IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 14 May 2024, accepted 16 May 2024, date of publication 20 May 2024, date of current version 28 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3402999

RESEARCH ARTICLE

Exploring Emotion and Emotional Variability as Digital Biomarkers in Frontotemporal Dementia Speech

YISHU GONG[1], FJONA PARLLAKU[2], KATERINA PLACEK[1], MARCO VILELA[1], BRIAN HAREL[1], ARTHUR SIMEN[1], BRIAN SUBIRANA[2], AMY BRODTMANN[3], ADAM VOGEL[4], AND BRIAN TRACEY[1]

[1]Takeda Pharmaceuticals Inc., Cambridge, MA 02142, USA
[2]Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3]Department of Neuroscience, School of Translational Medicine, Monash University, Melbourne, VIC 3800, Australia
[4]School of Health Sciences, The University of Melbourne, Melbourne, VIC 3052, Australia

Corresponding author: Yishu Gong (yishu.gong@takeda.com)

This work was supported by Takeda Development Center Americas, Inc., (Successor in Interest to Millennium Pharmaceuticals, Inc.).

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Human Research Ethics Committee of The University of Melbourne and Eastern Health, Australia.

ABSTRACT Frontotemporal Dementia (FTD) encompasses a diverse group of progressive neurodegenerative diseases that impact speech production and comprehension, higher-order cognition, behavior, and motor control. Traditional acoustic speech markers have been extensively studied in FTD, as have assessments capturing apathy and impairments in recognizing and expressing emotion. This work leverages machine learning to track changes in emotional content within the speech of individuals with FTD and healthy controls. The aim of the project is to develop tools for assessing and monitoring emotional changes in individuals with FTD, quantifying these subtle aspects of the disease and thus potentially providing insights for assessing future therapeutic interventions. A retrospective analysis was conducted on a dataset comprising standard elicited speech tasks performed by 78 individuals diagnosed with FTD and 55 healthy elderly controls. We employed an ensemble-based convolutional neural network (CNN) classifier trained on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset to extract emotion scores from processed speech samples. The classifier was applied with a sliding window to the FTD and healthy control narratives to facilitate a granular examination of emotional changes throughout longer speech samples. Analysis of variance (ANOVA) was used to test for group differences in average emotion scores as well as emotional variability over the duration of the speech samples. Compared to healthy controls, people with FTD demonstrated reduced emotional change in a monologue task describing a happy experience, as measured by the interquartile range (IQR) ($p < 0.005$) and slope of "happy" emotion scores vs. time ($p < 0.005$). During a picture description task, people with FTD displayed a slightly elevated average level of frustration ($p < 0.005$). Increased frustration levels in individuals with FTD could potentially indicate their difficulties in accomplishing the task. This study introduced the application of a pre-trained Speech Emotion Recognition (SER) model on overlapping short segments of extended speech samples, allowing for a detailed examination of emotional changes over time. Capturing the temporal evolution of emotional content offers a nuanced understanding of communication in individuals with FTD. Our findings lay the groundwork for further development of digital biomarkers to refine the assessment, monitoring, and understanding of the emotional and social communication impacts of FTD.

INDEX TERMS Artificial intelligence, biomarkers, bvFTD, dementia, emotion, FTD, speech, voice.

I. INTRODUCTION

Frontotemporal Dementia (FTD) encompasses a clinically heterogeneous spectrum of early-onset, progressive

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

neurodegenerative diseases affecting behaviour, speech, language, cognition and motor function, resulting in significant personal, social, and economic burden [1]. The disease has a significant impact on participants'' emotional processing; in particular, apathy, presenting clinically as a motivation and flatness of affect [2], relates to poorer disease prognosis including increased caregiver burden and progression of neurodegeneration [3]. Disease-modifying therapies targeting the underlying pathobiology of FTD are in development [4], driving the need for quantitative markers of disease presence and progression.

Digital speech biomarkers have shown significant potential in characterizing neurological and respiratory illness [5], [6], as well as in characterizing FTD subtype-specific deficits in acoustic and lexical aspects of speech [7], [8], [9]. The most common subtype, behavioral variant FTD (bvFTD) is clinically characterized by social-behavioral deficits, apathy, and executive dysfunction [2]; in terms of digital speech biomarkers, people with bvFTD demonstrate relative preservation of phonology, yet impaired prosody, frequent pausing, and reduced lexical diversity [10], [11], [12], [13]. Primary progressive aphasia (PPA) subtypes of FTD include the semantic variant (svPPA), characterized by word comprehension and confrontation naming deficits that lead for example to greater pronoun use, and the nonfluent variant (nfvPPA), which is characterized by apraxia, agrammatism and effortful, nonfluent speech leading to increased speech errors and pauses [8].

An important part of the burden of FTD, especially for caregivers, is that it causes deficits in recognizing or expressing emotion [14], [15], [16], [17] that differ among variants [18]. bvFTD is linked to impairments in emotion perception [19] and impaired processing of emotions like embarrassment is linked to behavioral disturbances in early bvFTD [20], [21]. Apathy is a particularly important aspect of emotional response in FTD, especially in bvFTD [3]. Apathy is currently assessed by clinicians through observation or clinician-administered testing [22], with multiple apathy scales in use [23], [24], [25], [26], [27]. Clinical-rated apathy shows less consistency than other measures used in identifying probable bvFTD [28]. It has been proposed that automated tools to objectively measure apathy and other emotional content of speech could be beneficial by allowing more wide-spread screening and assessment of participants [29], [30], [31].

Here, we sought to address these needs by applying recently developed tools for sentiment and emotion analysis. In many applications, sentiment is judged based on speech transcripts [32], [33]; for example, Friedman and Ballentine [32] analyzed transcripts to quantify effects of psychoactive substances. However, we believe audio has important advantages in our application, as emotion changes may occur at a fast time scale (within individual sentences) not easily captured through transcripts. Thus we focused instead on acoustic-based SER, which detects low-level latent features from acoustic data that can be used to categorize speaker emotions into discrete categories, including anger, boredom, disgust, surprise, fear, joy, happiness, neutral and sadness [34], [35], [36], [37]. Relevant to our work, Linz et al. [29] demonstrated an ability to predict clinical scores of apathy in a population of mild cognitive impairment (MCI) participants based on acoustic cues in speech. Other existing SER models based on audio input [35], [38], [39], [40], [41], [42], [43], [44] show are trained to predict emotion of an overall utterance but are not well suited to capturing temporal fluctuations in emotion (Appendix shows an example of over-training leading to rapid emotion switches; other issues in these models include discrepancies in audio clip duration and lack of a ''frustration'' class, a potentially important emotion in participants with cognitive difficulties). Recent work leverages labels generated by ChatGPT to further explore the intensity of emotions (instead of assigning a single class) [45]. However, the ChatGPT enhanced labels are still not a continuous output that can be easily used to track emotion changes over time.

We hypothesized that the impact of FTD on emotional expression would alter the perceived emotional content of speech in people with FTD, and that these impacts could be objectively quantified using automated SER analysis of speech tasks that elicit emotional content. Hence, in this study, we leveraged transfer learning techniques, by training a convolutional neural network (CNN)-based SER model to recognize emotions on a database of healthy controls, Interactive Emotional Dyadic Motion Capture (IEMOCAP) [46] and then applying this trained model it to a separate dataset of individuals with FTD and healthy elderly controls. Features designed to capture variability in expressed emotion were then extracted.

The main contribution of this paper lies in presenting a robust framework for tracking variability in expressed emotions; this framework leverages machine learning algorithms to analyze the temporal dynamics and variability of emotions during narratives via the slope and variability in scored emotions. This work builds on earlier work by members of our group [6], [47]. While we used a particular deep learning SER approach, our framework could be adapted to future SER models developed in this rapidly evolving field. A second contribution is that we compare these metrics in different participants (healthy elderly participants and those with FTD), demonstrating differences between the populations which suggest automated scoring of emotions has potential as a tool for clinical assessment.

The paper's structure is as follows: the Methods section covers dataset descriptions, audio data preprocessing steps, and the training and evaluation of the emotion recognition model through transfer learning on the widely used IEMOCAP dataset [46]. The Methods section also describes our approach for characterizing temporal variation in emotional content when our SER model (trained on IEMOCAP) is applied to new data. In the Results section, we first present performance of the emotion recognition model on IEMOCAP, then present findings from characterizing the

monologue and picture description tasks in FTD and healthy elderly participants. In the Discussion section, we discuss implications of the studys findings and as well as observed differences in emotional expression between FTD participants and healthy controls. In the appendix, we explore the limitations of using another pretrained emotion classifier model for our task (wav2vec2-IEMOCAP [35]).

Our results demonstrate that individuals with FTD demonstrate similar average levels of emotion as healthy controls during a monologue task but have significantly reduced variability vs. time in perceived emotion, consistent with a flatter, less emotionally engaged affect. In addition, individuals with FTD exhibit a small but significant increase in frustration scores during a picture description task. It is important to replicate these results in a dataset that includes clinician-rated scores of emotional processing and affect. Nevertheless, our results suggest that automated emotion scoring may be a useful tool for quantifying the impact of FTD and other disorders on participants'' ability to express emotion in daily life.

II. METHODS

A. DATA SOURCE

a: EMOTIONS TRAINING DATA

For SER model training, we used the IEMOCAP dataset to train machine learning classifiers for emotion categorization in audio recordings of speech [46]. The IEMOCAP dataset comprises audio recordings from professional actors who express a range of emotions through speech. It is organized into five sessions, each featuring a different pair of male and female actors engaged in both scripted and improvised dialogue.

