A Generalizable Speech Emotion Recognition Model Reveals Depression and Remission

Lasse Hansen[1,2,3,4], Yan-Ping Zhang[4], Detlef Wolf[4], Konstantinos Sechidis[5], Nicolai Ladegaard[1,2], Riccardo Fusaroli [6,7]

[1] *Department of Clinical Medicine, Aarhus University, Aarhus, Denmark*

[2] *Department of Affective Disorders, Aarhus University Hospital - Psychiatry, Aarhus, Denmark*

[3] *Center for Humanities Computing Aarhus, Aarhus University, Aarhus, Denmark*

[4] *Roche Pharmaceutical Research & Early Development Informatics, Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd., Basel, 4070, Switzerland*

[5] *Advanced Methodology and Data Science, Novartis Pharma AG, Basel, Switzerland*

[6] *Cognitive Science, School of Communication and Culture, Aarhus University, Aarhus Denmark*

[7] *The Interacting Minds Centre, Aarhus University, Aarhus Denmark*

Abstract

Objective

Affective disorders have long been associated with atypical voice patterns, however, current work on automated voice analysis often suffers from small sample sizes and untested generalizability. This study investigated a generalizable approach to aid clinical evaluation of depression and remission from voice.

Methods

A Mixture-of-Experts machine learning model was trained to infer happy/sad emotional state using three publicly available emotional speech corpora. We examined the model's predictive ability to classify the presence of depression on Danish speaking healthy controls (N = 42), patients with first-episode major depressive disorder (MDD) (N = 40), and the same patients in remission (N = 25) based on recorded clinical interviews. The model was evaluated on raw data, data cleaned for background noise, and speaker diarized data.

Results

The model showed reliable separation between healthy controls and depressed patients at the first visit, obtaining an AUC of 0.71. Further, we observed a reliable treatment effect in the depression group, with speech from patients in remission being indistinguishable from that of the control group. Model predictions were stable throughout the interview, suggesting that as little as 20-30 seconds of speech is enough to accurately screen a patient. Background noise (but not speaker diarization) heavily impacted predictions, suggesting that a controlled environment and consistent preprocessing pipelines are crucial for correct characterizations.

Conclusion

A generalizable speech emotion recognition model can effectively reveal changes in speaker depressive states before and after treatment in patients with MDD. Data collection settings and data cleaning are crucial when considering automated voice analysis for clinical purposes.

Keywords

Depression, Machine Learning, Speech Acoustics, Transfer learning, Speech Emotion Recognition

Significant outcomes

- Using a speech emotion recognition model trained on other languages, we predicted the presence of MDD with an AUC of 0.71.
- The speech emotion recognition model could accurately detect changes in voice after patients achieved remission from MDD.
- Preprocessing steps, particularly background noise removal, greatly influenced classification performance.

Limitations

- No data from non-remitters, meaning that changes to voice for that group could not be assessed.
- It is unclear how well the model would generalize beyond Germanic languages.

Data availability statement

Due to the nature of the data (autobiographical interviews in a clinical population), the recordings of the participants cannot be shared publicly. The aggregated model predictions and code used to run the analyses is available at https://github.com/HLasse/SERDepressionDetection.

Introduction

Major Depressive Disorder (MDD) is a mental disorder affecting more than 163 million people worldwide [1], and comprises symptoms related to abnormalities in mood, cognitive ability, and psychomotor function [2]. Current approaches to screening and monitoring of symptoms of depression primarily depend on self-reports, and are therefore often confounded by issues such as underestimating symptom severity, recency effects, and acquiescence [3–5]. More efficient and objective screening and monitoring of depressive clinical features has the potential for relieving the disease burden significantly by scaffolding more timely and personalized interventions or by providing a measure of efficacy of the current treatment [6].

Using voice as a marker for depression is appealing as analysis and assessment can be done cheaply, remotely, and non-invasively. Depressed speech is characterized by a slow speaking rate, reduced inflection and prosody, and low volume [7,8]. Correspondingly, a range of acoustic features have been identified as predictive of depression, ranging from changes to fundamental frequency (pitch) to more abstract spectral representations of speech [9]. Multiple studies have used these acoustic features in machine learning models to discriminate depressed patients and healthy controls. Classification performance is highly varying, and accuracies between chance level and up to 94% are found in the literature depending on model and feature choice, dataset size, language, and validation method [9,10]. However promising, it is still unclear how well these algorithms would actually perform across a broader variety of clinical contexts.

For instance, a recent systematic review found that most of the studies predicting depression from voice did not fully evaluate the generalizability of these results to new data, i.e., they measured performance on the training sample or in a cross-validated fashion but not on held-out validation sets [11]. This is probably motivated by small datasets being endemic to the field: high-quality clinical data is time-consuming and expensive to collect, and problematic to share. For instance, recent systematic reviews found that in related neuropsychiatric fields the median number of patients involved in studies of vocal markers were below 20 [12–14]. Therefore, reducing the size of already small training samples by creating held-out data might be seen as problematic. However, even cross-validated performance has been shown to be unreliable on small or improperly nested datasets [15,16]. It thus remains unclear how well these algorithms would perform on new data collected in similar conditions.

Further, deploying automated voice analysis in clinical settings involves analyzing recordings of very heterogeneous patient populations collected in very heterogeneous physical settings with a wide variety of equipment. In other words, it involves generalizing the trained algorithms to data that are potentially quite different from the original training set. We already know acoustic

feature extraction from audio to be strongly influenced by background noise and recording conditions. For instance, one study has found consistent differences in the acoustic environments between their diagnostic groups, which could predict group membership with nearly the same performance as models based on acoustic features [17]. Therefore, it might not be enough to validate an algorithm even on held-out data collected in similar conditions to the training data, as this might produce overly optimistic evaluations of performance. It is clear that before clinical implementation is possible, generalizable models that perform well on data from other sources are needed and validation practices have to include more heterogeneity.

To overcome these limitations of small training data and lack of testing for external validity, we train a speech emotion recognition (SER) model and directly apply it to depression detection. Using a SER model for predicting depression is motivated by the findings that prosodic expressions of emotions are inhibited in depression and that experimental studies have found positive effects of adding SER to depression detection models [18,19]. Datasets for SER are vastly more abundant, varied, and of higher quality than for depression assessment which is likely to produce more robust models. Further, by training a model solely for SER we are able to set aside our entire dataset of depressed patients and healthy controls for external validation thus providing a realistic view of generalizability.

Aims of the study

The aims of the study are three-fold:

- *Aim 1*: to investigate the feasibility of using a generalizable emotion recognition model to directly predict depression,
- *Aim 2*: to assess the stability of these predictions over time as well as changes following remission from Major Depressive Disorder, and
- *Aim 3*: to quantify the effect of preprocessing steps such as background noise removal, speaker diarization (removal of speech from the interviewer), and time-window size on the quality and consistency of model predictions.

