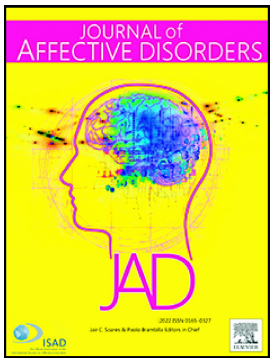
Multilingual markers of depression in remotely collected speech samples: A preliminary analysis



Nicholas Cummins, Judith Dineley, Pauline Conde, Faith Matcham, Sara Siddi, Femke Lamers, Ewan Carr, Grace Lavelle, Daniel Leightley, Katie M. White, Carolin Oetzmann, Edward L. Campbell, Sara Simblett, Stuart Bruce, Josep Maria Haro, Brenda W.J.H. Penninx, Yatharth Ranjan, Zulqarnain Rashid, Callum Stewart, Amos A. Folarin, Raquel Bailón, Björn W. Schuller, Til Wykes, Srinivasan Vairavan, Richard J.B. Dobson, Vaibhav A. Narayan, Matthew Hotopf, The RADAR-CNS Consortium

PII: S0165-0327(23)01076-5

DOI: https://doi.org/10.1016/j.jad.2023.08.097

Reference: JAD 16470

To appear in:

Received date: 6 May 2023

Revised date: 16 August 2023

Accepted date: 17 August 2023

Please cite this article as: N. Cummins, J. Dineley, P. Conde, et al., Multilingual markers of depression in remotely collected speech samples: A preliminary analysis, (2023), https://doi.org/10.1016/j.jad.2023.08.097

Multilingual markers of depression in remotely collected speech samples: A preliminary analysis

Nicholas Cummins[1,], Judith Dineley[1,2], Pauline Conde[1], Faith Matcham,[3,4], Sara Siddi[5], Femke Lamers[6], Ewan Carr[1], Grace Lavelle[3], Daniel Leightley[4], Katie M. White[4], Carolin Oetzmann[4], Edward L. Campbell[1,7], Sara Simblett[8], Stuart Bruce[9], Josep Maria Haro[5], Brenda W. J. H. Penninx[6], Yatharth Ranjan[1], Zulqarnain Rashid[1], Callum Stewart[1], Amos A. Folarin[1,10], Raquel Bailón[11,12], Björn W. Schuller[2,13], Til Wykes[8,10], Srinivasan Vairavan[14], Richard J.B. Dobson[1,15], Vaibhav A. Narayan[16], Matthew Hotopf[4,10], The RADAR-CNS Consortium[17]*

[1] Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[2] Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3] School of Psychology, University of Sussex, Falmer, UK
[4] Department of Psychological Medicine, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[5] Parc Sanitari Sant Joan de Déu, Fundació Sant Joan de Déu, CIBERSAM, Barcelona, Spain
[6] Department of Psychiatry, Amsterdam Public Health Research Institute and Amsterdam Neuroscience, Amsterdam University Medical Centre, Vrije Universiteit and GGZ InGeest, Amsterdam, NL
[7] GTM research group, AtlanTTic Research Center, University of Vigo, Spain
[8] Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London
[9] RADAR-CNS Patient Advisory Board, King's College London, UK
[10] NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK
[11] Biomedical Signal Interpretation and Computational Simulation (BSICoS) group, Aragon Institute for Engineering Research, University of Zaragoza, Zaragoza, Spain
[12] Biomedical Research Networking Center in Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Spain
[13] GLAM - Group on Language, Audio, & Music, Imperial College London, London, UK
[14] Janssen Research and Development LLC, Titusville, NJ, United States
[15] Institute of Health Informatics, University College London, London, UK
[16] Davos Alzheimer's Collaborative
[17] www.radar-cns.org

* Corresponding Author: nick.cummins@kcl.ac.uk, +44 (0)20 7848 0304

Abstract

Background: Speech contains neuromuscular, physiological and cognitive components, and so is a potential biomarker of mental disorders. Previous studies indicate that speaking rate and pausing are associated with major depressive disorder (MDD). However, results are inconclusive as many studies are small and underpowered and do not include clinical samples. These studies have also been unilingual and use speech collected in controlled settings. If speech markers are to help understand the onset and progress of MDD, we need to uncover markers that are robust to language and establish the strength of associations in real-world data.

Methods: We collected speech data in 585 participants with a history of MDD in the United Kingdom, Spain, and Netherlands as part of the RADAR-MDD study. Participants recorded their speech via smartphones every two weeks for 18 months. Linear mixed models were used to estimate the strength of specific markers of depression from a set of 28 speech features.

Results: Increased depressive symptoms were associated with speech rate, articulation rate and intensity of speech elicited from a scripted task. These features had consistently stronger effect sizes than pauses.

Limitations: Our findings are derived at the cohort level so may have limited impact on identifying intra-individual speech changes associated with changes in symptom severity. The analysis of features averaged over the entire recording may have underestimated the importance of some features.

Conclusions: Participants with more severe depressive symptoms spoke more slowly and quietly. Our findings are from a real-world, multilingual, clinical dataset so represent a step-change in the usefulness of speech as a digital phenotype of MDD.

1. Introduction

Speech is uniquely placed in digital health: no other signal contains a combination of cognitive, neuromuscular, and physiological information. Speech is relatively simple to collect in daily life via *remote measurement technologies* (RMT) such as smartphones. Speech phenotypes could therefore become scalable digital biomarkers of health, providing insights into both current and predicted future health outcomes. A growing body of research has demonstrated associations between depression and changes in specific acoustic and prosodic properties of speech (Cummins et al., 2015; Low et al., 2020). However, many of these findings are from small samples, are potentially underpowered, and should be treated as preliminary.

Most reported effects are observed in cross-sectional studies and include a range of prosodic and acoustic alterations such as flattened pitch contours and altered formant (measures of vocal tract resonances) trajectories (Cummins et al., 2015; Low et al., 2020). As well as being cross-sectional, most speech and depression research published in the last ten years has focused on model development using large multivariate feature spaces and machine learning paradigms while there have been comparatively few works in phenotype identification or speech feature characterisation (Cummins et al., 2015; Low et al., 2020). Most studies have also used only two publicly available datasets, the Audio-Visual Depressive Language (AViD) corpus and the Distress Analysis Interview Corpus (DAIC) (Gratch et al., 2014; Ringeval et al., 2019; Valstar et al., 2013). Both corpora are subject to two main limitations. Firstly, the speech collected cannot be regarded as clinical samples as they come from volunteers who have had their depression severity established through a questionnaire at the time of the study. Secondly, the metadata associated with both datasets is sparse, so potential confounding factors are unspecified.

Few observational speech depression studies have assessed the predictive power of individual speech features. In a six week study of 35 English-speaking participants, increases in pause time, and speaking rate, as well as decreases in the variation of second formant location were significantly associated with increasing depression severity (Mundt et al., 2007). A subsequent four-week study of 165 English-speaking participants observed that only increased speaking rate and pause time were associated with increased depression severity (Mundt et al., 2012). Association between these measures and depression severity have been replicated. Participants of a clinical trial for treatment response in depression (N=50) found that pauses became shorter and less variable as depression decreased (Yang et al., 2013). More recently, speech-to-pause ratio was associated with depression severity in a 4-week study of 18 English-speaking participants (Abbas et al., 2021). Finally, a 10-week study of 241 Japanese-speaking

participants also found slower speech rate and increased pause time were associated with greater depression severity (Yamamoto et al., 2020).