To prepare the data, audio recordings from IEMOCAP were truncated into utterances of average length 4.5 seconds, ensuring that each utterance contained speech from a single actor. Trained annotators then classified these recordings into ten emotion categories: ''angry,'' ''happy,'' ''disgusted,'' ''fear,'' ''frustrated,'' ''excited,'' ''neutral,'' ''sad,'' ''surprised,'' and ''others.'' The ''others'' category represents emotions outside the predefined set for the dataset. Additionally, the category ''xxx'' was used to indicate samples where annotators did not reach a consensus.

The original IEMOCAP dataset consists of 10,039 utterances, with the emotional category distribution shown in Fig. 1. However, we filtered the dataset by excluding the ''others'' and ''xxx'' categories as these have limited interpretability. Furthermore, we excluded the categories ''fear,'' ''disgusted,'' and ''surprised'' due to their small sample sizes. Last, following the approach of previous studies using IEMOCAP [39], [40], we combined the ''excited'' and ''happy'' categories into a single ''happy'' category. The filtered IEMOCAP dataset used in this study consists of 7,380 utterances and the distribution of emotion categories is illustrated in Fig. 2.

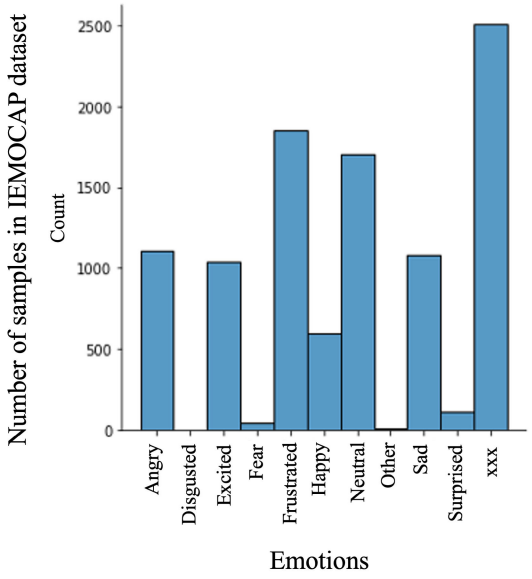

FIGURE 1. Distribution of all labeled emotion categories from the original IEMOCAP dataset. The category "xxx" represents samples where annotators did not reach a consensus; other emotion labels are as shown.

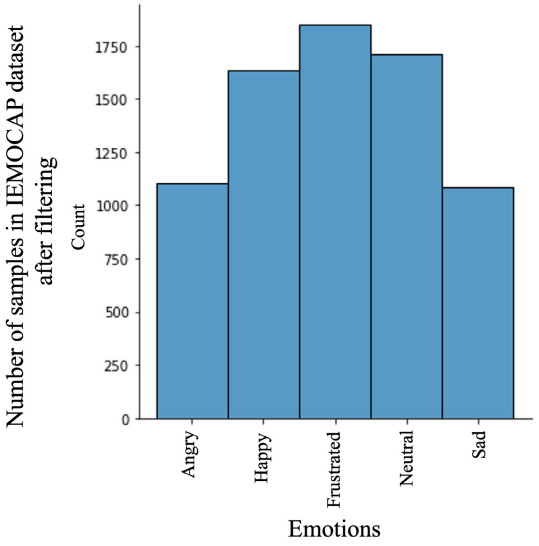

FIGURE 2. Distribution of filtered labeled emotion categories from the IEMOCAP dataset.

b: FTD AND HEALTHY PARTICIPANT DATA

We retrospectively analyzed an existing dataset containing audio recordings collected during elicited speech tasks from people with FTD and healthy elderly controls (denoted below as ''UMel dataset''). Participants provided informed consent and were seated in a sound-attenuated room (ambient noise in audio files <50dB$_A$; M = 36.5dB$_A$, SD = 1.27) [49]. Data from healthy elderly controls was collected at the University of Melbourne [50] and data from people with FTD was collected at the University of Melbourne and Monash University, both located in Melbourne, Australia [51]). We used the clinical diagnosis as our starting point, then checked the clinical features of each participant against the published

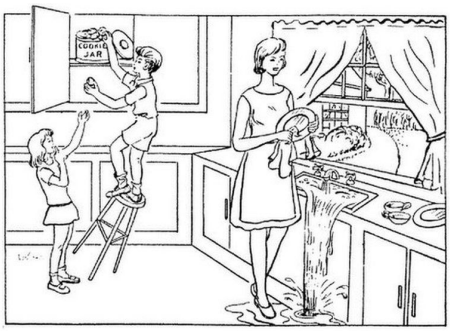

FIGURE 3. The "Cookie Theft" picture from the Boston Diagnostic Aphasia Examination [48].

criteria for each subtype [2], [52]. Only participants with a clinical diagnosis of bvFTD or primary progressive aphasia who fulfilled the published criteria for probable (not possible) or definite (for those with a known gene mutation) were included in this study. Participants were excluded if they presented with a behavioral disturbance better accounted for by a psychiatric diagnosis or biomarkers strongly indicative of Alzheimer's disease or other neurodegenerative process. Only participants with a clinical diagnosis of primary progressive aphasia (svPPA, lvPPA, nvPPA) according to consensus diagnostic criteria were included in the study [52]. No participants with a motor disorder (e.g., Corticobasal degeneration (CBD), Progressive Supranuclear Palsy (PSP), Amyotrophic lateral sclerosis (ALS)) were included in the study. Healthy elderly participants between 50 and 92 years of age were recruited through local networks and press releases in Melbourne. participants were excluded if they had a history of neurologic disease, traumatic brain injury, intellectual or hearing impairments, or any changes impairing the vocal tract. FTD participants were recruited from participants undergoing clinical care at the Eastern Cognitive Disorders Clinic, Box Hill Hospital, Melbourne, Australia. Recordings from people with FTD were collected during routine clinic visits (annually or less frequently), while recordings from healthy elderly controls were collected at a single visit. More specifically, samples were recorded using a Marantz PMD671 solid state recorder coupled with an AKG C520 cardioid head-mounted (frequency range, 20-20 KHz; sensitivity, 243 dB) condenser microphone positioned at a 458 angle 8 cm from the mouth. Recordings were sampled at 44.1 KHz and quantized at 8 bits [51]. The elicited speech tasks in this dataset measure acoustic, motoric, and linguistic aspects of speech and include picture description (Cookie Theft), monologue, sustained phonation, syllable repetition, word repetition, and days of the week.

In this study, we focused our analysis on participants with a diagnosis of bvFTD, nfvPPA, and svPPA. We chose to analyze only monologue and picture description tasks which have similar audio recording lengths as the IEMOCAP dataset utterances.

In the monologue task, participants are asked to describe a happy event of their own choice. In the picture description task, participants are shown a picture (Fig. 3) and asked to describe the picture in their own words [53]. All participants completed the monologue task, however, only a subset of people with FTD completed the picture description task.

The number of recordings per diagnosis for each task are shown in Table 1. In total, we analyzed audio recordings from 93 monologue and 44 picture description tasks from people with FTD, and audio recordings from 55 monologue and 58 picture description tasks from healthy elderly controls.

B. DATA PREPROCESSING

To maintain consistency between the IEMOCAP and UMel datasets, we first manually annotated and removed interviewer speech from audio recordings from the UMel dataset. This resulted in utterances containing only participant speech.

Next, we transformed all audio files from the IEMOCAP and UMel datasets to Mel Frequency Cepstral Coefficients (MFCC) representation using the librosa library [54]. Files were converted to a sampling rate of 22,050 Hz sampling rate and MFCCs were computed using default librosa parameters (2048-point Hanning windowed FFTs, 75% overlap, 20 MFCC coefficients, maximum frequency set to Nyquist rate). MFCC values were plotted to form images, which were first resized to 224×224 normalized with the mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225) for the three channels respectively. to meet the expected image size for machine learning classifiers. MFCC images had a fixed duration of five seconds, a parameter we chose with the rationale that the average length of an IEMOCAP utterances is 4.5 seconds. Shorter utterances were zero-padded to five seconds, while longer utterances were truncated to five seconds.

C. EMOTION RECOGNITION MODEL

Our analysis pipeline, as depicted in Fig. 4, involved two steps: 1) Training the emotion classifier, and 2) Conducting emotion tracking.

First, we employed standard transfer learning techniques to train five CNN source models: AlexNet [55], AlexNet-GAP [56], VGG11 [57], ResNet18, and ResNet50 [58]. Originally developed for image classification, these models were adapted to classify emotions using visual MFCC spectrograms from the IEMOCAP dataset. During the training phase, we designated session 5 of the IEMOCAP dataset as the test set to evaluate model performance. Sessions 1, 2, and 3 were used as the training sets. We performed hyperparameter tuning and model selection using session 4 as a validation set to optimize our models and select the best ensemble. To address class imbalance in the emotion categories, we employed the Synthetic Minority Oversampling Technique (SMOTE) [59] during model training. To prevent overtraining, we limited training epochs to 20, a significantly