Materials

Speech Emotion Recognition Corpora and Model

The SER model was trained following Sechidis et al. [20] on the public CREMA-D [21] , RAVDESS [22], and EMO-DB [23] datasets, all of which consist of recordings of sentences spoken by professional actors portraying different emotions. CREMA-D and RAVDESS include American English speech and EMO-DB German speech. A gradient boosted decision tree model was trained on each dataset separately to predict the probability of sounding happy or sad using Catboost [24] and combined in a Mixture of Experts (MoE) architecture [25] for ensemble prediction. MoE is a way of combining the predictions of multiple models (*ensembling)*, which weights model predictions based on the distance of the input data to the constituent models' training data. This way of ensembling allows one to train specialized models for e.g. different languages or recording conditions, and combine their knowledge at inference time hereby improving generalizability [26]. In a previous study, using different SER corpora for external validation, the MoE model outperformed all constituent models as well as a single model trained on the pooled data from all the different corpora [20]. Further details on feature extraction, training, and validation of the SER model are provided in the Supplementary Material, hence the following sections in *Materials* and *Methods* relate to the depression corpus.

Depressed Speech Corpus

The dataset used for depression assessment was collected at Aarhus University Hospital to investigate changes in social cognition in first-episode depression, and consists of 42 patients with first-episode MDD (two patients discarded due to missing data) and 42 healthy controls pairwise matched on gender, age, and educational level [27,28]. All participants were native speakers of Danish and met the following inclusion criteria: 1) first-episode major depression was the primary diagnosis, 2) the severity of depression was moderate to severe as measured by the 17-item Hamilton Rating Scale for Depression (HamD-17) [29], 3) patients were psychotropic drug-naïve. Physiological effects from psychotropic drugs can lead to changes in the voice [30], hence the inclusion of only psychotropic drug-naïve patients is crucial for the present study. Patients with head trauma, neurological illness, or substance use disorders were not permitted to the study. Exclusion criteria for healthy controls were the same as for depressed patients.

The dataset consists of audio recordings of the Indiana Psychiatric Illness Interview [31] conducted by a trained psychologist separately with each of the participants in Danish. Participants were asked to tell their life story in the first part of the interview, and in the second part to either reflect on their mental illness or on an emotionally distressing experience they have had in the last 2 years, depending on whether they were in the depression or control group. The interviews lasted between 20 and 50 minutes.

After the interview, the depressed patients underwent pharmacotherapy and individual psychotherapy. Those who entered remission within 6 months, defined as a HamD-17 score ≤ 7, were re-assessed with the same interview, along with the control group. As such, our dataset contains recordings of interviews with healthy controls at two timepoints six months apart (N=42/25), as well as interviews with 40 depressed individuals and a follow-up interview after six months with the subset who entered remission (N=25). Unfortunately, due to the focus of the original study, patients not in remission were not invited to the follow-up interview.

Methods

Table 1: Demographic characteristics

Diagnosis	Gender	N	Hamilton mean	Hamilton IQR	Age mean	Age Range
Visit 1						
Controls	Female	33	1.6	2.0	32.3	18.1-62.1
Controls	Male	9	1.8	2.0	36.3	22.4-54.3
Depression	Female	31	22.1	5.5	32.0	18.8-62.6
Depression	Male	9	21.8	4.0	34.0	21.1-53.9
Visit 2						
Controls	Female	21	1.5	3.0	37.1	19.7-62.1
Controls	Male	4	2.0	2.0	38.4	30.3-54.3
Depression	Female	20	22.0	6.2	29.9	18.8-62.6
Depression	Male	5	21.6	3.0	34.9	21.1-53.9

Data preprocessing

Background room noise, reverberation, and hum were removed from the audio recordings using iZotope RX 6 Elements [32]. Long-term average spectra for each recording were inspected for possible noise artefacts and further cleaned if any were found.

To ensure that only voice segments from patients and controls were included, all interviews were manually segmented to only contain audio from the interviewee. We further removed all segments of audio without voice activity defined by a loudness threshold of -40dB and a minimum duration of 100 ms.

Feature Extraction

From each audio recording, we extracted 13 mel-frequency cepstral coefficients (MFCC 0-12) with a window length of 25 ms and step size of 10 ms. Mel-frequency cepstral coefficients (MFCCs) have been widely used in both speaker recognition [33], SER [34], and depression detection [35], and have several desirable properties such as being independent of the energy of the acoustic signal and robustness across genders [36,37]. MFCCs represent movements of the vocal tract and

are designed to mimic how the human ear perceives sounds by having high resolution in the lower frequencies and less in higher frequencies [38].

The zeroth MFCC was discarded as it represents the average energy of the signal. Though energy is often found to be a reliable vocal marker of depression, it is easily confounded by inconsistent recording conditions and its inclusion might therefore reduce the generalizability of the model.

The MFCC features were summarized in windows of different sizes (2, 5, 10, 15, 20, 25, and 30 seconds of speech as well as the whole recording) with 11 descriptive statistics: mean, variance, kurtosis, skewness, mode, IQR, percentiles 10th, 25th, 50th, 75th, and 90th. Summarization often results in better predictive performance than using raw features [10,39], and is far less computationally expensive. Windows of 30 seconds of speech were used for the main analyses.

Though other acoustic features such as pitch and energy are often found to be predictive of depression [9], we chose to focus on MFCCs as they have previously shown good predictive performance, can be robustly extracted, and are independent of energy and gender.

Model Development

In order to assess whether SER predictions would discriminate depressed patients from controls (*aim 1*) and whether these predictions would be stable over time (*aim 2*), we built a Bayesian multilevel mixture model of the predictions. The SER model was applied to each speech sample from the depression corpus to obtain the log odds of sounding *happy* (1) or *sad* (0) for each time window, for each participant, at each visit. Given the heterogeneity of the symptom manifestations in depression [40], and only some patients entering remission, we adopted a mixture model of two gaussians to model the log-odds of sounding happy. In the mixture model, two Gaussian distributions best describing the data are estimated, and the probability of each speech sample to belong to one or the other can vary by participant - nested by diagnostic group as recommended in Valton et al. [41] to avoid pooling across groups - and visit. This also accounted for the presence of repeated measures, i.e., multiple predictions/time windows per participant. An interaction effect was used to assess the expectation that the depressed group should sound happier at the second visit as only those in remission were re-interviewed, while the voice of the control group should remain stable over time[1].

To estimate the performance of the SER model for discriminating between depressed patients and healthy controls, the area under the receiver operating characteristic curve (AUC) using

[1] See Appendix for further details on model building and assessment of the quality of the model fit.

the raw predictions was calculated. Further performance metrics were calculated using the decision threshold that optimized AUC.