These observational studies highlight the links between changes in depression severity and corresponding changes in speaking rate and pausing (Abbas et al., 2021; Mundt et al., 2012, 2007; Yamamoto et al., 2020; Yang et al., 2013). However, all studies were short-term, with the longest having only four observations over 21 weeks (Yang et al., 2013). They were also unilingual studies conducted on speech collected in highly controlled circumstances. The Voiceome Dataset has recently been released in which over 7,000 people completed speech recordings up to 4 times over a 12-week period using personal devices (Schwoebel et al., 2021). Voiceome has limited value in the study of depression with the vast majority of participants reporting as non-depressed. The mean 9-item Patient Health Questionnaire (PHQ-9) in Voiceome is 4.5, which sits on the border between no symptoms and mild symptom severity (Löwe et al., 2004).

To the best of the authors knowledge, no longitudinal studies have examined how specific languages, a known source of variability in speech (Ambikairajah et al., 2011), affect these markers. Identifying language-independent markers of depression would increase the clinical effectiveness of speech-phenotypes, for example, by opening them up for inclusion in large multinational clinical trials. A small number of cross-sectional studies have attempted to isolate language independent markers (Alghowinem et al., 2016; Kiss and Vicsi, 2017; Mitra et al., 2015). However, these studies compare speech samples collected with different elicitation and collection strategies, limiting the robustness of their findings. No studies to date, have investigated speech parameters and depression over multiple time points with speech collected via RMT. Evaluating such associations in real-world data is vital to understanding the role that speech analysis could ultimately play in the management of chronic conditions such as depression.

We used data collected in the major European Innovative Medicines Initiative (IMI2) *Remote Assessment of Disease and Relapse in Major Depressive Disorder* (RADAR-MDD) programme, a longitudinal cohort study examining the utility of multi-parametric RMT to predict changes in symptoms and relapse in people with MDD (Matcham et al., 2019). These data address previous limitations as the dataset (i) contains speech samples from the largest clinical cohort study utilising RMT (Matcham et al., 2022); (ii) collected longitudinally in the real world; and (iii) is multilingual.

We aimed to estimate cross-language associations of specific speech markers and depression from a smaller set of relevant features identified from the literature from remotely collected speech samples. We described the sociodemographic and clinical characteristics of the cohort and conducted analyses on these factors to identify potential biases in data availability.

2. Methods

2.1 Study Design

RADAR-MDD was an observational cohort study of individuals with established MDD from three recruitment sites: *King's College London* (KCL, London, United Kingdom); *Amsterdam UMC, Vrije Universiteit* (VUmc; Amsterdam, Netherlands); and *Centro de Investigación Biomédica en Red del Área Salud Mental* (CIBERSAM; Barcelona, Spain). The study protocol, eligibility and exclusion criteria have previously been reported (Matcham et al., 2022, 2019). Briefly, the core eligibility criteria were having met the DSM-5 diagnostic criteria for non-psychotic MDD within the past two years prior to enrolment and having recurrent MDD (lifetime history of at least two episodes). All participants were aged over 18 and able to give written informed consent.

We regard the RADAR-CNS protocol to be a clinical sample, as the study population were current or past clinical service users. All have been diagnosed and treated for MDD, with their most recent episode within 2 years of enrolment (Matcham et al., 2019). Each participant completed a Lifetime Depression Assessment – Self Report (LIDAS; (Bot et al., 2017)) at baseline as an additional layer of diagnosis confirmation.

2.2 Ethics

Ethical approval was obtained from the *Camberwell St. Giles Research Ethics Committee* (17/LO/1154) in London from the *Fundacio Sant Joan de Deu* Clinical Research Ethics Committee (CI: PIC-128-17) in Barcelona, and from the *Medische Ethische Toetsingscommissie VUmc* (2018.012–NL63557.029.17) in Amsterdam.

2.3 Speech collection

RADAR-MDD was already an active study when speech collection began (Matcham et al., 2022). Speech collection began in London in August 2019 and in Barcelona and Amsterdam in December 2019. Participants already enrolled when speech collection commenced were informed about speech collection in a newsletter and were provided with a link to a private YouTube instruction video. Participants who enrolled after the start of speech collection were

briefed either face-to-face or remotely. The most important instructions were also provided in the purpose-made study smartphone application (app) (Ranjan et al., 2019).

Participants were asked to record themselves, speaking in the language native to their recruitment site, completing two speech elicitation tasks every two weeks. The recordings were collected via the RADAR-base active RMT (aRMT) data collection app (Ranjan et al., 2019). The app produced notifications each time speech recordings were scheduled. Before recording, participants were reminded, via on-screen instructions, to find a quiet place to complete the recordings and speak in their normal voice.

The first activity was a *scripted speech task*, in which participants read aloud an extract from Aesop's fable, *The North Wind and The Sun* (International Phonetic Association, 1999); the extracts for each language are provided in Supplementary Tables 1-3. To minimise practice effects, the fable was split into three parts and participants were prompted to read a different extract at each recording. The second activity was a *free response speech task*, in which participants were asked to speak about something they were looking forward to in the next seven days (Mundt et al., 2007). Participants were given the option of re-recording their speech in each task, for example, if they were interrupted while recording and could skip the free-speech task.

As a safeguarding issue to discourage participants recording messages expressing suicidal ideation or intent, it was made clear to participants when they were introduced to the speech tasks that we would not be listening to the free speech audio while RADAR-MDD was an active study. Once recorded, the speech data were encrypted into a single file tagged with the participant's study ID number and sent to a secure server.

2.4 Speech data preparation

The collected data were decrypted into *Waveform Audio File Format* (WAV) files with a sampling frequency of 16 kHz and a 16-bit resolution; a separate file was created for each task. Some data could not be decrypted, so we define *audio files* as those files as WAV files that can be played on standard audio editing software such as Audacity (Franklin, 2006). Files that did not meet this criterion were not used in the analysis.

We then extracted a set of 28 speech features from the audio files. These features are categorised into three groups: (1) Speech Timing Measures, estimated via intensity thresholds (de Jong et al., 2021); (2) Prosodic and Phonation; and (3) Articulatory Measures. Features were extracted using Parselmouth (Jadoul et al., 2018), an open-source Python library that

enables the use of Praat, a software package for speech analysis (Boersma, 2001). All prosodic, phonation and articulatory measures were extracted using default Praat settings.

Details on the features are provided in Supplementary Tables 4-6.

The following three criteria were used to determine if a file was included in our analysis. Firstly, files shorter than 2 seconds were removed from the analysis on the assumption that they were less likely to contain analysable speech. Secondly, an audio file was included if Parselmouth could return a value for all 28 features, otherwise we assume that there was a considerable amount of corrupting noise in the file. The third criterion was that a participant had to supply a minimum of two audio files for each task, i.e., the minimum number of files necessary for the data to be considered longitudinal.