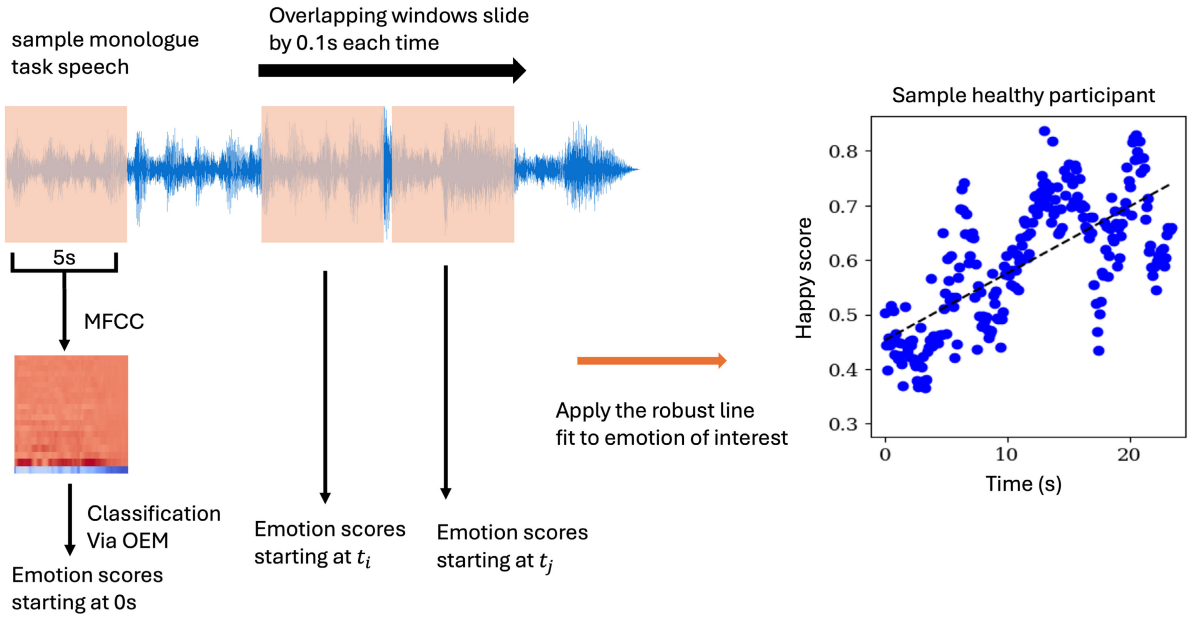

FIGURE 4. Flowchart illustrating the process of tracking emotions in a long speech sample from a monologue task, with the same process applied to speech samples from the PIC task. "MFCC" refers to "Mel Frequency Cepstral Coefficients", "OEM" refers to "Optimal Ensemble Model" described in III.

TABLE 1. Demographic characteristics of individuals with FTD and healthy elderly controls. "MONL" refers to the monologue task, and "PIC" refers to the picture description task.

Diagnosis	Gender	N (subjects) MONL	N (recordings) MONL	N (subjects) PIC	N (recordings) PIC	Age	Years since diagonsis
healthy	Female	29	29	31	31	64.5 ± 7.5	N/A
	Male	26	26	27	27	62.2 ± 7.8	N/A
bvFTD	Female	10	11	2	2	60.2± 5.2	3.5± 3.2
	Male	25	33	10	14	62.9± 7.5	4.5± 3.9
nfvPPA	Female	4	7	2	6	67.9± 3.1	2.4± 0.5
	Male	8	8	0	0	59.0± 5.5	1.8± 1.3
svPPA	Female	6	7	4	6	67.8± 9.4	5.4± 3.7
	Male	7	8	2	6	62.7± 7.0	3.2± 7.5

smaller value compared to similar works [34], [35], [36], [37]. The epoch size is chosen to be 32. The learning process is optimized using stochastic gradient descent (SGD) with a learning rate of 0.001 and a momentum factor of 0.9, exclusively applying updates to the parameters of the newly added fully connected layer. A learning rate scheduler is employed, reducing the learning rate by a factor of 0.1 every 7 epochs. Additionally, we utilized early stopping to determine the optimal epoch number. The ensemble was formed by averaging the probabilities generated by the five base models.

Next, we applied the trained emotion classifier to each recording in the UMel dataset to track the progression of emotions over time. For this purpose, we extracted emotion scores for five-second long windows and estimated the percentage of each emotion category (happy, neutral, frustrated, angry, and sad). We used sliding windows shifted by 0.1 seconds to smoothly capture emotions vs. time. This approach generated emotion densities that show the change in emotion percentages over the duration of the elicited speech tasks.

D. STATISTICAL ANALYSIS

We created custom Python code to statistically analyze the emotion percentages for each elicited speech task. We computed the mean, standard deviation, and (interquantile range) IQR of the emotion scores for each utterance. We employed Analysis of variance (ANOVA) to detect group differences. We also used the Games-Howell test for pairwise comparisons to determine which specific comparisons exhibited significant differences. In order to capture the changes in emotion scores, we performed robust line fits using the statsmodels packages with Huber distance [60]. The Huber distance is defined to be:

$$L_\delta(a) = \begin{cases} \frac{1}{2}a^2 & \text{for } |a| \le \delta \\ \delta \cdot (|a| - \frac{1}{2}\delta) & \text{otherwise.} \end{cases}$$

The Huber loss allow control over the treatment of ''inliers'' vs. ''outliers'' via one required user-specified parameter δ. Huber's parameter is interpreted as the maximum distance between the predicted value and the output value that would still be considered as an inlier (and thus use quadratic loss,

while the linear loss is used for outliers). In our case, the value of δ is chosen to be the median absolute deviation about the median. The Huber loss function is known to be relatively robust to outliers [61].

These line fits allowed us to estimate slopes and determine whether there were statistically significant differences between groups. Furthermore, we conducted binomial tests to evaluate whether the slopes of healthy elderly controls were outside the 95% range of slopes observed in people with FTD. Given the limited sample size in the picture description task, we combined all subtypes of FTD into a single FTD participant group for the statistical analyses.

III. RESULTS

A. EMOTION RECOGNITION MODEL PERFORMANCE

Model validation identified the best-performing model, referred to as the Optimal Ensemble Model (OEM), which combines AlexNet, ResNet18, and VGG11 architectures. This model classified five emotion categories (''angry'', ''happy'', ''sad'', ''neutral'', and ''frustrated'') with 50% accuracy (compared to the 20% expected for random guessing) and achieved a top-two class accuracy of 72%.

We compared our model to the wav2vec2-IEMOCAP model, a speech emotion recognition algorithm built upon the wav2vec2 base and fine-tuned on the IEMOCAP dataset. The wav2vec2-IEMOCAP model achieved 75% accuracy on the slightly simpler task of classifying four emotion categories (''angry'', ''happy'', ''sad'', and ''neutral'') [35]. However, limitations of the wav2vec2-IEMOCAP model in capturing changes in emotions over time are discussed in Appendix.

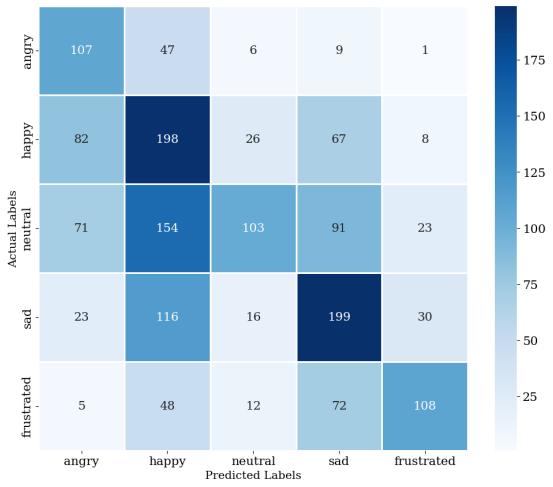

FIGURE 5. Confusion matrix of OEM, illustrating the model''s performance in accurately predicting emotion categories. The matrix displays the predicted emotion labels along the x-axis and the true emotion labels along the y-axis.

The confusion matrix for OEM depicts that that the model achieved the highest accuracies for the ''happy'' and ''sad'' emotions (Fig. 5). Receiver operating characteristic (ROC) analysis revealed a micro-averaged area under the curve (AUC) of 0.76.

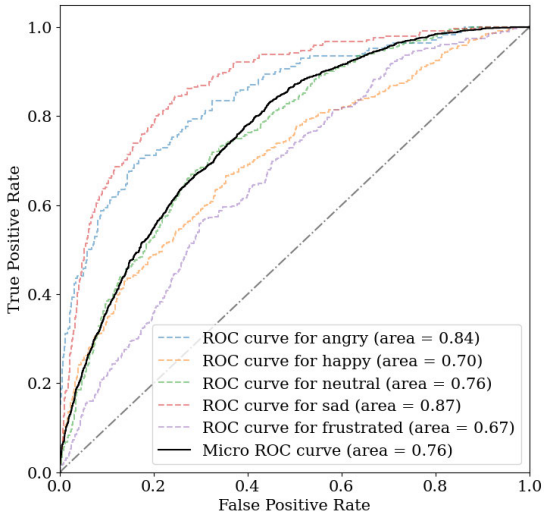

FIGURE 6. Receiver Operating Curve (ROC) of OEM, showing performance per emotion as well as micro-averaged performance.

The primary purpose of OEM was to accurately capture the distribution and density of emotion percentages and track the smooth transition of emotions over time, rather than to assign audio recordings to a single emotion category with high confidence. This particular use is highlighted with the subsequent application of OEM to the analysis of emotion over time in speech recordings ($\geq$30s).

B. EMOTION RECOGNITION IN FTD SPEECH

a: MONOLOGUE TASK

Across all diagnostic groups, OEM consistently identified ''happy'' as the emotion with the highest percentage, followed by ''neutral'' as the emotion with the second-highest percentage (Table 2). In contrast, the remaining emotions (''frustrated,'' ''angry,'' and ''sad'') had significantly lower percentages.

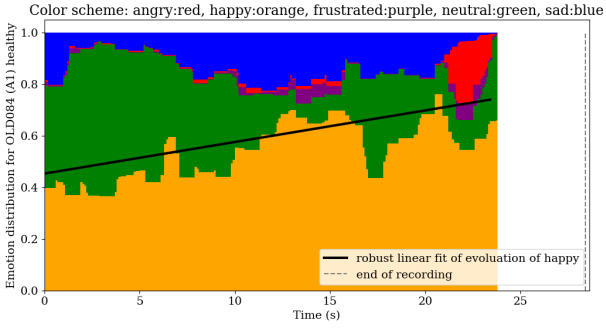

FIGURE 7. Emotion percentages over time for a healthy control participant completing the monologue task, obtained using OEM. The black line indicates the robust line fit of the ''happy'' emotion over time. Color scheme: red (''angry''), orange (''happy''), purple (''frustrated''), green (''neutral''), blue (''sad'').