To more directly investigate *aim 2*, we also trained a hierarchical Bayesian classification model (logistic regression) to predict prognosis (remission yes/no) based on the depressed patients' probability of sounding happy at the first visit. Baseline probabilities of sounding happy were modeled as varying by participant to account for repeated measures.

To assess the effects of the data preprocessing steps (*aim 3*), different datasets were created relying on the data from visit 1: a) data with background noise removal and without interviewer speech, b) data with background noise removal and including interviewer speech, c) data without background noise removal and including interviewer speech. How these preprocessing choices affected the difference in probability of sounding happy in patients and controls was assessed following the BEST approach suggested by Kruschke [42]. Further, the area under the receiver operating-characteristic curve for classifying depression from controls was calculated for different window sizes (2, 5, 10, 15, 20, 25, and 30 seconds, as well as no windowing) to investigate the impact of this choice.

All analyses were performed in RStudio 1.4.1103 [43], running R 4.0.3 [44] and relying on the brms v2.14.4 [45], pROC v1.17.0.1 [46] and Tidyverse v1.3.0 [47] packages.

Results

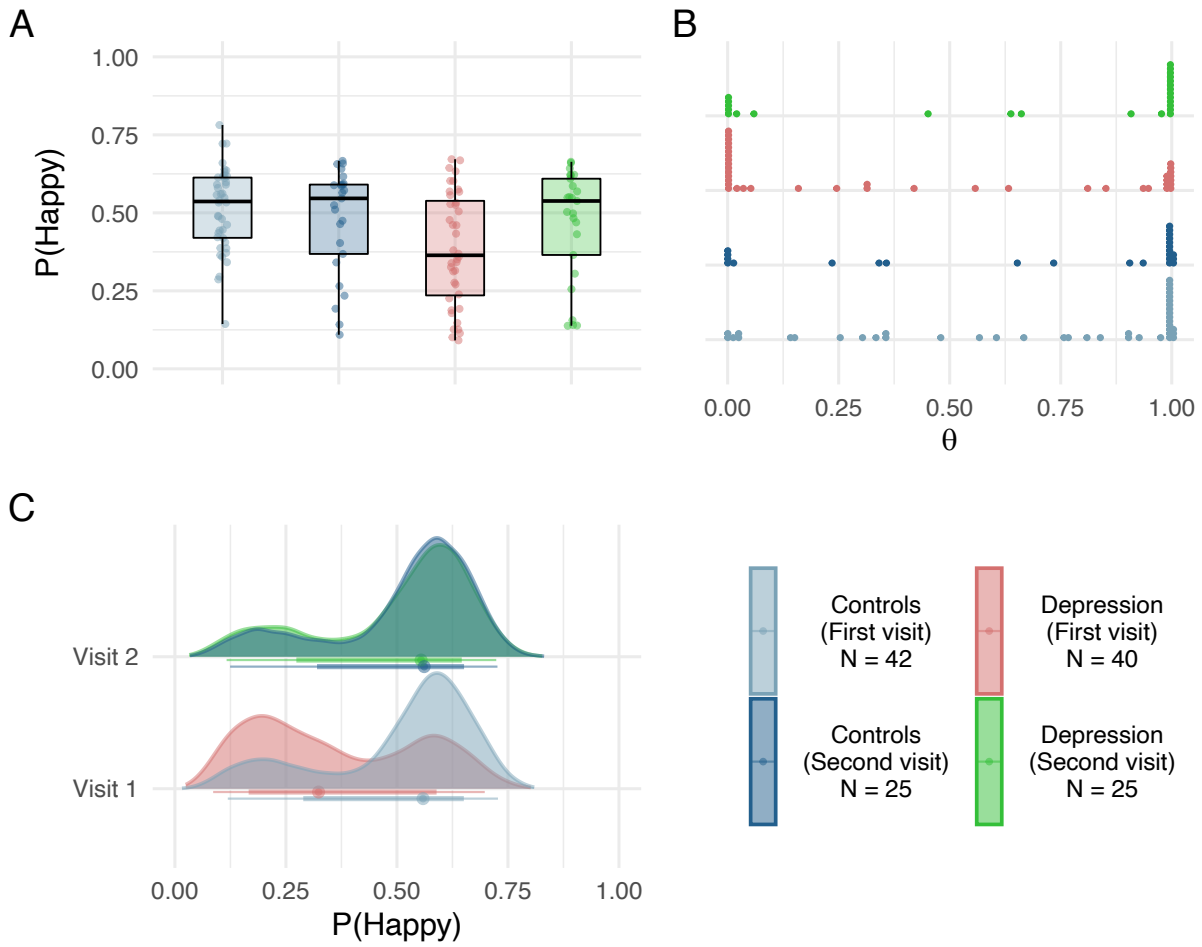

Figure 1: A) Distribution of mean predictions per individual at each visit. B) Probability of belonging to the 'happy' distribution by individual and diagnostic group. Dots show the mean mixing factor (theta) per individual. Most participants are in the extremes, which means that they are either exclusively in the happy (1) or sad distribution (0). C) Posterior probability of P(Happy) by diagnosis and visit from the multilevel Bayesian mixture model. Colored bands indicate 66% (thick line) and 95% (thin line) quantile intervals.

As shown in Figure 1a, the predicted probability of sounding happy is related to the diagnostic groups. The distribution of model predictions for the control group is stable across visits, and the distribution for patients in remission is similar to that of the control group. Predictions for the depressed group display larger variance, and are generally more sad sounding than controls.

The multilevel Bayesian mixture model supports these observations. As seen in Figure 1b, there are two nicely separated distributions with a population level difference in the probability of belonging to the happier distribution (Figure 1c). Most participants have mixing factors (theta)

very close to either 0 or 1, meaning that they are either exclusively in the *happy* or *sad* distribution. Mixing factors closer to 0.5 indicate greater uncertainty. The two Gaussian distributions identified by the mixture model are centered at 0.24 (sd=0.12) and 0.59 (sd=0.08). On the population level, depressed patients have a probability of sounding sad (theta) of 0.70 (95% CI: 0.38, 0.90), those in remission have a theta of 0.25 (95% CI: 0.07, 0.58), controls at visit 1 have a theta of 0.23 (95% CI: 0.10, 0.47), and controls at visit 2 have a theta of 0.22 (95% CI: 0.07, 0.58). The closer to 1, the larger a proportion of samples are from the *sad* distribution. The majority of depressed patients (approximately 70%) were thus identified as belonging to the *sad* distribution, whereas this number is only 22-25% for controls and patients in remission. The posterior distribution of patients in remission completely overlaps that of controls at both visits, which indicates that voice-based based symptoms of depression decrease following remission.