2.5 Depression Assessments

We used the Inventory of Depressive Symptomatology – Self Report (IDS-SR) (Rush et al., 2000) and the 8-item Patient Health Questionnaire (PHQ-8) scale (Kroenke et al., 2009). The IDR-SR was used to identify the presence of *depressive symptoms* at baseline. For our analysis, we define baseline depression to be the IDS-SR score obtained within six weeks of a participant first being scheduled to participate in the speech task. The PHQ-8, which gives a self-reported depressive symptom severity, was collected remotely and concurrently with the speech recordings, every two weeks, via the aRMT RADAR-Base app (Ranjan et al., 2019). Given speech samples and PHQ-8 were collected concurrently, we use the PHQ-8 scores to identify key cross-language speech based markers of depression.

2.6 Patient Involvement

The RADAR-MDD 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 has been discussed specifically with the RADAR-CNS *Patient Advisory Board* (PAB), and a member of PAB is also a co-author of this manuscript.

2.7 Cohort Description and Bias Assessment

We conducted analyses to identify potential biases in the composition of analysable speech data based on sociodemographic variables and depression severity (IDS-SR score). Sociodemographic factors assessed were, age, sex assigned at birth, height (as a proxy of

vocal tract length), and years of education (a proxy for reading ability and verbal IQ). These factors were considered as they are non-transient speaker characteristics which can affect acoustic and prosodic speech markers (Jefferson et al., 2011; Schuller et al., 2013).

We first describe each measure using medians and interquartile ranges, then used Chi-squared (sex assigned at birth) and Wilcoxon signed-rank (age, height, years of education, depression severity) tests to determine whether there were differences in the proportion of participants providing analysable speech samples. This analysis was conducted using IBM SPSS Statistics software, and was performed separately by language (English, Dutch, Spanish) and for the scripted and free response speech tasks individually.

2.8 Relationship of speech markers with depression

We used linear mixed effect models (LMEs) to estimate associations between PHQ-8 depression scores and 28 speech features from concurrently collected speech. Each speech feature was tested in a separate model. Since the amount of speech data varied between participants (Matcham et al., 2022), LMEs allowed us to analyse the varying amounts of data provided by different participants (Bagiella et al., 2000; Ibrahim and Molenberghs, 2009). We included random intercepts per participant to account for the intra-individual clustering of repeated fortnightly assessments. Each speech feature was standardised before modelling (mean = 0; standard deviation (SD) = 1) to improve estimation and interpretability. Following evidence from past studies, we included age, height, gender, and years spent in education as covariates. All models were estimated separately by language and by tasks. Our LMEs were estimated using the lme4 package for R (Bates et al., 2015).

As our aim is to estimate associations between specific speech markers and depression, estimates are reported as standardised coefficients and 95% bootstrap confidence intervals (Greenland et al., 2016). These represent the difference in the outcome (PHQ-8 score) per 1 SD difference in the respective speech feature, where negative differences represent lower feature values in the presence of increased depression symptom severity. Confidence intervals were estimated using a parametric percentile bootstrap with 1000 iterations, implemented using the confint.merMod method from the lme4 package (Bates et al., 2015).

3. Results

3.1 Cohort characteristics

A total of 585 participants were enrolled in RADAR-MDD during the speech collection period. The largest cohort was in the United Kingdom with 325 participants (56%), followed by Spain

with 143 participants (24%) and the Netherlands with 117 participants (20%). All cohorts have a larger female representation; 78% in the Netherlands, 76% in the United Kingdom, and 71% in Spain. The depression scores at baseline indicate in each country occur in the range classified as moderate severity for the IDS-SR. Full details of the distribution of our sociodemographic and clinical variables are given in Supplementary Table 7.

The final analytical sample contained 461 (79%) individuals who had analysable data on one or both tasks (457/585 (78%) with scripted task data, 435/585 (74%) with free response task data, 431/585 (74%) provided information for both tasks – see Figure 1). No baseline demographic and clinical depression characteristics were associated with who did or did not provide analysable speech data in either speech tasks for the British (Supplementary Table 8) and Dutch (Supplementary Table 9) cohorts. Baseline depression severity was significantly higher (p=.024) for the Spanish participants who provided analysable scripted speech versus those who did not. Years in Education was also significantly higher (p=.009) for the Spanish participants who provided analysable free-response speech versus those who did not (Supplementary Table 10).

Figure 1

Speech collection was active in RADAR-MDD for a period of 620 days. The median speech collection period was 433 days (interquartile range (IQR): 358-473 days, range: 4-590 days). As speech recording was scheduled once every two weeks, this resulted in a median of 31 recording opportunities (interquartile range (IQR): 26-34, range: 1-43). A more detailed breakdown of the number of files analysed for the scripted and free response task for each country, as well as descriptive statistics of the corresponding PHQ-8 files are given in Supplementary Table 11. A comparison of the PHQ-8 distributions (per language, per speech task) is given in Supplementary Figure 1.

3.2 Associations between speech features and depression severity

Heatmaps per collection site, per speech task correlation are given for all features in Supplementary Figures 2-7. As expected, correlations are higher within each feature group; i.e., speech timing features are more correlated with themselves than with prosodic or articulatory features.

Speech timing: For the scripted task, we found speaking and articulation rates to be strictly negatively associated with depressive severity in all three languages (Figure 3; Supplementary Table 12). This observation indicates that participants with more severe depressive symptoms

spoke more slowly, regardless of the language being spoken. For example, a 1 SD increase in speaking rate was associated with a 0.20 unit (95% CI [-0.32, -0.07]), 0.44 unit (95% CI [-0.69, -0.21]) and a 0.27 unit (95% CI [-0.51, -0.02]) decrease in the subsequent PHQ-8 score, in the UK, Dutch and Spanish cohorts, respectively.

Recording duration was positively associated with depressive severity in the UK and Dutch cohorts, however this trend is not as clear in the Spanish data. Phonation ratio (positive) and number of pauses (negative) were associated with depressive severity for the UK cohort only, with similar trends observable in the Dutch and Spanish cohorts but with associated confidence intervals crossing zero.

For the *free response task*, phonation ratio, speaking rate and articulation rate were negatively associated with depressive severity in the UK and Dutch cohorts, while average syllable duration was positively associated in these countries (Figure 4; Supplementary Table 12). A larger negative β coefficient indicates a similar trend for phonation ratio in the Spanish data but there is no evidence of associations in the other features. Recording duration (negative), phonation time (negative), number of syllables (negative), mean length run (negative), and average pause duration (positive) were associated with depressive severity in the UK cohort only. Similar trends can be seen for phonation time (Dutch, Spanish), number of syllables (Dutch, Spanish), mean length run (Dutch), and average pause duration (Dutch) in the other language groups, though with associated confidence intervals crossing zero.

Prosodic and Phonation: In the scripted task, as with speaking rate and articulation rate, we found that intensity was strictly negatively associated with depressive severity in all three languages (Figure 3; Supplementary Table 13).