Fig. 7 displays the output for a single healthy elderly control participant, while Fig. 8 illustrates an example of the model output for a single bvFTD participant. In both

TABLE 2. Mean and standard deviation of average emotion percentage over time for the monologue task obtained using OEM.

FTD subtype	Gender	N (recordings)	angry	happy	neutral	sad	frustrated
healthy	Female	29	$2.0\% \pm 0.7\%$	$59.5\% \pm 11.7\%$	$26.7\% \pm 9.1\%$	$5.5\% \pm 6.2\%$	$7.7\% \pm 4.5\%$
	Male	26	$2.9\% \pm 1.4\%$	$60.1\% \pm 12.0\%$	$25.2\% \pm 10.7\%$	$2.9\% \pm 2.8\%$	$8.8\% \pm 5.3\%$
bvFTD	Female	11	$3.5\% \pm 2.2\%$	$55.9\% \pm 14.0\%$	$29.7\% \pm 14.6\%$	$3.1\% \pm 2.3\%$	$7.7\% \pm 4.4\%$
	Male	33	$2.4\% \pm 2.0\%$	$52.4\% \pm 14.3\%$	$33.9\% \pm 12.2\%$	$4.7\% \pm 4.0\%$	$6.6\% \pm 4.4\%$
nfvPPA	Female	7	$2.3\% \pm 1.2\%$	$61.2\% \pm 16.1\%$	$30.4\% \pm 14.3\%$	$2.0\% \pm 1.9\%$	$4.2\% \pm 2.8\%$
	Male	8	$4.8\% \pm 1.7\%$	$58.8\% \pm 15.9\%$	$26.3\% \pm 12.6\%$	$1.7\% \pm 1.6\%$	$8.3\% \pm 8.0\%$
svPPA	Female	7	$2.5\% \pm 1.6\%$	$60.1\% \pm 17.8\%$	$29.4\% \pm 14.8\%$	$1.1\% \pm 0.6\%$	$6.8\% \pm 3.6\%$
	Male	8	$2.7\% \pm 2.7\%$	$43.3\% \pm 10.4\%$	$42.7\% \pm 9.8\%$	$4.0\% \pm 2.2\%$	$7.4\% \pm 5.0\%$

TABLE 3. Mean and standard deviation of average emotion over time for the picture description task obtained using OEM.

Health Status	Gender	N (recordings)	angry	happy	neutral	sad	frustrated
Healthy	Female	31	$10.9\% \pm 0.5\%$	$23.8\% \pm 1.3\%$	$22.3\% \pm 1.0\%$	$28.6\% \pm 0.9\%$	$14.4\% \pm 0.5\%$
	Male	27	$11.0\% \pm 0.6\%$	$22.4\% \pm 1.2\%$	$23.1\% \pm 0.9\%$	$29.2\% \pm 1.0\%$	$14.4\% \pm 0.5\%$
FTD	Female	18	$10.1\% \pm 0.8\%$	$22.5\% \pm 1.6\%$	$22.4\% \pm 0.7\%$	$27.9\% \pm 1.5\%$	$16.3\% \pm 2.0\%$
	Male	21	$11.3\% \pm 0.8\%$	$22.2\% \pm 1.3\%$	$22.4\% \pm 1.2\%$	$28.4\% \pm 1.2\%$	$15.5\% \pm 1.1\%$

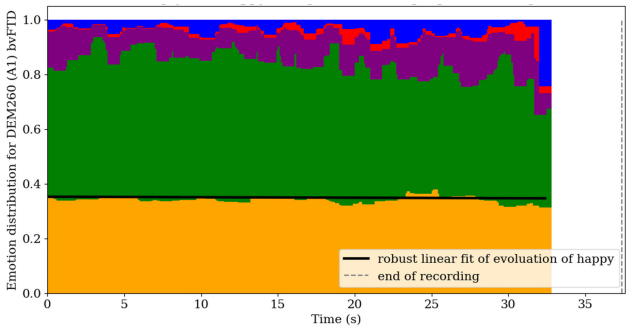

FIGURE 8. Emotion percentages over time for a bvFTD participant completing the monologue task, obtained using OEM. The black line indicates the robust line fit of the "happy" emotion over time. Color scheme: red ("angry"), orange ("happy"), purple ("frustrated"), green ("neutral"), blue ("sad").

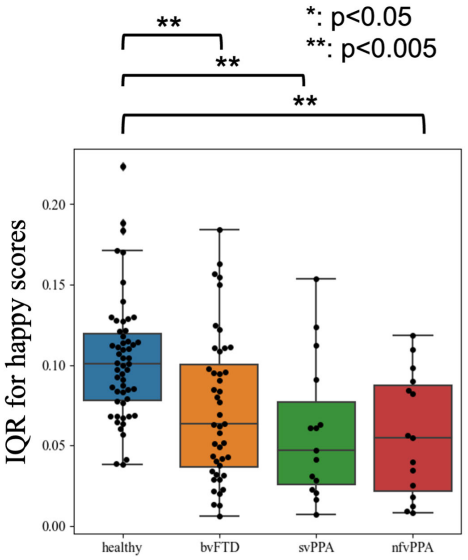

FIGURE 9. IQR of the "happy" emotion from the monologue task obtained using OEM.

cases, the model predictions exhibit smooth transitions over the time series, with the dominant emotion being "happy." This observation aligns with the nature of the monologue task, in which participants are instructed to describe a happy experience. In addition, the healthy elderly control participant had more noticeable variability in the happy emotion score over the duration of the audio recording, a point examined below in detail.

No statistically significant differences were observed in the mean and standard deviation of the "happy" emotion percentage over time among the different diagnostic groups. However, the healthy elderly controls exhibited a wider IQR for the "happy" emotion compared to bvFTD, svPPA, and nfvPPA, as depicted in Fig. 9 with statistical significance ($p < 0.005$) observed between healthy elderly and each FTD subtypes.

To further investigate the relationship between the variability in the "happy" emotion percentage over time and the observed difference in the IQR between FTD participants and healthy elderly controls, we detrended the time series data. After detrending, we observed that only one participant group (svPPA) retained a statistically significant difference in IQR from the healthy elderly controls. This indicates that the between-group differences in IQR are mainly explained by the overall trends in the "happy" emotion percentage over time, rather than fluctuations around the trend.

We next explored changes in the "happy" emotion percentage over time as a novel measure of emotional variability. To quantify this, we calculated the absolute value of the change in the "happy" emotion percentage over time (as depicted in Fig. 11). Notably, the slopes of the "happy" emotion percentage indicated change over time in healthy elderly controls, while in people with FTD the slopes were relatively stable. Robust line fit analysis indicated that the "happy" emotion percentage slopes from healthy elderly controls deviated from the 95% confidence intervals of the "happy" emotion percentage slopes from FTD participants (Fig. 10). This difference was found to be

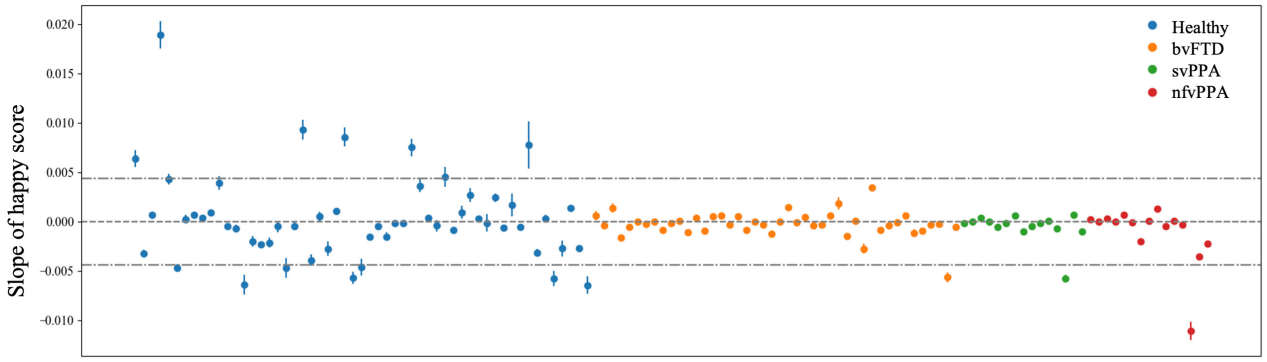

FIGURE 10. Slope (with 95% confident interval) for the robust line fit of "happy" emotion percentage over time for the monologue task obtained using OEM. The dotted-dashed line denotes the range that contains 95% of the slope of FTD participants.

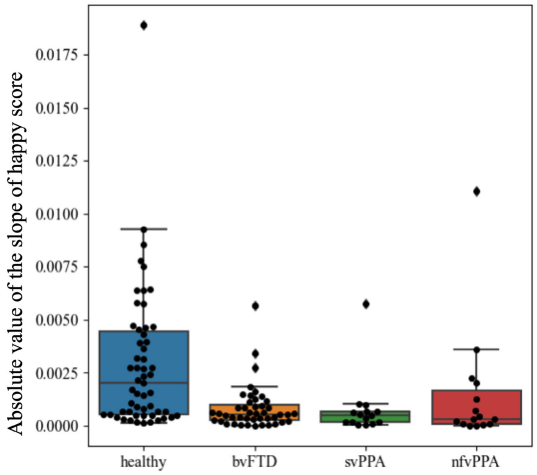

FIGURE 11. Absolute value of the slope of "happy" emotion over time on the monologue task obtained using OEM.