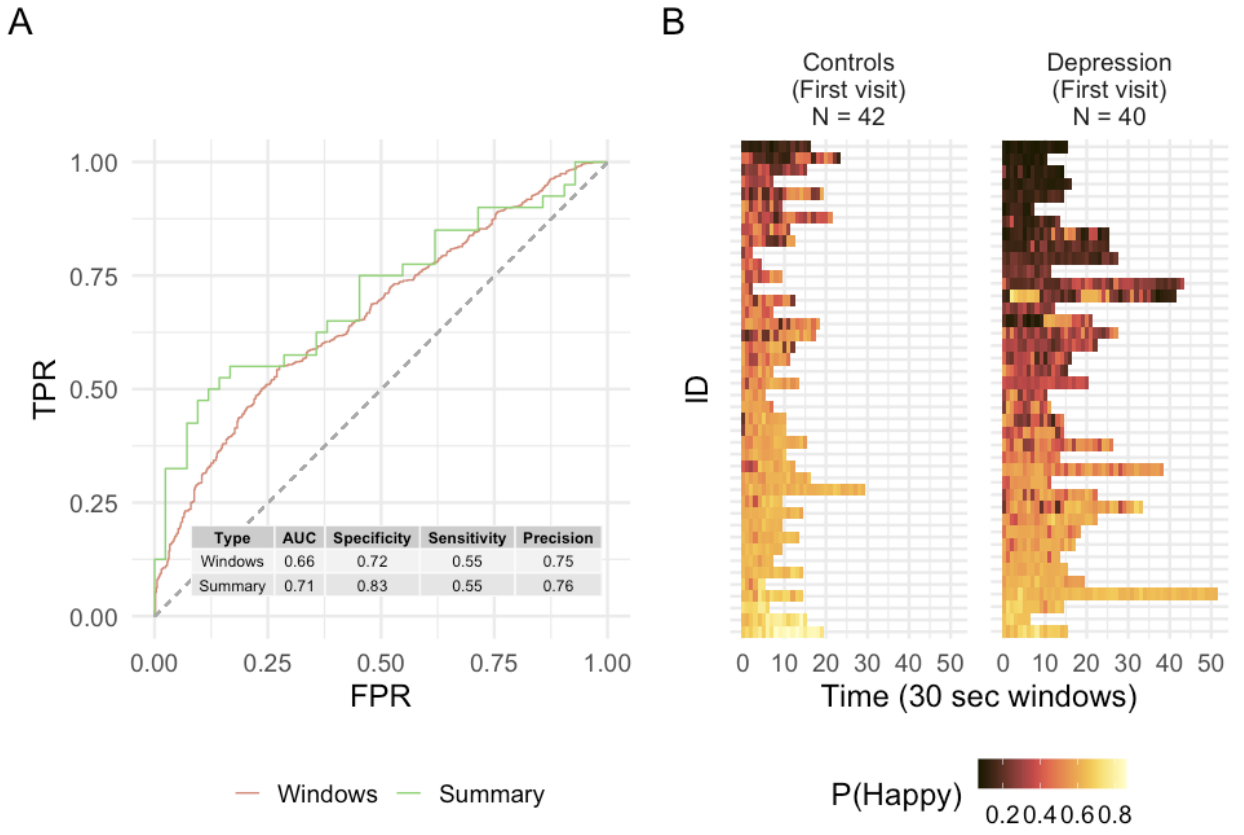

Figure 2: A) ROC curve displaying the separation between depression and healthy controls at visit 1. The red line is calculated on a per time-bin basis, i.e., the model is evaluated on each 30-second bin. The green line is calculated on a per participant basis, ie. predictions are summarized using the mean for each participant. B) Heatmap showing predictions for each individual throughout the interview at visit 1 sorted in ascending order. Each row corresponds to one participant. For most participants, predictions are highly stable over the course of the interview.

Predictions from the SER model can discriminate between healthy controls and depressed patients obtaining an AUC of 0.71 (95% CI: 0.59-0.82) using the mean prediction for each participant. Defining the classification threshold to the one that optimizes AUC (threshold=0.38) leads to a specificity of 0.83, sensitivity of 0.55, and positive predictive value of 0.76. In other words, given an optimized decision threshold the model is able to correctly classify 83% of the control group and 55% of the depressed group, while 76% of those predicted as being depressed are correctly classified. Figure 2a shows the ROC curve for this task using both 30 second time windows and the mean prediction per participant. Predictions seem to be stable over the course of the interview when using 30 second time windows, as seen in Figure 2b[2]. The time-windowing serves to smooth small changes in the participants' speech, and thus derives a time-independent emotional state. Further, predictions from the SER model seem sensitive to changes in the participant's depressive state: patients in remission are more likely to be identified as controls than during their earlier depressive state (72% vs. 45% using the optimal cutoff defined by AUC).

The prognosis model did not find any reliable differences in voice at visit 1 between those who subsequently entered remission and those who did not. Further details are reported in Supplementary Material Figure S4.

[2] See Supplementary Figure S5 and Supplementary Table S4 for individual and group level standard deviations.