A 1 SD increase in speaking rate was associated with a 0.28 unit (95% CI [-0.45, -0.10]), 0.43 unit (95% CI [-0.73, -0.13]) and a 0.34 unit (95% CI [-0.63, -0.08]) decrease in the subsequent PHQ-8 score, in the UK, Dutch and Spanish cohorts, respectively. Harmonic to noise ratio (HNR) was positively associated with depressive severity in the UK and Dutch cohorts, while shimmer was negatively associated. A similar, but not clear trend can be seen for jitter in the Spanish cohort. Finally, mean pitch was negatively associated with depressive severity for the UK cohort only, with a negative trend visible in the Dutch cohort.

For the free response task, mean pitch (negative), intensity (negative), and fraction of unvoiced frames (positive) were associated with depressive severity in the UK and Dutch cohorts (Figure 4; Supplementary Table 13). Mean pitch and intensity also display negative tendencies in the

Spanish cohort. Jitter was positivity associated with depressive severity in the UK and Spanish cohorts. The number of voice breaks (UK) and pitch standard deviation (Dutch) were negatively associated with depression.

Articulatory Measures: Associations between articulatory features and depression severity were observed within single languages only. Each country returned one associated formant change in the *scripted task* (Figure 3; Supplementary Table 14). Decreasing mean F1 frequency, decreasing standard deviation in F2 bandwidth, and increasing standard deviation in F2 frequency with increasing depression were observed in the UK, Dutch, and Spanish data respectively. There is evidence of decreasing mean F1 frequency in the Spanish data with increasing depression, again though the associated confidence interval crosses zero. The only association observed in the *free response task* was an increase in mean F1 bandwidth with increasing depression in the Dutch cohort (Figure 4; Supplementary Table 14).

4. Discussion

The RADAR-MDD speech dataset is unique in its scale, longitudinal duration and the number of languages recorded. Participants provided speech samples for around 62 weeks, compared to a maximum 21 weeks in previous studies (Yang et al., 2013). The dataset also provides speech recordings in three languages, all collected using the same speech elicitation tasks and software platform in contrast to other datasets that contain only one language (Cummins et al., 2015; Low et al., 2020). We found that as depression increases, participants spoke slower and more quietly, whichever language they used. A decrease in speaking rate has been observed in previous longitudinal studies (Abbas et al., 2021; Mundt et al., 2012, 2007; Yamamoto et al., 2020; Yang et al., 2013), but this is the first time that a decrease in intensity with increasing levels of depression has been observed in a longitudinal study.

There are two ways slow speech rate can occur; with the insertion of longer pauses, or by decreasing the rate of speech sound production (Cannizzaro et al., 2004; Cummins et al., 2015). The insertion of pauses is linked with cognitive impairments, while a decrease in the rate of speech sound production is more reflective of psychomotor impairments (Cannizzaro et al., 2004). As decreases in speech rate and articulation rate have stronger effects than any of the pausing measures, we can infer that decreases in speech rate are due to increases in phonation time rather than increases in pause rate. The changes we observed are therefore more likely to be due to increases in neuromuscular impairment affecting the rate of speech production. Decreases in intensity with increased levels of depression are not universally reported in the literature (Cummins et al., 2015; Low et al., 2020). However, most studies which

do not report a significant association are based on small samples and so may be underpowered.

Out of the pausing measures tested, number of pauses (scripted) and average pause duration (free response) were consistently (i.e., CI not crossing zero) associated with depression in the UK cohort. Most other studies that have reported this finding were conducted in English (Abbas et al., 2021; Mundt et al., 2012, 2007; Yamamoto et al., 2020; Yang et al., 2013), and as pausing characteristics can vary between languages (Werner et al., 2022), it is conceivable that this finding only applies to English speaking countries. Further research on non-English corpora is required to verify this conjecture. We found no evidence of associations between pause rate and depressive symptoms in any speech tasks or language, in contrast to previous findings (Abbas et al., 2021; Mundt et al., 2012, 2007; Yang et al., 2013). This discrepancy may be due to statistics methodology (we took account of clustering of repeated measures within individuals) and none of the other studies were conducted in real world settings, meaning the speech collected could also be subject to the observer effects (Wagner et al., 2015). This is a phenomenon in which the speaking style of a participant changes due to the presence of a researcher or clinician during the recording session.

Aside from speaking intensity, we did not observe effects in our prosodic and phonation features consistently across all three countries. Prosodic patterns vary between languages, even within language families, and while phonemes are shared across different languages; their characteristics depend on the phonetic constraints of their underlying language (Ambikairajah et al., 2011; Moskvina, 2013). Similarly, while we did observe a small number of formant changes associated with depression severity, there were no consistent cross-language results for these features. Studies that have previously reported significant formant findings have either used different speech elicitation methods, e.g., extended vowel sounds (Mundt et al., 2007) or were cross-sectional (Scherer et al., 2016).

We have identified the following limitations of this study. Firstly, these findings are demonstrated at the cohort level and so have limited impact on identifying changes within an individual, which is the focus of our planned future research. Secondly, we did not collect the data in a laboratory, so our observed intensity effects could be artefacts in participant interactions with the recording equipment and conditions (Dineley et al., 2023). However, given the number of recordings collected, this is unlikely. Thirdly, we analysed features extracted over the entire duration of the recordings and a phonetic transcription of the data would enable a more fine-grained analysis of different voice quality and formant effects, so we may have underestimated their importance.

Fourthly, neuromuscular impairment tends to be higher in severe cases of MDD, while our participants have, on average more moderate levels. Therefore, more marked effects of depression on speech may exist that are not observable in our dataset. Fifthly, we do not directly control for the reading ability and verbal IQ of our participants. The relationship between verbal fluency, verbal IQ and general executive impairment in depression is complicated and not well understood (Henry and Crawford, 2005). Therefore, it is difficult to speculate how this could have affected our results. Finally, despite the long data collection period, 125 participants (21%) did not provide longitudinal data for either speech task, highlighting the need to understand facilitators and barriers of remote speech collection (Brederoo et al., 2021; Dineley et al., 2021). Future work will include in depth analysis to understand if there are specific groups who are less willing to provide speech data.

To conclude, as our findings are based on multilingual data they represent a considerable step-change in the usefulness of speech as a digital phenotype of MDD. Importantly, as the identified associations were observed with the scripted task, they are potentially more favourable to participants in future studies from a privacy perspective. Combining the results of this study with previously presented analyses (6-10), there is strong evidence to support the use of speech-rate measures as digital phenotypes of MDD in larger scale research projects. A range of future works are planned to build on the analysis presented in this paper. We plan to re-examine the relationships found under our explanatory paradigm with predictive models (Shmueli, 2010) and expand the analysis to consider symptom severity (Fara et al., 2022). We also plan to assess speech feature stability and minimal detectable change in speech features with respect to depression and individual symptom severity (Kothare et al., 2022). These future works will increase our understanding in how changes in depression severity affects speech production and by association, the underlying mechanisms it represents.

Code Availability

Please contact the lead corresponding author for a copy of the feature extraction code used in this work.

Data Availability

Due to the confidential nature of speech data, we are unable to make this 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 corresponding author for details.