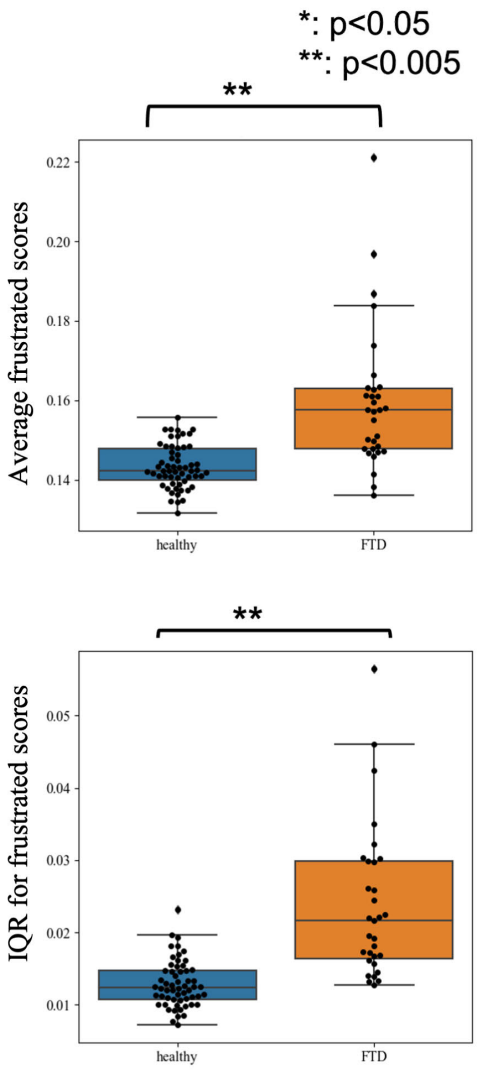

FIGURE 12. Average and IQR (per-recording) of "frustrated" emotion over time for the picture description task obtained using OEM. "**" means Bonferroni corrected p-values < 0.005, "*" means Bonferroni corrected p-values < 0.05, "N.S" means not statistically significant.

statistically significant from binomial testing (Binomial test p-value = 0.0008).

As a sensitivity analysis, we conducted a re-analysis of the monologue task using the emotions classifier described previously [35]. This classifier is known for its high accuracy in categorical classification of emotion categories, although it is not specifically designed to capture mixtures of emotions as OEM does through emotion percentages. The results presented in the appendix validated our observation of reduced variability in the "happy" emotion for FTD participants, thereby supporting the robustness of our findings.

b: PICTURE DESCRIPTION TASK

In contrast to the monologue task, the picture description task did not exhibit a dominant emotion in any group, as indicated in Table 3. However, when comparing the combined FTD group (bvFTD, svPPA, and nfvPPA) to healthy elderly controls, we observed a statistically significant higher average, standard deviation, and IQR for the "frustrated" emotion percentage, as illustrated in Fig. 12.

A robust line fit analysis of the "frustrated" emotion percentage showed that 95% of all participants, including both FTD and healthy elderly controls, had negligible changes over time, with absolute slopes of the "frustrated" emotion percentage below 0.001. This suggests that the

observed differences in the ''frustrated'' emotion between groups are not driven by temporal variation but rather reflect inherent distinctions between the groups. It is possible that the decline in formal language abilities experienced by participants contributes to a sense of loss of control, which can lead to their frustration and anger [62], [63]. These may contribute to the subtle differences in the ''frustrated'' emotion percentage observed.

IV. DISCUSSION

The research community currently lacks objective digital biomarkers for assessing emotional response in neurological disorders such as FTD. Hence, our goal was to capture the dynamic nature of emotions in long narratives from people with FTD so that we can assess the emotional characteristics that have been observed by caregivers and family of people with FTD. Previous SER algorithms have focused on assigning a single categorical emotion label to each utterance. In contrast, we sought to capture variations in expressed emotions over time. Due to this different goal, we found the other pre-trained SER algorithms unsuitable for tracking emotion changes (e.g. discrepancies in audio clip durations, overtraining leading to rapid emotion switches, and overlooking frustration, see Appendix). Therefore, we developed a new model following standard transfer learning techniques, and then applied our ensemble-based method for SER with a sliding window to long narratives from people with FTD and healthy elderly controls so that we can capture granular temporal variations in expressed emotions. Statistical analysis of model-assigned emotion percentages from bvFTD and healthy participants showed differing results by task. In monologue tasks where participants were asked to describe a happy experience, both FTD and healthy elderly controls demonstrated high levels of ''happy'' and ''neutral'' emotions, but people with FTD exhibited less emotional variability relative to healthy elderly controls. In particular, people with FTD uniformly showed flat trajectories, i.e., the slope of line fits over time, in the dominant ''happy'' emotion. In contrast, healthy elderly controls were statistically much more likely to exhibit variability in the ''happy'' emotion percentage assigned by the model. In the picture description task (Cookie Theft), there was no dominant emotion, and neither FTD nor healthy elderly control participants exhibited noticeable emotional variation over time. However, people with FTD exhibited higher average percentages of ''frustration'', potentially because this task is more difficult for individuals with FTD. It is important to note that our available dataset for picture description was much smaller than for monologue, which impacts our ability to analyze FTD subtypes.

It is important to note that SER scores can only capture *perceived* emotion, which is relevant for social interaction but may differ from true emotions experienced by participants. For example, because FTD may affect motor control of speech [51], it is possible that people with FTD may have altered ability to express emotions that they may be experiencing. However, the general trend to more negative emotions like frustration and to less emotional variability is consistent with clinical assessments of FTD. People with bvFTD are known to exhibit increased apathy [3], which may be related to the flatness of happy emotion percentages we observed in the monologue data for bvFTD participants.

Our work has several important limitations related to the dataset used. Our dataset does not include the published batteries for assessing participants' apathy or emotional processing. Analyzing a dataset which does include these ratings would allow us to better validate our approach relative to gold standard speaking difficulties. It would be also interesting to see how well the results correlate for example with speaking difficulties or with caregiver-reported behavior from people with FTD.

A second limitation is that our dataset is not large (especially for picture description tasks), so it is important that these results be verified in additional datasets and in related neurological conditions, and that the elicited speech task dependence we report above should be further explored. A related limitation concerns the variability in the manifestations of FTD within and across subtypes. The analysis of the picture description test which combined all FTD participants into a single category is especially affected by this. In addition, we lacked longitudinal data, which limited our ability to look for changes in emotion expression as the disease progresses. Identifying or collecting longitudinal datasets in FTD would allow researchers to understand how the emotional content of FTD speech changes longitudinally over disease progression, similar to [64].

A final technical limitation is that the training of the emotion classifier was performed on the IEMOCAP dataset which consists of speech samples from US English speakers, while classification was performed in an Australian cohort (note, we were not able to identify an emotions training dataset for speakers of Australian English). While there is evidence that basic emotions are communicated across cultures [65], and US and Australian speech appear to be similar enough that Americans and Australians can generally recognize each other's emotions, recent work suggests the possible existence of ''emotional accents'' [66]. Thus, this train-test accent mismatch could potentially impact emotion scoring results. However, while average emotions scores could be affected by this mismatch in accent, features related to variability of emotions within a single recording (important in the monologue task) should be more robust to this accent mismatch. The future scope involves addressing the gaps noted above, most importantly by replicating the work in a larger, ideally longitudinal dataset which includes clinical batteries for rating apathy and emotional processing. A second area for future work involves adapting this emotion tracking framework to the continuously developing field of SER. This adaptation may even encompass the integration of diverse modalities, such as transcriptions and video recordings, enabling a more nuanced and comprehensive

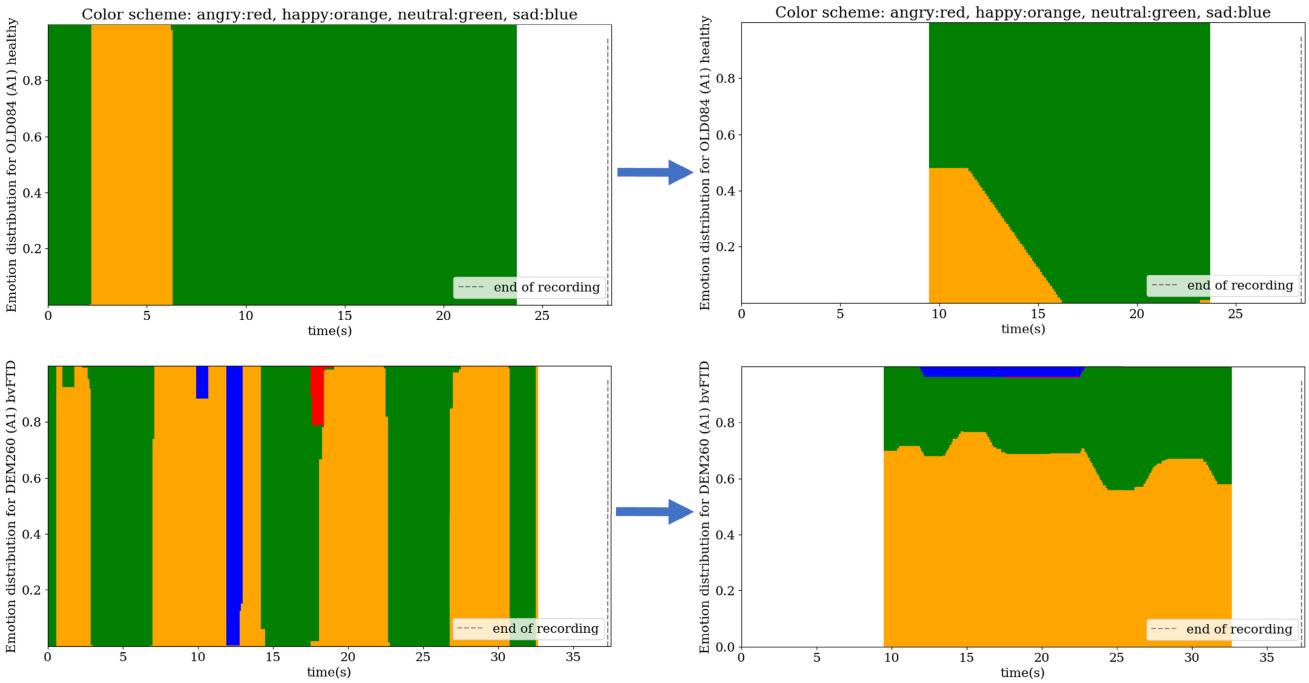

FIGURE 13. Emotion tracking over time for a bvFTD participant and a healthy participant completing the monologue task using wav2vec2-IEMOCAP. A moving average filter of size 10 seconds was used to create a smooth transition between emotions. Color scheme: red ("angry"), orange ("happy"), green ("neutral"), blue ("sad").