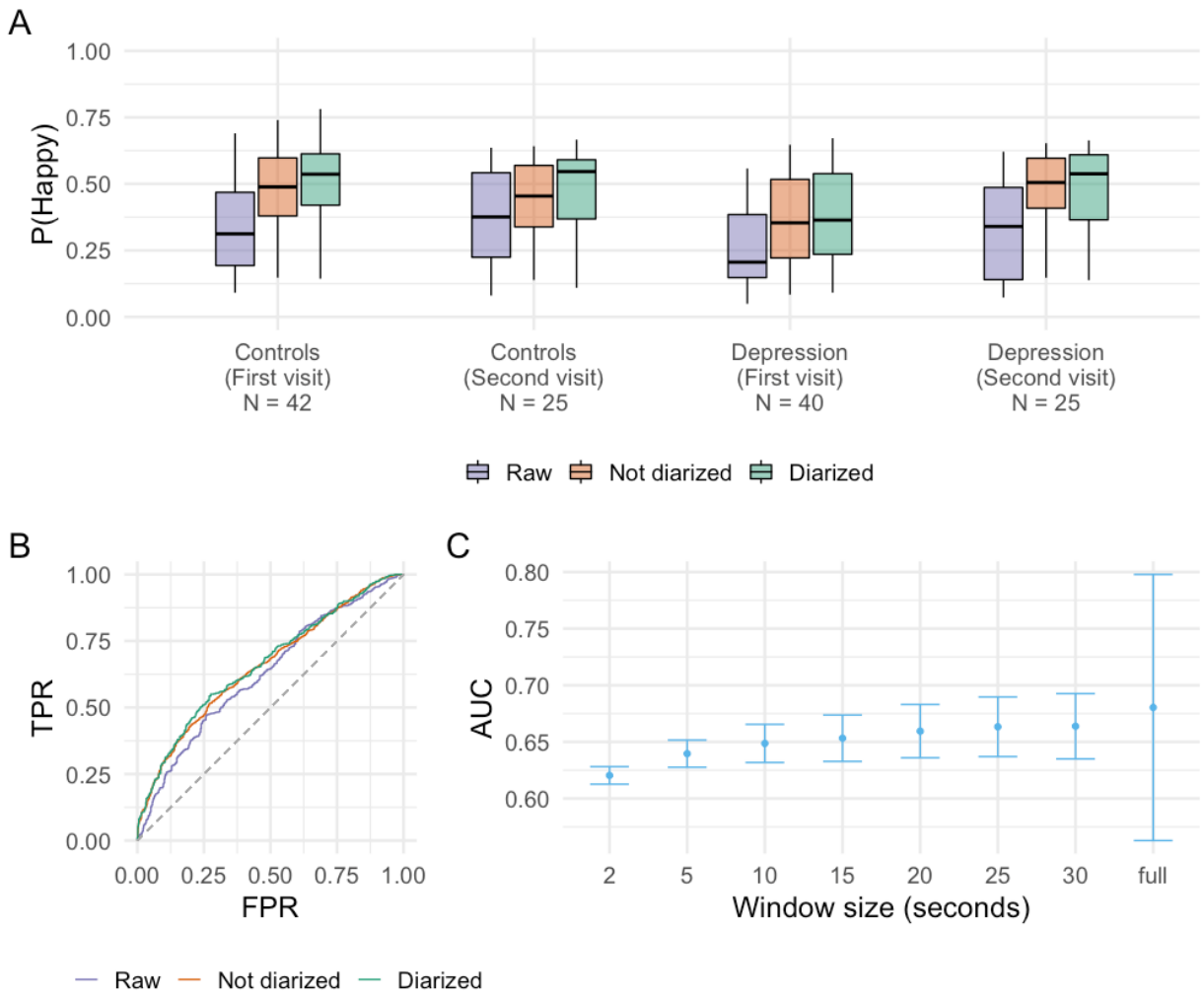

Figure 3: A) Difference in predictions with different preprocessing strategies. Both the diarized and non-diarized datasets were cleaned from background noise. B) ROC-curve for the task of predicting depression or control at visit 1 with different preprocessing strategies. Using 30 second windows. C) AUC for the task of predicting depression or control at visit 1 using different time-window sizes at a per window basis. Error bars display 95% bootstrapped confidence intervals. Note that the last bar presents higher uncertainty as it is evaluated on a per-participant basis where the others are evaluated on a per-time-window basis.

Figure 3a shows the distribution of predictions using the different preprocessing procedures. The magnitude of the effect of background noise removal and speaker diarization differs between the groups, highlighting the importance of these procedures for consistent inferences.

The effect of the preprocessing steps on the model's ability to discriminate between healthy controls and depressed patients is visualized in Figure 3b. The lowest AUC is obtained using raw data (AUC: 0.63, 95% CI: 0.60-0.66), followed by background noise removal and no

speaker diarization (AUC: 0.66, 95% CI: 0.63-0.68), with data using both speaker diarization and background noise removal obtaining the highest AUC (AUC: 0.66, 95% CI: 0.63-0.69).[3]

Correspondingly, the model following the BEST approach [42] found marked differences between models trained on data with different levels of preprocessing, although with large uncertainty estimates. The model trained on raw data had the lowest effect size for the difference between controls and depression (Cohen's d 0.39, 95% CI: 0.08-0.71), background noise removal but no speaker diarization was in the middle (Cohen's d 0.49, 95% CI: 0.18-0.77), and background noise removal with speaker diarization had the largest effect size (Cohen's d 0.55, 95% CI: 0.24-0.86). Although the effect of speaker diarization is not as extreme as background noise, performing speaker diarization improves model performance.[4]

As shown in Figure 3c, AUC gradually increases with larger time-window sizes, however there seems to be a ceiling effect around windows of 20 seconds. The same figure shows that using the whole recording (no windowing) provides superior performance than assessing on a per-window basis. However, taking the mean prediction per participant using 20-30 second windows was found to be better than using the whole recording (no windowing: AUC: 0.68, 95% CI: 0.56, 0.80; mean of 30 second windows: AUC: 0.71, 95% CI: 0.59, 0.82).

Discussion

This study set out to investigate three main aims: 1) whether a speech emotion recognition model is useful for identifying depression 2) whether these predictions can be used for assessing changes in voice following remission, and 3) how much preprocessing steps impact the quality of model predictions.

We found that the speech emotion recognition model was able to accurately discriminate between healthy controls and depressed patients, obtaining an AUC of 0.71 (95% CI: 0.59-0.82). While voice patterns during the course of the disorder did not predict whether a patient would remit in the future, remission itself could be observed in the voice. Indeed, a treatment effect was observed, as the voice of patients in remission was indistinguishable from that of healthy controls and was credibly more *happy* sounding than during the disease. Model predictions were stable over the course of the interviews when using a 30-second time window, indicating that robust, long-term representations of voice are captured. Pre-processing steps had an impact on the

[3] See Supplementary Table S2 for more performance metrics.
[4] See Supplementary Table S3 and Supplementary Figure S3 for full model report.

model predictions: background noise had a large effect on model performance, and must be controlled for making meaningful inferences. Removing speech from the interviewer (speaker diarization) did not reliably affect model performance, however, it slightly increased effect size. Model performance was found to increase with larger window sizes at the expense of increased variance, and to stabilize around 20-30 seconds.

Heterogeneity in Depression

Although we found a credible difference in the voice of depressed patients and healthy controls there seem to be two subpopulations of people with depression. Approximately 30% of the patients in the depressed group were predicted as having similar emotional valence (from sad to happy) of their voices as the control group, whereas the remaining 70% have markedly more 'sad' sounding voices (see Figure 2). This large variability potentially arises from the fact that MDD is a highly heterogeneous disorder [48,49], and patients might therefore express disparate symptoms while still falling under the umbrella of MDD [40]. To partly account for this, MDD can be subdivided into a melancholic, anxious, and atypical type based on distinct symptomatic profiles [50]. The melancholic subtype might be of particular relevance for depression detection from speech, as it is characterized by severe anhedonia without mood reactivity, psychomotor disturbance, and neurovegetative symptoms [51]. Investigating whether specific depression subtypes are better captured by speech analysis than others is a field of further research. However, approximately 44% of patients can not be assigned a specific subtype, and only 15% have the melancholic subtype [50]. As a consequence, significant unexplained heterogeneity between patients remains and a more granular perspective based on specific symptoms might be better poised at describing this heterogeneity [52]. Models of depression from speech are inherently constrained by this factor, which underscores that the primary area of application for such systems should be screening and disease monitoring, not diagnosis.