Acknowledgement:

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 European Union'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 European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding body have not been involved in the design of the study, the collection or analysis of data, or the interpretation of data.

Participant recruitment in Amsterdam was partially accomplished through Hersenonderzoek.nl, a Dutch online registry that facilitates participant recruitment for neuroscience studies (https://hersenonderzoek.nl/). Hersenonderzoek.nl is funded by ZonMw-Memorabel (project no 73305095003), a project in the context of the Dutch Deltaplan Dementie, Gieskes-Strijbis Foundation, the Alzheimer's Society in the Netherlands and Brain Foundation Netherlands Participants in Spain were recruited through the following institutions: Parc Sanitari Sant Joan de Déu network of mental health services (Barcelona); Institut Català de la Salut primary care services (Barcelona); Institut Pere Mata-Mental Health Care (Tarrassa); Hospital Clínico San Carlos (Madrid). This paper represents independent research part 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, the NIHR or the Department of Health and Social Care.

We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design. This research was reviewed by a team with experience of mental health problems and their carers who have been specially trained to advise on research proposals and documentation through the Feasibility and Acceptability Support Team for Researchers (FAST-R): a free, confidential service in England provided by the National Institute for Health Research Maudsley Biomedical Research Centre via King's College London and South London and Maudsley NHS Foundation Trust.

We thank all GLAD Study volunteers for their participation, and gratefully acknowledge the NIHR BioResource, NIHR BioResource centres, NHS Trusts and staff for their contribution. We also acknowledge NIHR BRC, King's College London, South London and Maudsley NHS Trust and King's Health Partners. We thank the National Institute for Health Research, NHS Blood

and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Programme.

We thank our colleagues both within the RADAR-CNS consortium and across all involved institutions for their contribution to the development of this protocol. We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design.

References

Abbas, A., Sauder, C., Yadav, V., Koesmahargyo, V., Aghjayan, A., Marecki, S., Evans, M., Galatzer-Levy, I.R., 2021. Remote Digital Measurement of Facial and Vocal Markers of Major Depressive Disorder Severity and Treatment Response: A Pilot Study. Frontiers in Digital Health 3, 610006.

Alghowinem, S., Goecke, R., Epps, J., Wagner, M., Cohn, J.F., 2016. Cross-cultural depression recognition from vocal biomarkers., in: Proceedings INTERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association. ISCA, San Francisco, CA, USA, pp. 1943–1947.

Ambikairajah, E., Li, H., Wang, L., Yin, B., Sethu, V., 2011. Language Identification: A Tutorial. IEEE Circuits and Systems Magazine 11, 82–108.

Bagiella, E., Sloan, R.P., Heitjan, D.F., 2000. Mixed-effects models in psychophysiology. Psychophysiology 37, 13–20.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48.

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot International 5, 341–345.

Bot, M., Middeldorp, C.M., de Geus, E.J.C., Lau, H.M., Sinke, M., van Nieuwenhuizen, B., Smit, J.H., Boomsma, D.I., Penninx, B.W.J.H., 2017. Validity of LIDAS (LIfetime Depression Assessment Self-report): a self-report online assessment of lifetime major depressive disorder. Psychological Medicine 47, 279–289.

Brederoo, S.G., Nadema, F.G., Goedhart, F.G., Voppel, A.E., De Boer, J.N., Wouts, J., Koops, S., Sommer, I.E.C., 2021. Implementation of automatic speech analysis for early detection of psychiatric symptoms: What do patients want? Journal of Psychiatric Research 142, 299–301.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., Snyder, P.J., 2004. Voice acoustical measurement of the severity of major depression. Brain and cognition 56, 30–35.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., Quatieri, T.F., 2015. A review of depression and suicide risk assessment using speech analysis. Speech Communication 71, 10–49.

de Jong, N.H., Pacilly, J., Heeren, W., 2021. PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically. Assessment in Education: Principles, Policy & Practice 28, 456–476.

Dineley, J., Carr, E., Matcham, F., Downs, J., Dobson, R., Quatieri, T.F., Cummins, N., 2023. Towards robust paralinguistic assessment for real-world mobile health (mHealth) monitoring: an initial study of reverberation effects on speech. accepted for publication at INTERSPEECH 2023, Dublin, Ireland.

Dineley, J., Lavelle, G., Leightley, D., Matcham, F., Siddi, S., Peñarrubia-María, M.T., White, K.M., Ivan, A., Oetzmann, C., Simblett, S., Dawe-Lane, E., Bruce, S., Stahl, D., Ranjan, Y., Rashid, Z., Conde, P., Folarin, A.A., Haro, J.M., Wykes, T., Dobson, R.J.B., Narayan, V.A., Hotopf, M., Schuller, B.W., Cummins, N., The RADAR-CNS Consortium, 2021. Remote Smartphone-Based Speech Collection: Acceptance and Barriers in Individuals with Major Depressive Disorder. INTERSPEECH 2021.

Fara, S., Goria, S., Molimpakis, E., Cummins, N., 2022. 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: Proceedings INTERSPEECH 2022, 23rd Annual Conference of the International Speech Communication Association. ISCA, Incheon, Korea, pp. 1911–1915.

Franklin, J., 2006. The Sheer Audacity: How to Get More, in Less Time, from the Audacity Digital Audio Editing Software, in: 2006 IEEE International Professional Communication Conference. IEEE, Saratoga Springs, NY, USA, pp. 92–105.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., Morency, L.-P., 2014. The Distress Analysis Interview Corpus of human and computer interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation.

Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N., Altman, D.G., 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology 31, 337–350.

Henry, J.D., Crawford, J.R., 2005. A Meta-Analytic Review of Verbal Fluency Deficits in Depression. Journal of Clinical and Experimental Neuropsychology 27, 78–101.

Ibrahim, J.G., Molenberghs, G., 2009. Missing data methods in longitudinal studies: a review. Test 18, 1–43.

International Phonetic Association, 1999. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge University Press.

Jadoul, Y., Thompson, B., de Boer, B., 2018. Introducing Parselmouth: A Python interface to Praat. Journal of Phonetics 71, 1–15.

Jefferson, A.L., Gibbons, L.E., Rentz, D.M., Carvalho, J.O., Manly, J., Bennett, D.A., Jones, R.N., 2011. A life course model of cognitive activities, socioeconomic status, education, reading ability, and cognition. Journal of the American Geriatrics Society 59, 1403–1411.

Kiss, G., Vicsi, K., 2017. Mono- and multi-lingual depression prediction based on speech processing. International Journal of Speech Technology 20, 919–935.

Kothare, H., Neumann, M., Liscombe, J., Roesler, O., Burke, W., Exner, A., Snyder, S., Cornish, A., Habberstad, D., Pautler, D., 2022. Statistical and clinical utility of multimodal dialogue-based speech and facial metrics for Parkinson's disease assessment, in: Proceedings INTERSPEECH 2022, 23rd Annual Conference of the International Speech Communication Association. ISCA, Incheon, Korea, pp. 3658–3662.