understanding of the emotions exhibited by FTD participants. Further development of SER-based emotion recognition tools that focuses on detecting apathy and levels of apathy may be a valuable tool for future research and clinical trials as noted in [31]. If such tools prove clinically useful, it will be critical to overcome the known limitations of current speech processing methods in low resource languages [67] and when data quality is non-ideal [68]. If robust methods are demonstrated in multiple datasets, SER-based biomarkers could provide clinicians with a valuable new tool for evaluating the quality of life and social interactions in people with FTD and related disorders.

APPENDIX
EMOTION TRACKING USING WAV2VEC2-IEMOCAP

In this appendix, we present the results of a sensitivity study that employs the previously developed wav2vec2-IEMOCAP classifier [35] on our FTD Monologue task dataset, following the methodology outlined in the Methods section.

As previously mentioned, wav2vec2-IEMOCAP demonstrates superior performance in classifying emotions within the IEMOCAP dataset. However, this sensitivity study reveals the limitations of previously developed emotion classifiers when applied to our dataset, particularly: a) the absence of frustration as a class, and b) overfitting, leading to rapid switching between emotions. These shortcomings underscore the need for the development of the OEM classifier. Nevertheless, the wav2vec2-IEMOCAP results support our overall finding that FTD participants exhibit "flatter" emotional trajectories during monologues, even when assessed using a different classifier.

Typically, as machine learning classifiers undergo more training epochs, they become more confident in their predictions, resulting in a higher probability of classifying an audio clip accurately. Here we present the predicted emotions over time in Fig. 13 for the same two participants demonstrated earlier in the main text in Fig. 7 and Fig. 8.

In the healthy example, emotions shift rapidly from "neutral" to "happy", and back to "neutral" shown on the top of Fig. 13. Rapid changes can also be seen in the bvFTD example in the bottom of Fig. 13, the participant's emotion changed from "happy" to "sad", back to "happy", then to "neutral" within 5 second changes, even though these adjacent 5 second time windows have a significant overlap.

Similar to results in the main body of the paper, we observe "happy" and "neutral" as the dominant emotions throughout the monologue. However, rather than showing a happy-neutral mixture, the wav2vec2-IEMOCAP model shows rapid switching between emotions. The interpretation of these switches becomes challenging since the 5 second time windows only differs by 0.1s. This contrasts with the smoother transition observed in Fig. s 7 and 8. However, by applying a moving average filter with a window size of 10 seconds, we can obtain more gradual transitions between emotions from the wav2vec2-IEMOCAP output. We explored varying the length of the moving average window from 5 to 15 seconds and found that differences

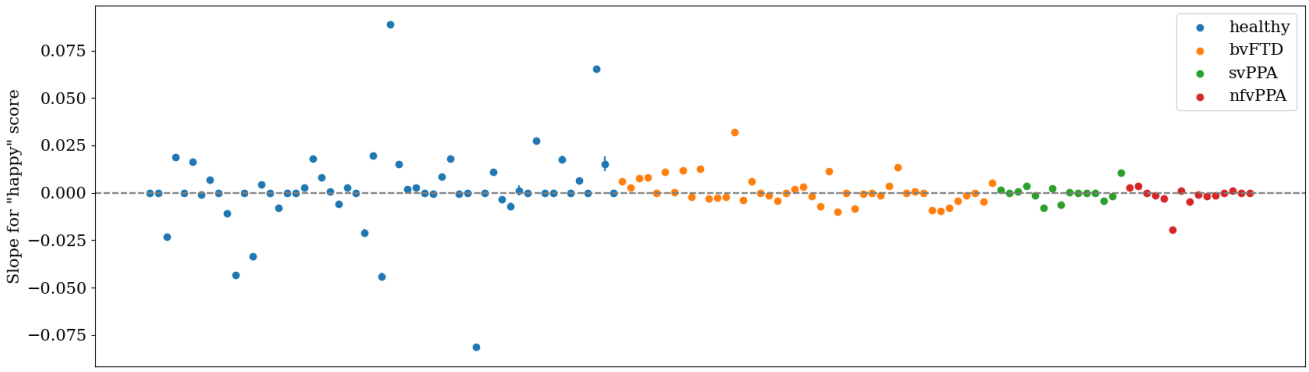

FIGURE 14. Slope for the robust line fit of "happy" percentage over time for all participants performing monologue task. Happy percentage was obtained using wav2vec2-IEMOCAP and a moving average filter of size 10 seconds.

between FTD and control participants were robust to the window length. It is important to note that this approach results in the loss of information from the first 10 seconds of the recordings. With this modified analysis, we can perform similar analyses as with OEM proposed in the main body. However, this time-smoothing approach is an ad-hoc fix to address the unrealistically rapid switching predicted by the wav2vec2-iemocap model. Thus we feel the proposed OEM, which is specifically trained to avoid overly confident predictions, is preferable.

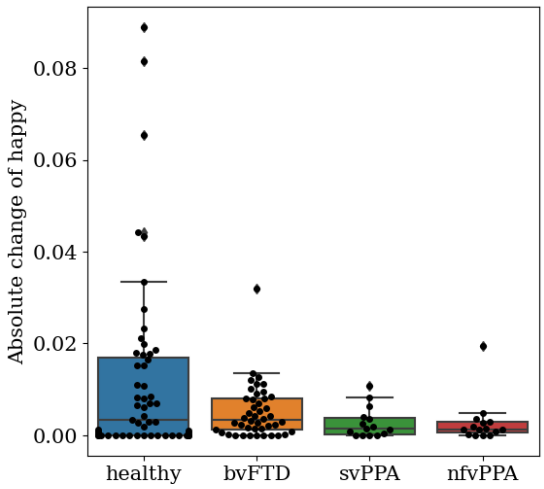

FIGURE 15. "Absolute change in "happy" score over time for all participants performing the monologue task, grouped by type. "Happy" percentage was obtained using wav2vec2-IEMOCAP with a moving average filter of 10 seconds.

Fig. 14 illustrates that a significant number of healthy individuals exhibited either a positive or negative trend in their "happy" percentage as they described their happy experiences. In contrast, most people with bvFTD and its subtypes appeared to have a relatively stable "happy" percentage over time. To quantify this observation, we calculated the absolute value of the slope for each participant, which is presented in Fig. 15.

Comparing with Fig. 11 in the main body of the paper, we observe that healthy elderly demonstrates more variability than people with bvFTD, svPPA, and nfvPPA.

ACKNOWLEDGMENT

The authors would like to extend their sincere appreciation to Yuan Gong and Jim Glass for their valuable suggestions and review. Their valuable input and expertise were instrumental in the completion of this work. *(Yishu Gong and Fjona Parllaku are co-first authors.)*

REFERENCES

[1] J. E. Galvin, D. H. Howard, S. S. Denny, S. Dickinson, and N. Tatton, "The social and economic burden of frontotemporal degeneration," *Neurology*, vol. 89, no. 20, pp. 2049–2056, Nov. 2017.

[2] K. Rascovsky et al., "Sensitivity of revised diagnostic criteria for the behavioural variant of frontotemporal dementia," *Brain*, vol. 134, no. 9, pp. 2456–2477, 2011.

[3] L. Massimo, J. P. Powers, L. K. Evans, C. T. Mcmillan, K. Rascovsky, P. Eslinger, M. Ersek, D. J. Irwin, and M. Grossman, "Apathy in frontotemporal degeneration: Neuroanatomical evidence of impaired goal-directed behavior," *Frontiers Hum. Neurosci.*, vol. 9, p. 611, Nov. 2015.

[4] S. Amin, G. Carling, and L. Gan, "New insights and therapeutic opportunities for progranulin-deficient frontotemporal dementia," *Current Opinion Neurobiol.*, vol. 72, pp. 131–139, Feb. 2022.

[5] A. Tsanas, M. A. Little, P. E. McSharry, J. Spielman, and L. O. Ramig, "Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1264–1271, May 2012.

[6] J. Laguarta, F. Hueto, and B. Subirana, "COVID-19 artificial intelligence diagnosis using only cough recordings," *IEEE Open J. Eng. Med. Biol.*, vol. 1, pp. 275–281, 2020.

[7] A. Geraudie, P. Battista, A. M. García, I. E. Allen, Z. A. Miller, M. L. Gorno-Tempini, and M. Montembeault, "Speech and language impairments in behavioral variant frontotemporal dementia: A systematic review," *Neurosci. Biobehav. Rev.*, vol. 131, pp. 1076–1095, Dec. 2021.

[8] N. Nevler, S. Ash, D. J. Irwin, M. Liberman, and M. Grossman, "Validated automatic speech biomarkers in primary progressive aphasia," *Ann. Clin. Transl. Neurol.*, vol. 6, no. 1, pp. 4–14, Jan. 2019.