Generalizability

The main focus of our work was to improve generalizability and robustness of depression detection from voice. In this regard, factors relating to intrapersonal variability, method of speech elicitation, and language must be discussed.

The intrapersonal variability among participants was heterogeneous. Within each interview, the probability of sounding happy varied on average from 5.7 to 7.9 percentual points depending on the diagnostic group.[5] The distribution of this value was long-tailed, meaning that predictions were relatively stable for the majority of participants, with a minority being very variable. Though clear trends were visible on the group aggregate level, the extent to which each participant's voice changed between visits differed markedly. Parts of the variability might stem from participants being in a different emotional state, from slightly different recording settings, or from different changes in depressive symptom profiles across visits. For example, one depressed patient might have started with a low number of symptoms influencing the voice and therefore not show a large change at visit two, whereas another might be in the opposite situation which would lead to a large change in predictions at the second visit. To increase robustness of the method, we advise practical applications to perform multiple recordings over multiple days. Further, to better understand the symptom profiles and patient cohorts which might benefit from voice-based depression detection systems, further studies should strive to include variables relating to symptom expression.

Several previous studies have found the method of speech elicitation to impact the patterns extracted from speech [53,54]. Patterns of pathological voice are expressed to a greater extent in more social and cognitively demanding tasks such as free speech than in read speech or vocal exercises [9,13]. A clinical interview can be considered an extremely social and cognitively demanding task, and might therefore provide an exceptionally strong signal for detecting depression from voice. Whether our model works equally well on voice elicited from other tasks remains to be tested.

Our model was trained to predict emotion in voice from English and German speech and validated on Danish speech for the task of detecting depression, thereby generalizing to new participants, tasks, and even to a new language. As a consequence, our results are highly likely to generalize to other languages and datasets of depressed speech. This is valuable for clinical implementations, as models need to handle a variety of languages, dialects, and accents. However, our model was only trained and tested on Germanic languages which leaves the extent of generalizability across language families unknown. The finding that emotional valence of speech (from sad to happy) from patients in remission is similar to healthy controls increases the credibility of our model, and suggests that acoustic features of speech could be used as an effective marker for depression. However, it should be noted that we only had access to 25 patients in

[5] See Supplementary Figure S5 and Supplementary Table S4.

remission and that we were not able to make comparisons with the group who remained depressed.

Effect of Preprocessing

This paper sought to investigate the effect of different preprocessing methods for producing reliable predictions, namely background noise removal, speaker diarization, and window size.

Background noise was found to drastically impact model performance by differentially distorting predictions based on the level of noise in the recordings. This has major implications for clinical implementations as recordings must be made under relatively noise-free conditions to avoid excessive false positives. For screening and monitoring, participants can be advised to perform the audio recordings in quiet surroundings, however automated quality checks and noise reduction methods are likely necessary.

Speaker diarization had some effect on model performance, but not to the same extent as noise removal. The audio recordings used in this study consisted of interviews conducted by the same psychologist with all the participants. In such a dyadic setting, the psychologist is likely to align his way of speaking with that of the participants thereby decreasing the effect of speaker diarization [55–57]. However, audio recorded in environments with multiple speakers or in more naturalistic settings (e.g. with sounds from people in adjacent rooms) will likely present more severe confounds without a speaker diarization process.

Larger window sizes, i.e the size of the time bin used for summarization of acoustic features, afforded better predictions until a ceiling effect at window sizes of 20-30 seconds of speech. Using the whole recording without windowing performed marginally better than windowing when evaluating on a per-window basis. However, when summarizing the prediction for each window into a single prediction per participant, the windowed summarization outperformed the whole recording, again with a ceiling effect at window sizes of 20-30 seconds. This suggests that recordings as short as 20-30 seconds of speech might be enough to provide high quality predictions, but longer recordings, if available, will slightly improve performance.

Limitations

The results of our study need to be interpreted in light of the following limitations. First, this study mainly served to investigate the usefulness of directly applying transfer learning from SER to the task of predicting depression under optimal conditions. In this regard, manual preprocessing procedures such as background noise removal and speaker diarization and corresponding quality checks were taken, but are not feasible for clinical implementations to the same extent. A large

part of the preprocessing pipeline can be automated, but performance is likely to suffer without a human in the loop.

Second, given that the focus of our study has been more on generalizability than predictive performance, we decided to only train our model on MFCCs although several other acoustic features have been found predictive of depression [9]. As reviewed, MFCCs have several desirable properties for generalization, whereas features such as pitch are highly gender dependent and therefore might harm generalizability. However, once larger and higher quality datasets of depressed speech become available, it might be beneficial to include features more specific to depression. In a similar vein, the model proposed here could easily be extended by training depression-specific experts and adding to the ensemble. Further, owing to the success of transformer-based neural network models in fields such as Natural Language Processing, it might prove beneficial to use speech models such as wav2vec 2.0 [58], or its multilingual variant [59], for SER and depression detection.

Third, our models are sensitive to the preprocessing steps and recording environment. Data must be carefully cleaned to ensure consistent noise profiles and inferences. Though our model was robust against speech from the interviewer, background noise from other people might further influence model predictions.

Fourth, though our model was validated on a relatively large dataset - at least compared to field standards -, there is a pronounced need for larger, high-quality, longitudinal datasets from diverse languages. Current publicly available databases often contain a low number of participants, sparse information on symptoms and demographic characteristics, and are primarily in English or other Germanic languages. Longitudinal studies following the same patient group over the course of their treatment could provide insights into the effectiveness and robustness of voice-based depression measures and potentially cast light on which subpopulation of patients the models work best for. Datasets from more diverse languages are required to assess cross-lingual performance. Our study provides an attempt at this, by training and testing on different languages.

Conclusions

Voice-based systems have the advantage of being less prone to biases related to self-reports and human ratings, and can be used remotely, cheaply, and non-invasively. Successful implementation of voice-based depression screening and monitoring has potential for providing earlier diagnosis and a more granular view of treatment effect, thereby facilitating improved prognosis of major depressive disorder. To reach this aim we showed the potential of transfer learning to identify the presence of depression, and identified conditions of applicability: at least 30 seconds of

voiced speech, strong attention to background noise and recording conditions, but no huge impact of diarization.

Conflicts of Interest

Lasse Hansen was an intern at F. Hoffmann-La Roche while conducting this research, and Riccardo Fusaroli has been a consultant for F. Hoffmann-La Roche on related topics. The remaining authors declare no competing interests.