Kroenke, K., Strine, T.W., Spitzer, R.L., Williams, J.B.W., Berry, J.T., Mokdad, A.H., 2009. The PHQ-8 as a measure of current depression in the general population. Journal of Affective Disorders 114, 163–173.

Low, D.M., Bentley, K.H., Ghosh, S.S., 2020. Automated assessment of psychiatric disorders using speech: A systematic review. Laryngoscope Investigative Otolaryngology 5, 96–116.

Löwe, B., Kroenke, K., Herzog, W., Grafe, K., 2004. Measuring depression outcome with a brief self-report instrument: sensitivity to change of the Patient Health Questionnaire (PHQ-9). Journal of Affective Disorders 81, 61–66.

Matcham, F., Barattieri di San Pietro, C., Bulgari, V., de Girolamo, G., Dobson, R., Eriksson, H., Folarin, A.A., Haro, J.M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V, Mohr, D.C., Myin-Germeys, I., Narayan, V., BWJH, P., Ranjan, Y., Rashid, Z, Rintala, A., Siddi, S., Simblett, S.K., Wykes, T., Hotopf, M., DiFrancesco, S., White, K., Ivan, A., Polhemus, A., Ferrao, J., Ringkjøbing-Elema, M., Nobilia, F., Viechtbauer, W., Peelen, S., Rashid, Zulqarnain, Boere, J., Cummins, N., Meyer, N., Consortium, on behalf of the R.-C., 2019. Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. BMC Psychiatry 19, e72.

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K.M., Annas, P., de Girolamo, G., Difrancesco, S., Haro, J.M., Horsfall, M., Ivan, A., Lavelle, G., Li, Q., Lombardini, F., Mohr, D.C., Narayan, V., Oetzmann, C., Penninx, B.W.J.H., Bruce, S., Nica, R., Simblett, S.K., Wykes, T., Brasen, J.C., Myin-Germeys, I., Rintala, A., Conde, P., Dobson, R.J.B., Folarin,

A.A., Stewart, C., Ranjan, Y., Rashid, Z., Cummins, N., Manyakov, N. V, Vairavan, S., Hotopf, M., 2022. Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. BMC Psychiatry 22, 136.

Mitra, V., Shriberg, E., Vergyri, D., Knoth, B., Salomon, R.M., 2015. Cross-corpus depression prediction from speech, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, Brisbane, Australia, pp. 4769–4773.

Moskvina, A., 2013. Comparative Study of English, Dutch and German Prosodic Features (Fundamental Frequency and Intensity) as Means of Speech, in: Železný, M., Habernal, I., Ronzhin, A. (Eds.), International Conference on Speech and Computer. Springer International Publishing, Pilsen, Czech Republic, pp. 86–91.

Mundt, J.C., Snyder, P.J., Cannizzaro, M.S., Chappie, K., Geralts, D.S., 2007. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. Journal of Neurolinguistics 20, 50–64.

Mundt, J.C., Vogel, A.P., Feltner, D.E., Lenderking, W.R., 2012. Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. Biological Psychiatry 72, 580–587.

Ranjan, Y., Rashid, Z., Stewart, C., Conde, P., Begale, M., Verbeeck, D., Boettcher, S., Dobson, R., Folarin, A., 2019. RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. JMIR Mhealth Uhealth 7, e11734.

Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., Pantic, M., 2019. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition, in: Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop. ACM, Nice, France, pp. 3–12.

Rush, A.J., Carmody, T., Reimitz, P.-E., 2000. 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 9, 45–59.

Scherer, S., Lucas, G.M., Gratch, J., Skip Rizzo, A., Morency, L., 2016. Self-Reported Symptoms of Depression and PTSD Are Associated with Reduced Vowel Space in Screening Interviews. IEEE Transactions on Affective Computing 7, 59–73.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2013. Paralinguistics in speech and language—State-of-the-art and the challenge. Computer

Speech & Language 27, 4–39.

Schwoebel, J.W., Schwartz, J., Warrenburg, L.A., Brown, R., Awasthi, A., New, A., Butler, M., Moss, M., Pissadaki, E.K., 2021. A longitudinal normative dataset and protocol for speech and language biomarker research. medrxiv. https://doi.org/10.1101/2021.08.16.21262125

Shmueli, G., 2010. To Explain or to Predict? Statistical Science 25, 289–310.

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

Wagner, P., Trouvain, J., Zimmerer, F., 2015. In defense of stylistic diversity in speech research. Journal of Phonetics 48, 1–12.

Werner, R., Trouvain, J., Moebius, B., 2022. Optionality and variability of speech pauses in read speech across languages and rates, in: Speech Prosody. ISCA, Lisbon, Portugal, pp. 312–316.

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K., Fujita, T., Mimura, M., Kishimoto, T., 2020. Using speech recognition technology to investigate the association between timing-related speech features and depression severity. PLOS ONE 15, e0238726.

Yang, Y., Fairbairn, C., Cohn, J.F., 2013. Detecting Depression Severity from Vocal Prosody. IEEE Transactions on Affective Computing 4, 142–150.

Figure 1: Overview of the key steps in the data processing pipeline. Participants in RADAR-MDD were invited to record speech samples once every two weeks at the same time they completed an 8-item Patient Health Questionnaire (PHQ-8) scale (25) to assess their depression severity. All audio samples collected were then decrypted into WAV files. Feature extraction was undertaken using Parselmouth (22). We then used Linear Mixed Effect models to estimate the association between the speech features and PHQ-8 scores.

