data sharing, as data are processed and aggregated in a way that removes any potentially sensitive participant information.

We observed comparable predictive performance for RI-derived features from different reference datasets (Objective 2). Datasets with very short speech recordings (such as CV) may be difficult to use for valid RI generation due to the inability to determine some features. Further developments to RIs in speech should evaluate the characteristics of potential reference datasets before proceeding with RI generation.

5. 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. 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. We thank all RADAR-CNS patient advisory board members for contributing to the device selection procedures and providing 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. Lauren White, Pre-Doctoral Fellow NIHR303473, is funded by the NIHR for this research project. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, NHS or the UK Department of Health and Social Care.

6. References

[1] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, Jan. 2020, doi: 10.1002/lio2.354.

[2] 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, Nov. 2023, doi: 10.1016/j.jad.2023.08.097.

[3] C. Kraus, B. Kadriu, R. Lanzenberger, C. A. Zarate, and S. Kasper, "Prognosis and improved outcomes in major depression: a review," *Translational Psychiatry*, vol. 9, no. 1, Apr. 2019, doi: 10.1038/s41398-019-0460-3.

[4] E. E. Thomas *et al.*, "Factors influencing the effectiveness of remote patient monitoring interventions: a realist review," *BMJ Open*, vol. 11, no. 8, Aug. 2021, doi: 10.1136/bmjopen-2021-051844.

[5] V. Ramanarayanan, A. C. Lammert, H. P. Rowe, T. F. Quatieri, and J. R. Green. "Speech as a biomarker: opportunities, interpretability, and challenges." *Perspectives of the ASHA Special Interest Groups* vol 7, no. 1 pp. 276-283, Jan. 2022, doi: 10.1044/2021_PERSP-21-00174

[6] J. D. Singh Sara, D. Orbelo, E. Maor, L. O. Lerman, and A. Lerman, "Guess What We Can Hear – Novel Voice Biomarkers for the Remote Detection of Disease," *Mayo Clinic Proceedings*, vol. 98, no. 9, Mar. 2023, doi: 10.1016/j.mayocp.2023.03.007.

[7] C. Oetzmann *et al.*, "Identifying depression subtypes and investigating their consistency and transitions in a 1-year cohort analysis," *PLOS ONE*, vol. 20, no. 1, p. e0314604, Jan. 2025, doi: 10.1371/journal.pone.0314604.

[8] C. Botelho, A. Abad, T. Schultz, and I. Trancoso, "Speech as a Biomarker for Disease Detection," *IEEE Access*, vol. 12, pp. 1–1, Jan. 2024, doi: doi.org/10.1109/access.2024.3506433.

[9] C. Botelho, A. Abad, Schultz. T, and I. Trancoso, "Towards reference speech characeterization for health applications,' in *Proc. Interspeech 2023.* ISCA, Dublin, Ireland, pp. 2363-2367, 2023. doi: 10.21437/Interspeech.2023-1435.

[10] Y. Ozarda, "Reference intervals: current status, recent developments and future considerations," *Biochemia Medica*, vol. 26, no. 1, pp. 5–16, 2016, doi: 10.11613/bm.2016.001.

[11] Y. Ozarda, K. Sikaris, T. Streichert, and J. Macri, "Distinguishing reference intervals and clinical decision limits – A review by the IFCC Committee on Reference Intervals and Decision Limits," *Critical Reviews in Clinical Laboratory Sciences*, vol. 55, no. 6, pp. 420–431, Jul. 2018, doi: 10.1080/10408363.2018.1482256.

[12] G. R. D. Jones *et al.*, "Indirect methods for reference interval determination – review and recommendations," *Clinical Chemistry and Laboratory Medicine (CCLM)*, vol. 57, no. 1, pp. 20–29, Apr. 2018, doi: 10.1515/cclm-2018-0073.

[13] M. Brockmann-Bauser and M. F. de Paula Soares, "Do We Get What We Need from Clinical Acoustic Voice Measurements?," *Applied Sciences*, vol. 13, no. 2, p. 941, Jan. 2023, doi: 10.3390/app13020941.

[14] S. Fara, S. Goria, E. Molimpakis, N. Cummins, "Speech and the n-Back task as a lens into depression. How combining both may allow us to isolate different core symptoms of depression," in *Proc. Interspeech 2022.* ISCA, Incheon, Korea, pp. 1911-1915, 2022. doi: 10.21437/Interspeech.2022-10393.

[15] R. Haulcy, and J. Glass, "CLAC: A Speech Corpus of Healthy English Speakers," in *Proc. Interspeech 2021.* ISCA, Brno, Czechia, pp. 2966-2970, 2021. doi: 10.21437/Interspeech.2021-1810.

[16] R. Ardila *et al.*, "Common Voice: A Massively-Multilingual Speech Corpus," *arXiv.org*, Mar. 05, 2020. https://arxiv.org/abs/1912.06670.

[17] 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, Feb. 2019, doi: 10.1186/s12888-019-2049-z.

[18] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. W. Williams, J. T. Berry, and A. H. Mokdad, "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, Apr. 2009, doi: 10.1016/j.jad.2008.06.026.

[19] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, Nov. 2018, doi: 10.1016/j.wocn.2018.07.001.

[20] P. C. Mahalanobis, "Reprint of: Mahalanobis, P.C. (1936) 'On the Generalised Distance in Statistics.,'" *Sankhya A*, vol. 80, no. S1, pp. 1–7, Dec. 2018, doi: 10.1007/s13171-019-00164-5.

[21] P. Torre and J. A. Barlow, "Age-related changes in acoustic characteristics of adult speech," *Journal of Communication Disorders*, vol. 42, no. 5, pp. 324–333, Sep. 2009, doi: 10.1016/j.jcomdis.2009.03.001.

[22] A. L. Jefferson *et al.*, "A Life Course Model of Cognitive Activities, Socioeconomic Status, Education, Reading Ability, and Cognition," *Journal of the American Geriatrics Society*, vol. 59, no. 8, pp. 1403–1411, Jul. 2011, doi: 10.1111/j.1532-5415.2011.03499.x.

[23] S. A. Almaghrabi, S. R. Clark, and M. Baumert, "Bio-acoustic features of depression: A review," *Biomedical Signal Processing and Control*, vol. 85, p. 105020, Aug. 2023, doi: 10.1016/j.bspc.2023.105020.

[24] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, Oct. 2011, Available: https://jmlr.org/papers/v12/pedregosa11a.html

[25] G. Cawley and N. Talbot, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079–2107, Jul. 2010, Available: https://jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

[26] R. Polle, S. Fara, S. Goria, and N. Cummins, "Revealing Confounding Biases: A Novel Benchmarking Approach for Aggregate-Level Performance Metrics in Health Assessments," in *Proc. Interspeech 2024.* ISCA, Kos, Greece, pp. 1440-1444, 2024. doi: 10.21437/Interspeech.2024-1092.

[27] V. Berisha, C. Krantsevich, G. Stegmann, S. Hahn, and J. Liss "Are reported accuracies in the clinical speech machine learning literature overoptimistic?," in *Proc. Interspeech 2022.* ISCA, Incheon, Korea, pp. 2453-2457. doi: 10.21437/Interspeech.2022-691.

[28] E. Larsen et al., "Validating the efficacy and value proposition of mental fitness vocal biomarkers in a psychiatric population: prospective cohort study," *Frontiers in Psychiatry*, vol. 15, Mar. 2024, doi: 10.3389/fpsyt.2024.1342835.