References

1. James SL, Abate D, Abate KH, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*. 2018;392(10159):1789-1858. doi:10.1016/S0140-6736(18)32279-7

2. Fava M, Kendler KS. Major Depressive Disorder. *Neuron*. 2000;28(2):335-341. doi:10.1016/S0896-6273(00)00112-4

3. Krosnick JA. Survey research. *Annu Rev Psychol*. 1999;50(1):537-567.

4. Eaton WW, Neufeld K, Chen LS, Cai G. A comparison of self-report and clinical diagnostic interviews for depression: diagnostic interview schedule and schedules for clinical assessment in neuropsychiatry in the Baltimore epidemiologic catchment area follow-up. *Arch Gen Psychiatry*. 2000;57(3):217-222. doi:10.1001/archpsyc.57.3.217

5. Baumgartner H, Steenkamp J-BE. Response styles in marketing research: A cross-national investigation. *J Mark Res*. 2001;38(2):143-156.

6. Maj M, Stein DJ, Parker G, et al. The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry*. 2020;19(3):269-293. doi:https://doi.org/10.1002/wps.20771

7. Sobin C, Sackeim HA. Psychomotor symptoms of depression. *Am J Psychiatry*. 1997;154(1):4-17. doi:10.1176/ajp.154.1.4

8. Buyukdura JS, McClintock SM, Croarkin PE. Psychomotor retardation in depression: biological underpinnings, measurement, and treatment. *Prog Neuropsychopharmacol Biol Psychiatry*. 2011;35(2):395-409. doi:10.1016/j.pnpbp.2010.10.019

9. Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF. A review of depression and suicide risk assessment using speech analysis. *Speech Commun*. 2015;71:10-49.

10. Afshan A, Guo J, Park SJ, Ravi V, Flint J, Alwan A. Effectiveness of Voice Quality Features in Detecting Depression. In: *Interspeech*. ; 2018:1676-1680.

11. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig Otolaryngol*. 2020;5(1):96-116.

12. Fusaroli R, Lambrechts A, Bang D, Bowler DM, Gaigg SB. Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis. *Autism Res*. 2017;10(3):384-407.

13. Parola A, Simonsen A, Bliksted V, Fusaroli R. Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis. *Schizophr Res*. 2020;216:24-40. doi:10.1016/j.schres.2019.11.031

14. Weed E, Fusaroli R. Acoustic Measures of Prosody in Right-Hemisphere Damage: A Systematic Review and Meta-Analysis. *J Speech Lang Hear Res*. 2020;63(6):1762-1775.

15. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media; 2009.

16. Beleites C, Neugebauer U, Bocklitz T, Krafft C, Popp J. Sample size planning for classification models. *Anal Chim Acta*. 2013;760:25-33. doi:10.1016/j.aca.2012.11.007

17. Bone D, Chaspari T, Audhkhasi K, et al. Classifying language-related developmental disorders from speech cues: the promise and the potential confounds. In: *INTERSPEECH*. ; 2013:182-186.

18. Stasak B, Epps J, Cummins N, Goecke R. An Investigation of Emotional Speech in Depression Classification. In: ; 2016:485-489. doi:10.21437/Interspeech.2016-867

19. Harati S, Crowell A, Mayberg H, Nemati S. Depression Severity Classification from Speech Emotion. In: Vol 2018. ; 2018:5763-5766. doi:10.1109/EMBC.2018.8513610

20. Sechidis K, Fusaroli R, Orozco-Arroyave JR, Wolf D, Zhang Y-P. A machine learning perspective on the emotional content of Parkinsonian speech. *Artif Intell Med*. 2021;115:102061.

21. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans Affect Comput*. 2014;5(4):377-390.

22. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*. 2018;13(5):e0196391. doi:10.1371/journal.pone.0196391

23. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: *Ninth European Conference on Speech Communication and Technology*. ; 2005.

24. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *ArXiv170609516 Cs*. Published online January 20, 2019. Accessed February 15, 2021. http://arxiv.org/abs/1706.09516

25. Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. *Neural Comput*. 1991;3(1):79-87.

26. Masoudnia S, Ebrahimpour R. Mixture of experts: A literature survey. *Artif Intell Rev*. 2014;42. doi:10.1007/s10462-012-9338-y

27. Ladegaard N, Larsen ER, Videbech P, Lysaker PH. Higher-order social cognition in first-episode major depression. *Psychiatry Res*. 2014;216(1):37-43. doi:10.1016/j.psychres.2013.12.010

28. Ladegaard N, Videbech P, Lysaker PH, Larsen ER. The course of social cognitive and metacognitive ability in depression: Deficit are only partially normalized after full remission of first episode major depression. *Br J Clin Psychol*. 2016;55(3):269-286.

29. Hamilton M. The Hamilton rating scale for depression. In: *Assessment of Depression*. Springer; 1986:143-152.

30. Thompson AR. Pharmacological agents with effects on voice. *Am J Otolaryngol*. 1995;16(1):12-18. doi:10.1016/0196-0709(95)90003-9

31. Lysaker PH, Clements CA, Plascak-Hallberg CD, Knipscheer SJ, Wright DE. Insight and personal narratives of illness in schizophrenia. *Psychiatry Interpers Biol Process*. 2002;65(3):197-206.

32. iZotope. *RX 6 Elements*. iZotope; 2017. https://www.izotope.com/en/products/repair-and-edit/rx/rx-elements.html

33. Tiwari V. MFCC and its applications in speaker recognition. *Int J Emerg Technol*. 2010;1(1):19-22.

34. Akçay MB, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Commun*. 2020;116:56-76. doi:10.1016/j.specom.2019.12.001

35. Stolar M. Acoustic and conversational speech analysis of depressed adolescents and their parents. Published online 2016.

36. Zheng F, Zhang G. Integrating the energy information into MFCC. In: *Sixth International Conference on Spoken Language Processing*. ; 2000.

37. Taguchi T, Tachikawa H, Nemoto K, et al. Major depressive disorder discrimination using vocal acoustic features. *J Affect Disord*. 2018;225:214-220. doi:10.1016/j.jad.2017.08.038

38. Prasanth PS. Speaker Recognition Using Vocal Tract Features. *Int J Eng Invent*. 2013;3(1):26-30.

39. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognit*. 2011;44(3):572-587.

40. Fried EI. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *J Affect Disord*. 2017;208:191-197. doi:10.1016/j.jad.2016.10.019

41. Valton V, Wise T, Robinson OJ. Recommendations for Bayesian hierarchical model specifications for case-control studies in mental health. *ArXiv201101725 Cs Stat*. Published online November 3, 2020. Accessed March 6, 2021. http://arxiv.org/abs/2011.01725

42. Kruschke JK. Bayesian estimation supersedes the t test. *J Exp Psychol Gen*. 2013;142(2):573.

43. RStudio Team. *RStudio: Integrated Development for R*. RStudio, Inc.; 2016. http://www.rstudio.com/

44. R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing; 2018. https://www.R-project.org/

45. Bürkner P-C. brms: An R package for Bayesian multilevel models using Stan. *J Stat Softw*. 2016;80(1):1-28.

46. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77.

47. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *J Open Source Softw*. 2019;4(43):1686.

48. Goldberg D. The heterogeneity of "major depression." *World Psychiatry*. 2011;10(3):226-228.

49. Fried EI, Nesse RM. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. *J Affect Disord*. 2015;172:96. doi:10.1016/j.jad.2014.10.010

50. Musil R, Seemüller F, Meyer S, et al. Subtypes of depression and their overlap in a naturalistic inpatient sample of major depressive disorder. *Int J Methods Psychiatr Res*. 2018;27(1). doi:10.1002/mpr.1569

51. Association AP. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Pub; 2013.

52. Borsboom D, Cramer AOJ, Kalis A. Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behav Brain Sci*. 2019;42. doi:10.1017/S0140525X17002266

53. Calev A, Nigal D, Chazan S. Retrieval from semantic memory using meaningful and meaningless constructs by depressed, stable bipolar and manic patients. *Br J Clin Psychol*. 1989;28(1):67-73. doi:https://doi.org/10.1111/j.2044-8260.1989.tb00813.x

54. Vanger P, Summerfield AB, Rosen BK, Watson JP. Effects of communication content on speech behavior of depressives. *Compr Psychiatry*. 1992;33(1):39-41. doi:10.1016/0010-440X(92)90077-4

55. Pickering MJ, Garrod S. Toward a mechanistic psychology of dialogue. *Behav Brain Sci*. 2004;27(2):169-190.

56. Dale R, Fusaroli R, Duran ND, Richardson DC. The self-organization of human interaction. In: *Psychology of Learning and Motivation*. Vol 59. Elsevier; 2013:43-95.

57. Bone D, Lee C-C, Black MP, et al. The Psychologist as an Interlocutor in Autism Spectrum Disorder Assessment: Insights From a Study of Spontaneous Prosody. *J Speech Lang Hear Res JSLHR*. 2014;57(4):1162-1177. doi:10.1044/2014_JSLHR-S-13-0062

58. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *ArXiv200611477 Cs Eess*. Published online October 22, 2020. Accessed December 20, 2020. http://arxiv.org/abs/2006.11477

59. Conneau A, Baevski A, Collobert R, Mohamed A, Auli M. Unsupervised Cross-lingual Representation Learning for Speech Recognition. *ArXiv200613979 Cs Eess*. Published online December 15, 2020. Accessed June 11, 2021. http://arxiv.org/abs/2006.13979

Appendix

Speech Emotion Recognition with Mixture of Experts

We made use of the model trained in Sechidis et al. [1] for SER and briefly describe it here for completeness. The SER model was trained on the CREMA-D [2], RAVDESS [3], and EMO-DB [4] datasets and validated on EMOVO [5], TESS [6], and SAVEE [7]. Each corpus contains recordings from professional actors who repeat sentences while changing which emotion they convey. MFCC coefficients were extracted from each recording (10 ms windows), and summarized using the 11 descriptive statistics mentioned in the Methods section (mean, variance, kurtosis, skewness, mode, IQR, percentiles 10th, 25th, 50th, 75th, and 90th) per utterance.

For each training dataset, a gradient boosted decision tree model was fitted using CatBoost. The optimal hyperparameters for each model were estimated using subject-wise cross-validation stratified by gender. The three models were combined in a Mixture of Experts (MoE) architecture using the Mahalanobis distance as similarity metric. In practice, this entails that for each new sample to predict, the prediction from each constituent model is weighted in terms of how similar the new datapoint is to the model's training data.

To assess the effectiveness of the MoE, its performance was tested against each constituent model on their own, as well as a model trained on pooled data from all three datasets. Performance was assessed on the EMOVO, TESS, and SAVEE dataset on which the MoE was found to achieve superior performance.

Model Building

The Bayesian mixture model trained to assess the difference in the probability of sounding happy from the interaction between diagnosis and visit was trained using the *brms* R package [8]. The probability of sounding happy was transformed to log odds to improve model fit, and subsequently modelled as a mixture of two Gaussian distributions. Weakly regularizing priors were used for all parameters:

$$\mu_1, \mu_2 \sim Normal(0, 1)$$
$$\sigma_1, \sigma_2 \sim Normal(1, 0.5)$$
$$\theta_1 \sim Normal(0, 1)$$
$$sd(\theta_1) \sim Normal(3, 1)$$

The model was trained for 4000 iterations, including 1000 warmup iterations, on 4 chains, with adapt_delta set to 0.99 to improve convergence. All $\hat{R}$ [9] were below 1.001 and chains were visually

inspected for convergence with no obvious issue being found. See Supplementary Material Figure S6 and S7 for prior and posterior predictive checks and posterior update plots.

The Bayesian multilevel model trained to assess the difference in the probability between the diagnostic groups varying by level of preprocessing were also trained using the *brms* R package. One model was fit for each dataset (with different levels of preprocessing), with the same settings with regards to priors and samples. The models were fit following Kruschke [10], and modelled as a T-distribution. The priors for the betas were set as Gaussians with mean 0.5, and standard deviation 0.5, bounded at 0 and 1. Prior for the *nu* parameter for the normality of the T-distribution was an exponential with parameter 1/29 following Kruschke's [10] recommendation. Models were run for 6000 iterations on 4 chains, and passed all checks for convergence.

References

1. Sechidis K, Fusaroli R, Orozco-Arroyave JR, Wolf D, Zhang Y-P. A machine learning perspective on the emotional content of Parkinsonian speech. *Artif Intell Med*. 2021;115:102061.

2. Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Trans Affect Comput*. 2014;5(4):377-390.

3. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*. 2018;13(5):e0196391. doi:10.1371/journal.pone.0196391

4. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier WF, Weiss B. A database of German emotional speech. In: *Ninth European Conference on Speech Communication and Technology*. ; 2005.

5. Costantini G, Iaderola I, Paoloni A, Todisco M. EMOVO corpus: an Italian emotional speech database. In: *International Conference on Language Resources and Evaluation (LREC 2014)*. European Language Resources Association (ELRA); 2014:3501-3504.

6. Pichora-Fuller MK, Dupuis K. Toronto emotional speech set (TESS). Published online February 13, 2020. doi:10.5683/SP2/E8H2MF

7. Haq S, Jackson PJ, Edge J. Speaker-dependent audio-visual emotion recognition. In: *AVSP*. ; 2009:53-58.

8. Bürkner P-C. brms: An R package for Bayesian multilevel models using Stan. *J Stat Softw*. 2016;80(1):1-28.

9. Vehtari A, Gelman A, Simpson D, Carpenter B, Bürkner P-C. Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC. *Bayesian Anal*. 2021;16(2). doi:10.1214/20-BA1221

10. Kruschke JK. Bayesian estimation supersedes the t test. *J Exp Psychol Gen*. 2013;142(2):573.