[9] M. L. Poole, A. Brodtmann, D. Darby, and A. P. Vogel, "Motor speech phenotypes of frontotemporal dementia, primary progressive aphasia, and progressive apraxia of speech," *J. Speech, Lang., Hearing Res.*, vol. 60, no. 4, pp. 897–911, Apr. 2017.

[10] A. Geraudie et al., "Speech and language impairments in behavioral variant frontotemporal dementia: A systematic review," *Neurosci. Biobehavioral Rev.*, vol. 131, pp. 1076–1095, 2021.

[11] N. Nevler, S. Ash, C. Jester, D. J. Irwin, M. Liberman, and M. Grossman, "Automatic measurement of prosody in behavioral variant FTD," *Neurology*, vol. 89, no. 7, pp. 650–656, 2017.

[12] Y. Yunusova, N. L. Graham, S. Shellikeri, K. Phuong, M. Kulkarni, E. Rochon, D. F. Tang-Wai, T. W. Chow, S. E. Black, L. H. Zinman, and J. R. Green, "Profiling speech and pausing in amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD)," *PLoS ONE*, vol. 11, no. 1, Jan. 2016, Art. no. e0147573.

[13] S. Cho, N. Nevler, S. Ash, S. Shellikeri, D. J. Irwin, L. Massimo, K. Rascovsky, C. Olm, M. Grossman, and M. Liberman, "Automated analysis of lexical features in frontotemporal degeneration," *Cortex*, vol. 137, pp. 215–231, Jan. 2021.

[14] A. Wright, S. Saxena, S. M. Sheppard, and A. E. Hillis, "Selective impairments in components of affective prosody in neurologically impaired individuals," *Brain Cognition*, vol. 124, pp. 29–36, Jul. 2018.

[15] K. P. Rankin, A. Salazar, M. L. Gorno-Tempini, M. Sollberger, S. M. Wilson, D. Pavlic, C. M. Stanley, S. Glenn, M. W. Weiner, and B. L. Miller, "Detecting sarcasm from paralinguistic cues: Anatomic and cognitive correlates in neurodegenerative disease," *NeuroImage*, vol. 47, no. 4, pp. 2005–2015, Oct. 2009.

[16] C. Dara, L. Kirsch-Darrow, E. Ochfeld, J. Slenz, A. Agranovich, A. Vasconcellos-Faria, E. Ross, A. E. Hillis, and K. B. Kortte, "Impaired emotion processing from vocal and facial cues in frontotemporal dementia compared to right hemisphere stroke," *Neurocase*, vol. 19, no. 6, pp. 521–529, Dec. 2013.

[17] P. Desmarais, K. L. Lanctôt, M. Masellis, S. E. Black, and N. Herrmann, "Social inappropriateness in neurodegenerative disorders," *Int. Psychogeriatrics*, vol. 30, no. 2, pp. 197–207, Feb. 2018.

[18] J. S. Snowden, "Distinct behavioural profiles in frontotemporal dementia and semantic dementia," *J. Neurol., Neurosurgery Psychiatry*, vol. 70, no. 3, pp. 323–332, Mar. 2001.

[19] C. Strikwerda-Brown, S. Ramanan, Z.-L. Goldberg, A. Mothakunnel, J. R. Hodges, R. M. Ahmed, O. Piguet, and M. Irish, "The interplay of emotional and social conceptual processes during moral reasoning in frontotemporal dementia," *Brain*, vol. 144, no. 3, pp. 938–952, Apr. 2021.

[20] V. E. Sturm, E. A. Ascher, B. L. Miller, and R. W. Levenson, "Diminished self-conscious emotional responding in frontotemporal lobar degeneration patients," *Emotion*, vol. 8, no. 6, pp. 861–869, Dec. 2008.

[21] R. W. Levenson and B. L. Miller, "Loss of cells—Loss of self: Frontotemporal lobar degeneration and human emotion," *Current Direction Psychol. Sci.*, vol. 16, no. 6, pp. 289–294, Dec. 2007.

[22] F. Arshad, A. Paplikar, S. Mekala, F. Varghese, V. V. Purushothaman, D. J. Kumar, L. Shingavi, S. Vengalil, S. Ramakrishnan, R. Yadav, P. K. Pal, A. Nalini, and S. Alladi, "Social cognition deficits are pervasive across both classical and overlap frontotemporal dementia syndromes," *Dementia Geriatric Cognit. Disorders Extra*, vol. 10, no. 3, pp. 115–126, Nov. 2020.

[23] R. S. Marin, R. C. Biedrzycki, and S. Firinciogullari, "Reliability and validity of the apathy evaluation scale," *Psychiatry Res.*, vol. 38, no. 2, pp. 143–162, Aug. 1991.

[24] Y.-S. Ang, P. Lockwood, M. A. J. Apps, K. Muhammed, and M. Husain, "Distinct subtypes of apathy revealed by the apathy motivation index," *PLoS ONE*, vol. 12, no. 1, Jan. 2017, Art. no. e0169938.

[25] R. Levy and B. Dubois, "Apathy and the functional anatomy of the prefrontal cortex–basal ganglia circuits," *Cerebral Cortex*, vol. 16, no. 7, pp. 916–928, Jul. 2006.

[26] W. Fitts, L. Massimo, N. Lim, M. Grossman, and N. Dahodwala, "Computerized assessment of goal-directed behavior in Parkinson's disease," *J. Clin. Experim. Neuropsychol.*, vol. 38, no. 9, pp. 1015–1025, Oct. 2016.

[27] P. Sockeel, "The Lille apathy rating scale (LARS), a new instrument for detecting and quantifying apathy: Validation in Parkinson's disease," *J. Neurol., Neurosurg. Psychiatry*, vol. 77, no. 5, pp. 579–584, May 2006.

[28] A. K. LaMarre et al., "Interrater reliability of the new criteria for behavioral variant frontotemporal dementia," *Neurology*, vol. 80, no. 21, pp. 1973–1977, 2013.

[29] N. Linz, X. Klinge, J. Tröger, J. Alexandersson, R. Zeghari, R. Philippe, and A. König, "Automatic detection of apathy using acoustic markers extracted from free emotional speech," in *Proc. IJCAI*, 2018.

[30] P. Robert et al., "Is it time to revise the diagnostic criteria for apathy in brain disorders? The 2018 international consensus group," *Eur. Psychiatry*, vol. 54, pp. 71–76, Oct. 2018.

[31] K. L. Lanctôt, L. Agüera-Ortiz, H. Brodaty, P. T. Francis, Y. E. Geda, Z. Ismail, G. A. Marshall, M. E. Mortby, C. U. Onyike, P. R. Padala, A. M. Politis, P. B. Rosenberg, E. Siegel, D. L. Sultzer, and E. H. Abraham, "Apathy associated with neurocognitive disorders: Recent progress and future directions," *Alzheimer's Dementia*, vol. 13, no. 1, pp. 84–100, Jan. 2017.

[32] S. F. Friedman and G. Ballentine, "Language models learn sentiment and substance from 11,000 psychoactive experiences," *BioRxiv*, 2022, doi: 10.1101/2022.06.02.494544.

[33] H.-C. Chou, W.-C. Lin, C.-C. Lee, and C. Busso, "Exploiting annotators' typed description of emotion perception to maximize utilization of ratings for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 7717–7721.

[34] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[35] M. Ravanelli et al., "SpeechBrain: A general-purpose speech toolkit," 2021, *arXiv:2106.04624*.

[36] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D log-mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[37] M. B. Er, "A novel approach for classification of speech emotions based on deep and acoustic features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020.

[38] R. Radakovic, S. Colville, D. Cranley, J. M. Starr, S. Pal, and S. Abrahams, "Multidimensional apathy in behavioral variant frontotemporal dementia, primary progressive aphasia, and Alzheimer disease," *J. Geriatric Psychiatry Neurol.*, vol. 34, no. 5, pp. 349–356, Sep. 2021.

[39] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL based multimodal emotion recognition and sentiment analysis," in *Proc. IEEE 16th Int. Conf. Data Mining (ICDM)*, Dec. 2016, pp. 439–448.

[40] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 7390–7394.

[41] Z. Li, F. Tang, M. Zhao, and Y. Zhu, "EmoCaps: Emotion capsule based model for conversational emotion recognition," 2022, *arXiv:2203.13504*.

[42] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," 2022, *arXiv:2211.11256*.

[43] T. Kim and P. Vossen, "EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa," 2021, *arXiv:2108.12009*.

[44] J. Shor, A. Jansen, R. Maor, O. Lang, O. Tuval, F. de Chaumont Quitry, M. Tagliasacchi, I. Shavitt, D. Emanuel, and Y. Haviv, "Towards learning a universal non-semantic representation of speech," 2020, *arXiv:2002.12764*.

[45] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S. R. Jang, C.-C. Lee, and H.-Y. Lee, "EMO-SUPERB: An in-depth look at speech emotion recognition," 2024, *arXiv:2402.13018*.

[46] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008.

[47] F. Parllaku, "Speech-based artificial intelligence emotion biomarkers in frontotemporal dementia," M.S. thesis, Dept. Elect. Eng. omput. Sci., Massachusetts Inst. Technol., 2022.

[48] L. Cummings, "Describing the cookie theft picture: Sources of breakdown in Alzheimer's dementia," *Pragmatics Soc.*, vol. 10, no. 2, pp. 153–176, Jul. 2019.

[49] J. G. Švec and S. Granqvist, "Tutorial and guidelines on measurement of sound pressure level in voice and speech," *J. Speech, Lang., Hearing Res.*, vol. 61, no. 3, pp. 441–461, Mar. 2018.