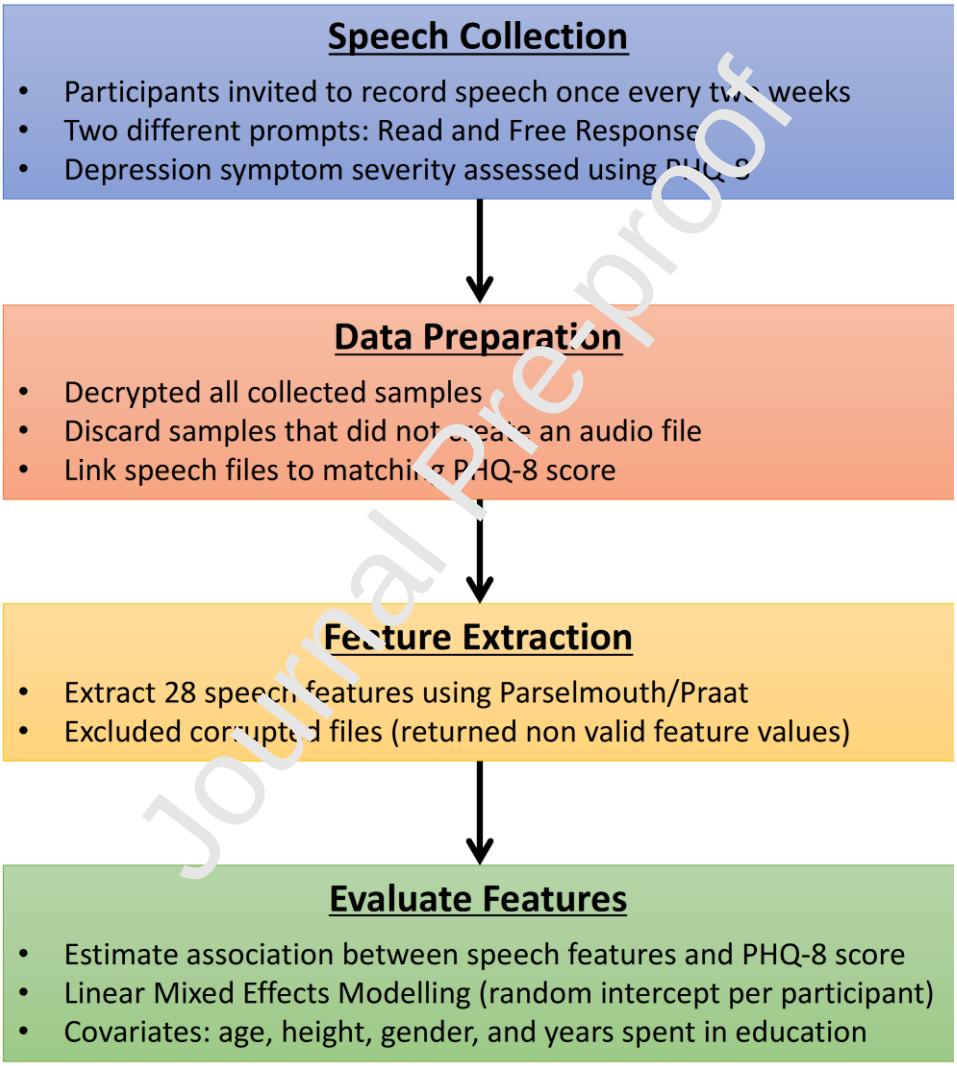


Figure 2: Breakdown of number of participants supplying longitudinal speech data in RADAR-MDD. *Never Attempt* denotes the number participants who never attempted that task; *Corrupt Files Only*, are participants who attempted the task, but created only corrupt files; *Attempted Once*, are participants provided only one file; and *No Analysable Data* are participants excluded as their files failed the paper's inclusion of being both over 2 seconds in length and returning usable Parselmouth features.

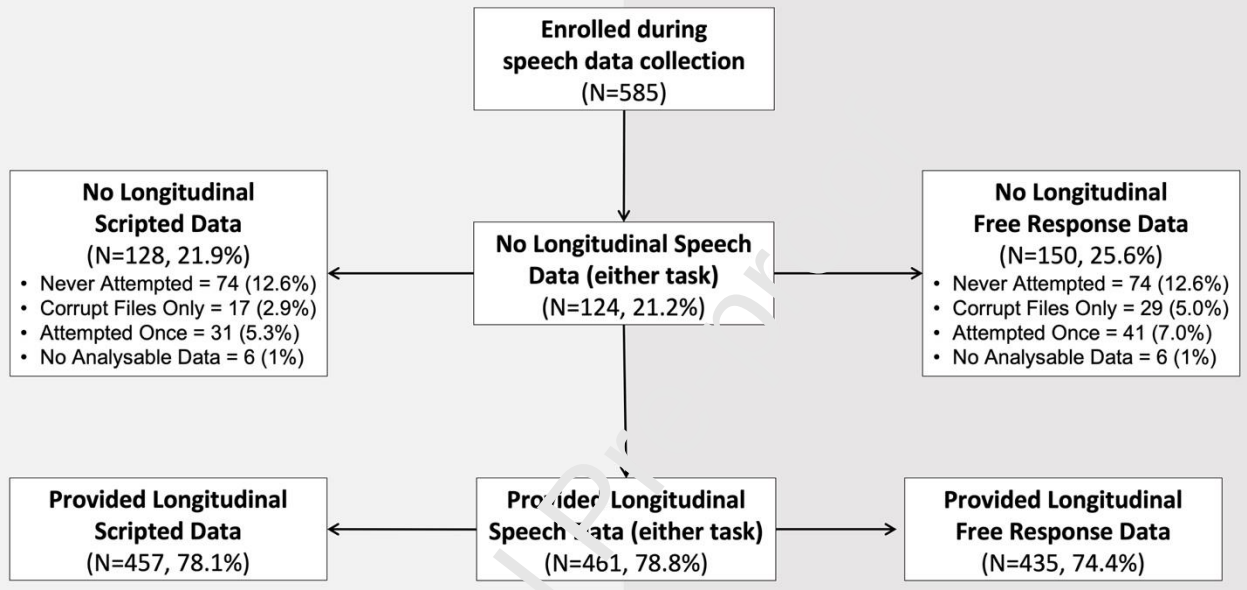


Figure 3: Association of speech features with PHQ-8 score for the scripted task (n=457 individuals; 7356 observations). *Notes.* Points represent the difference in PHQ-8 per 1 SD difference in each feature. 95% confidence intervals obtained using bootstrap resampling (1000 samples)