[50] B. G. Schultz, S. Rojas, M. S. John, E. Kefalianos, and A. P. Vogel, "A cross-sectional study of perceptual and acoustic voice characteristics in healthy aging," *J. Voice*, vol. 37, no. 6, pp. 969.e23–969.e41, Nov. 2023.

[51] A. P. Vogel, M. L. Poole, H. Pemberton, M. W. J. Caverlé, F. M. C. Boonstra, E. Low, D. Darby, and A. Brodtmann, "Motor speech signature of behavioral variant frontotemporal dementia: Refining the phenotype," *Neurology*, vol. 89, no. 8, pp. 837–844, Aug. 2017.

[52] M. L. Gorno-Tempini, A. E. Hillis, S. Weintraub, A. Kertesz, M. Mendez, S. F. Cappa, J. M. Ogar, J. D. Rohrer, S. Black, and B. F. Boeve, "Classification of primary progressive aphasia and its variants," *Neurology*, vol. 76, no. 11, pp. 1006–1014, 2011.
[53] H. Goodglass, E. Kaplan, and S. Weintraub, *BDAE: The Boston Diagnostic Aphasia Examination*. Philadelphia, PA, USA: Ippincott Williams & Wilkins, 2001.
[54] B. McFee, C. Raffel, D. Liang, D. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "Librosa: Audio and music signal analysis in Python," in *Proc. 14th Python Sci. Conf.*, 2015, pp. 18–25.
[55] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," 2014, *arXiv:1404.5997*.
[56] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*.
[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, *arXiv:1512.03385*.
[59] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[60] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," *SciPy*, vol. 7, no. 1, 2010.
[61] P. J. Huber, "Robust estimation of a location parameter," in *Breakthroughs Statistics: Methodology Distribution*. Cham, Switzerland: Springer, 1992, pp. 492–518.
[62] A. M. Landes, "Prevalence of apathy, dysphoria, and depression in relation to dementia severity in Alzheimer's disease," *J. Neuropsychiatry*, vol. 17, no. 3, pp. 342–349, Aug. 2005.
[63] F. Mirakhori, M. Moafi, M. Milanifard, and H. Tahernia, "Diagnosis and treatment methods in Alzheimer's patients based on modern techniques: The orginal article," *J. Pharmaceutical Negative Results*, vol. 13, no. 1, pp. 1889–1907, Jan. 2022.
[64] Y. Gong, L. Yang, J. Zhang, Z. Chen, S. He, X. Zhang, and W. Zhang, "Using speech emotion recognition as a longitudinal biomarker for Alzheimer's disease," *Int. J. Biomed. Biol. Eng.*, vol. 17, no. 11, pp. 267–272, 2023.
[65] K. R. Scherer, "A cross-cultural investigation of emotion inferences from voice and speech: Implications for speech technology," in *Proc. Interspeech*, vol. 4, 2000, pp. 379–382.
[66] S. Kanwal, S. Asghar, A. Hussain, and A. Rafique, "Identifying the evidence of speech emotional dialects using artificial intelligence: A cross-cultural study," *PLoS ONE*, vol. 17, no. 3, Mar. 2022, Art. no. e0265199.
[67] C. Chakraborty, T. K. Dash, G. Panda, and S. S. Solanki, "Phase-based cepstral features for automatic speech emotion recognition of low resource Indian languages," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 2022, Sep. 2022.
[68] T. K. Dash, C. Chakraborty, S. Mahapatra, and G. Panda, "Mitigating information interruptions by COVID-19 face masks: A three-stage speech enhancement scheme," *IEEE Trans. Computat. Social Syst.*, pp. 1–10, Oct. 2022, doi: 10.1109/TCSS.2022.3210988.

YISHU GONG received the B.S. degree in physics and applied mathematics from UCLA, Los Angeles, CA, USA, and the M.S. degree in computer science and the Ph.D. degree in mathematics (focus on mathematical biology) from Duke University, Durham, NC, USA. She is currently the Manager of statistics with the Statistical and Quantitative Sciences Group, a part of the Data Science Institute, Takeda Pharmaceuticals Inc., Cambridge, MA, USA. Her research interests include utilizing quantitative methods to conduct digital biomarker analysis.

FJONA PARLLAKU was born in Tirana, Albania, in February 1998. She received the Bachelor of Science degree in electrical engineering and computer science, and mathematics from MIT, in 2021, and the Master of Engineering degree from MIT, in 2022, with a focus on artificial intelligence. She is currently a Data Scientist with 91life and a Teaching Assistant of an introductory course in AI in Harvard. Her research throughout her years at MIT has been heavily focused on using machine learning techniques in real life problems in various areas, mainly in the medical one.

KATERINA PLACEK received the B.A. degree in psychology from the St. Mary's College of Maryland, the M.S. degree in experimental psychology from the University of Oxford, and the Ph.D. degree in neuroscience from the University of Pennsylvania. She is currently the Senior Implementation Manager of the DHS Sensing and Measurement Group, Data Sciences Institute, Takeda Pharmaceuticals Inc. She is responsible for the evaluation, selection, and implementation of digital health tools in clinical trials, with a specific focus on neurologic, gastrointestinal, and oncologic indications.

MARCO VILELA received the B.Eng. degree in control engineering from the Federal University of Itajuba, the M.Sc. degree from the National Laboratory for Scientific Computing (LNCC), Brazil, and the Ph.D. degree from ITQB/New University of Lisbon, Portugal. He was a Visiting Student with The University of Texas MD Anderson Cancer Center and Georgia Tech, from 2006 to 2009, a Postdoctoral Fellow with Harvard Medical School, from 2010 to 2014, an Instructor with UT Southwestern, from 2014 to 2015, and a Senior Research Associate with the BrainGate Laboratory, Brown University, from 2015 to 2018. After a couple of years in a startup company, he joined Takeda Pharmaceuticals Inc., in 2020, involved on developing digital endpoints.

BRIAN HAREL received the Ph.D. degree in clinical psychology with a specialty in neuropsychology from the University of Connecticut. He is currently in internship with the Clinical Neuropsychology, Ann Arbor VA Healthcare System/University of Michigan Medical School, and a Postdoctoral Residency with the Clinical Neuropsychology, Department of Neurology, University of Iowa Hospitals and Clinics. He is a Clinical Neuropsychologist with clinical and research experience in academia and industry. He was previously with Pfizer, as the Associate Director Clinician of the Neuroscience and Pain Research Unit. He is also the Scientific Director-Neuropsychology of the Clinical Science Group, Neuroscience Therapeutic Area Unit (TAU).

ARTHUR SIMEN received the Ph.D./M.D. degree from the University of Chicago, Chicago IL, USA. He is currently the Executive Medical Director in Neuroscience of Takeda Pharmaceuticals Inc., Cambridge, MA, USA. Prior to Takeda Pharmaceuticals Inc., he was a Faculty Member with the School of Medicine, Yale University, and the Director of Pfizer. He is a Board Certified Physician and a Neurobiologist with a strong background in clinical and preclinical neuroscience and biostatistics, and has drug development experience that includes collaboration with discovery scientists on target discovery and preclinical biology and clinical development experience from phase I through phase III. He is passionate about the use of precision medicine to address unmet participant needs. His research interests include Alzheimer's disease, frontotemporal dementia, and cognitive aspects of Parkinson's disease.

BRIAN SUBIRANA received the M.B.A. degree from the MIT Sloan School of Management, Cambridge, MA, USA, and the Ph.D. degree in artificial intelligence (AI) from the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge. He is currently the Faculty Member of Massachusetts Institute of Technology and Harvard University, Cambridge, and a Full Professor of artificial intelligence with the EADA Business School, Barcelona, Spain. He has been with MIT, for over two decades, including being he Director of the MIT Auto-ID Laboratory, where the research reported here was conducted. His research interests include the crossroads of the Internet of Things (IoT) and artificial intelligence (AI).

AMY BRODTMANN is currently a Professor of neurology and leads the Cognitive Health Initiative, Department of Neuroscience, School of Translational Medicine, Monash University. She leads Cognitive Neurology Services, Eastern Cognitive Disorders Clinic, Box Hill Hospital, and the Melbourne Health Cognitive Neurology Service, Royal Melbourne Hospital. She was previously a Professor and the Co-Head of Dementia with the Florey Institute of Neuroscience and Mental Health, Melbourne Brain Centre, and a Consultant Neurologist with Austin Health.

ADAM VOGEL received the degree in psychology and speech science from the University of Queensland, Australia, and the Ph.D. degree in behavioural neuroscience from The University of Melbourne. He is currently a Professor and the Head of Speech Pathology with the School of Health Sciences, The University of Melbourne, and an Australian Research Council Future Fellow. He is a Humboldt Fellow with the Hertie Institute for Clinical Brain Research, Tübingen, Germany. Concurrently, he is the Chief Science Officer of Redenlab Inc., a neuroscience technology company using speech and language biometrics to enhance decision making in clinical trials. Redenlab Inc., work with pharmaceutical companies across 25 countries, and more than 250 sites globally. His research interests include improving speech, language, and swallowing function in people with progressive and acquired neurological conditions.

BRIAN TRACEY received the B.A. degree in physics from Kalamazoo College, Kalamazoo, MI, USA, and the Ph.D. degree in oceanographic engineering (focusing on acoustics) from the MIT/Woods Hole Joint Program, Cambridge, MA USA. He has also with medical device industry and was a Staff with the MIT Lincoln Laboratory. He is currently the Director of the Biostatistics in the Statistical and Quantitative Sciences Group, a part of the Data Science Institute, Takeda Pharmaceuticals Inc., Cambridge. In addition, he is a part-time Lecturer with Tufts University, Medford, MA, USA, where he previously was a Professor of the practice and performed research on imaging and biomedical signal processing. His research interests include development of signal processing, machine learning, and statistical methods for analyzing speech and wearable sensor data to gain insight into neurological and other disease.