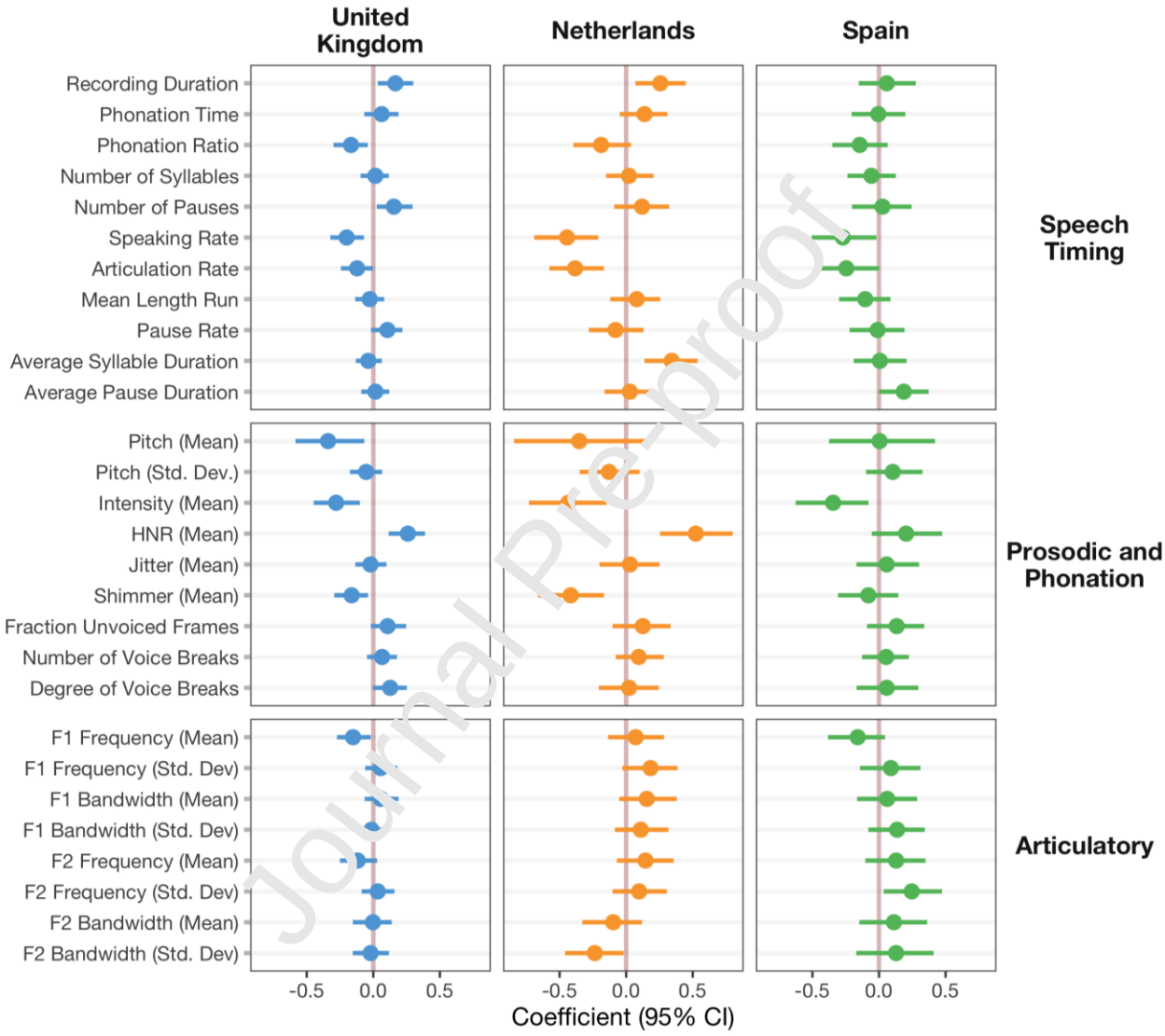
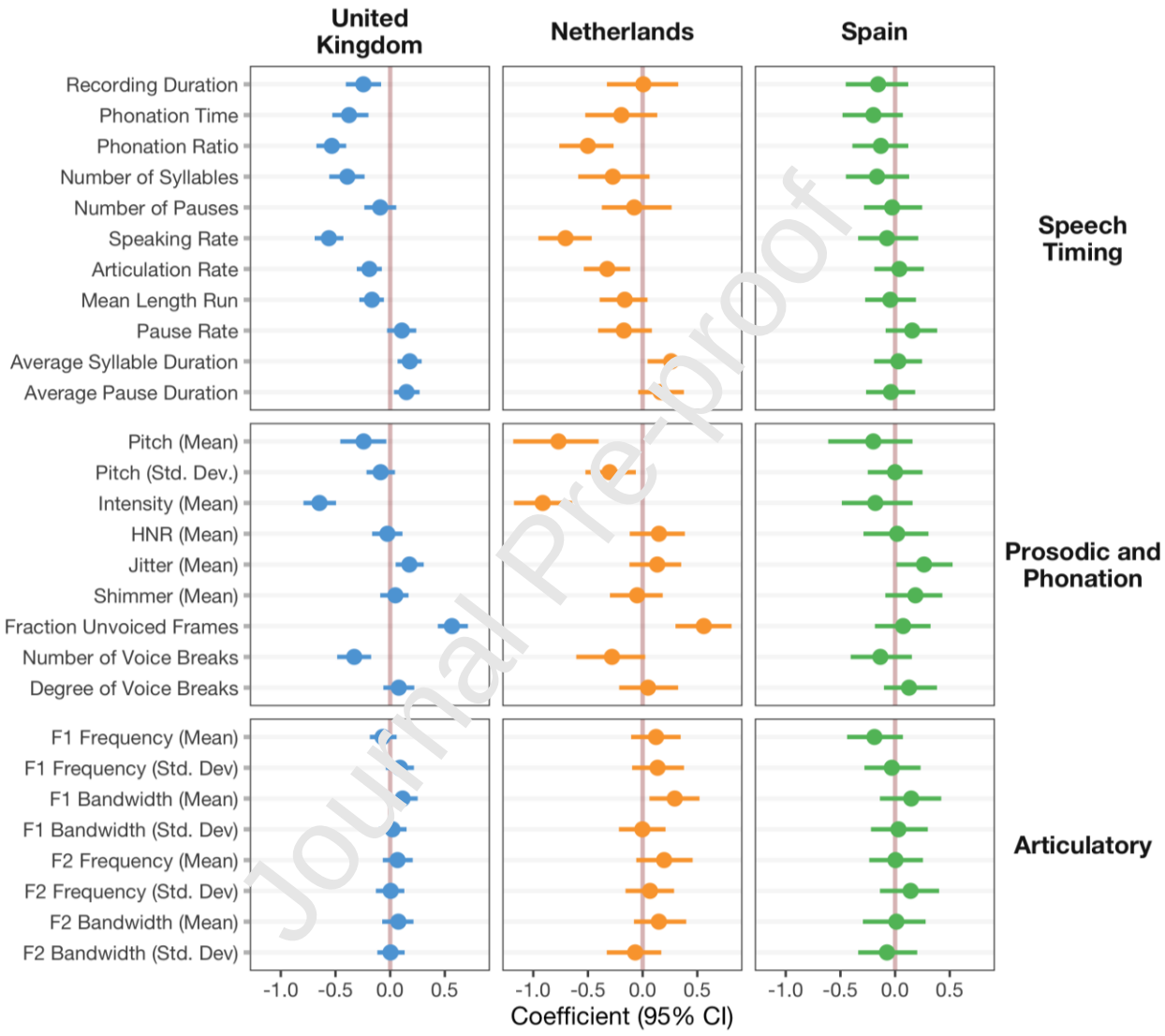


Figure 4: Association of speech features with PHQ-8 score for the free response task (n=435 individuals; 6106 observations) *Notes.* Points represent the difference in PHQ-8 per 1 SD difference in each feature. 95% confidence intervals obtained using bootstrap resampling (1000 samples).



Author Contributions

NC: Conceptualization, Methodology, Formal Analysis, Writing Original Draft; JD: Conceptualization, Methodology, Data Curation, Writing – Original Draft; PC: Data Curation, Software; FM: Conceptualization, Methodology, Project administration, Writing – Review; SS and FL: Data Curation, Project administration, Writing – Review; EC: Methodology, Validation, Writing – Review; GL, DL, KW, CO: Data Curation, Writing – Review; ELC: Investigation; SS, SB: Conceptualization, Writing – Review; JMH: Project administration, Funding acquisition; BP: Writing – Review, YR, ZR, CS: Software, AF: Software, Project administration, Writing – Review; RB: Project administration, Writing – Review; BS: Project administration, Funding acquisition, Writing – Review; TW: Conceptualization, Project administration, Funding acquisition, Writing – Review; SV: Methodology, Project administration, Writing – Review; RB: Methodology, Software, Project administration, Funding acquisition, Writing – Review; VN: Project administration, Funding acquisition, Writing – Review; MH: Methodology, Project administration, Funding acquisition, Writing – Review.

Declaration of ompeting interest

The Authors declare no Competing Financial or Non-Financial Interests.

Highlights

- Collected of a unique multilingual clinical speech dataset.
- Identified three multilingual speech phenotypes of MDD.
- Pausing not universally impacted by MDD across the three languages.
- Slowing of speech is most likely due to psychomotor impairments.