frontiers | Research Topics

Linguistic biomarkers of neurological, cognitive, and psychiatric disorders: Verification, analytical validation, clinical validation, and machine learning

Edited by
Ratree Wayland, Si Chen and Kevin Tang

Published in
Frontiers in Psychology
Frontiers in Neuroscience

ISSN 1664-8714
ISBN 978-2-8325-5297-1
DOI 10.3389/978-2-8325-5297-1

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Linguistic biomarkers of neurological, cognitive, and psychiatric disorders: Verification, analytical validation, clinical validation, and machine learning

Topic editors

Ratree Wayland — University of Florida, United States
Si Chen — Hong Kong Polytechnic University, Hong Kong, SAR China
Kevin Tang — Heinrich Heine University of Düsseldorf, Germany

Citation

Wayland, R., Chen, S., Tang, K., eds. (2024). *Linguistic biomarkers of neurological, cognitive, and psychiatric disorders: Verification, analytical validation, clinical validation, and machine learning*. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-8325-5297-1

Table of contents

TYPE Editorial
PUBLISHED 26 July 2024
DOI 10.3389/fpsyg.2024.1454225

Editorial: Linguistic biomarkers of neurological, cognitive, and psychiatric disorders: verification, analytical validation, clinical validation, and machine learning

Ratree Wayland[1]*, Kevin Tang[1,2] and Si Chen[3]

[1]Department of Linguistics, College of Liberal Arts and Sciences, University of Florida, Gainesville, FL, United States, [2]Department of English Language and Linguistics, Institute of English and American Studies, Faculty of Arts and Humanities, Heinrich-Heine-Universität Düsseldorf, Düsseldorf, Germany, [3]Department of Chinese and Bilingual Studies, Faculty of Humanities, Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China

OPEN ACCESS

EDITED AND REVIEWED BY
Xiaolin Zhou,
Peking University, China

*CORRESPONDENCE
Ratree Wayland
✉ ratree@ufl.edu

RECEIVED 24 June 2024
ACCEPTED 12 July 2024
PUBLISHED 26 July 2024

CITATION
Wayland R, Tang K and Chen S (2024) Editorial: Linguistic biomarkers of neurological, cognitive, and psychiatric disorders: verification, analytical validation, clinical validation, and machine learning. *Front. Psychol.* 15:1454225.
doi: 10.3389/fpsyg.2024.1454225

KEYWORDS

neurodegenerative disease, autism, cognitive impairments, dementia, mental health stuttering, concussion, respiratory health

Editorial on the Research Topic
Linguistic biomarkers of neurological, cognitive, and psychiatric disorders: verification, analytical validation, clinical validation, and machine learning

Introduction

Speech production is a complex process involving the coordination of over 100 muscles across the respiratory, articulatory, and phonation systems. This intricate coordination makes speech a valuable source of biomarkers for various diseases. By analyzing speech production, we can gain insights into neuromuscular and psychological conditions, making it a powerful tool for the early detection and monitoring of these disorders as evidenced by the diverse studies in this Research Topic. These studies leverage and develop innovative methodologies to uncover the diagnostic potential of speech characteristics.

Nine Original Research articles were accepted in this Research Topic out of 17 submissions. With each paper having on average 5.2 authors, the interdisciplinary nature of this Research Topic is apparent. The nine articles cover seven broad types of disorders, including neurodegenerative diseases (Dash et al.; Roland et al.), neurodevelopmental disorders (Hong et al.), cognitive impairments (Oh et al.), psychological and emotional disorders (Cohen et al.; Chao et al.), respiratory health (Zeng et al.), concussion (Patel et al.) and stuttering (Barrett et al.). To illustrate what the nine articles cover, a word cloud (Figure 1) was generated showing the 200 most frequent words found in the abstracts of the articles.

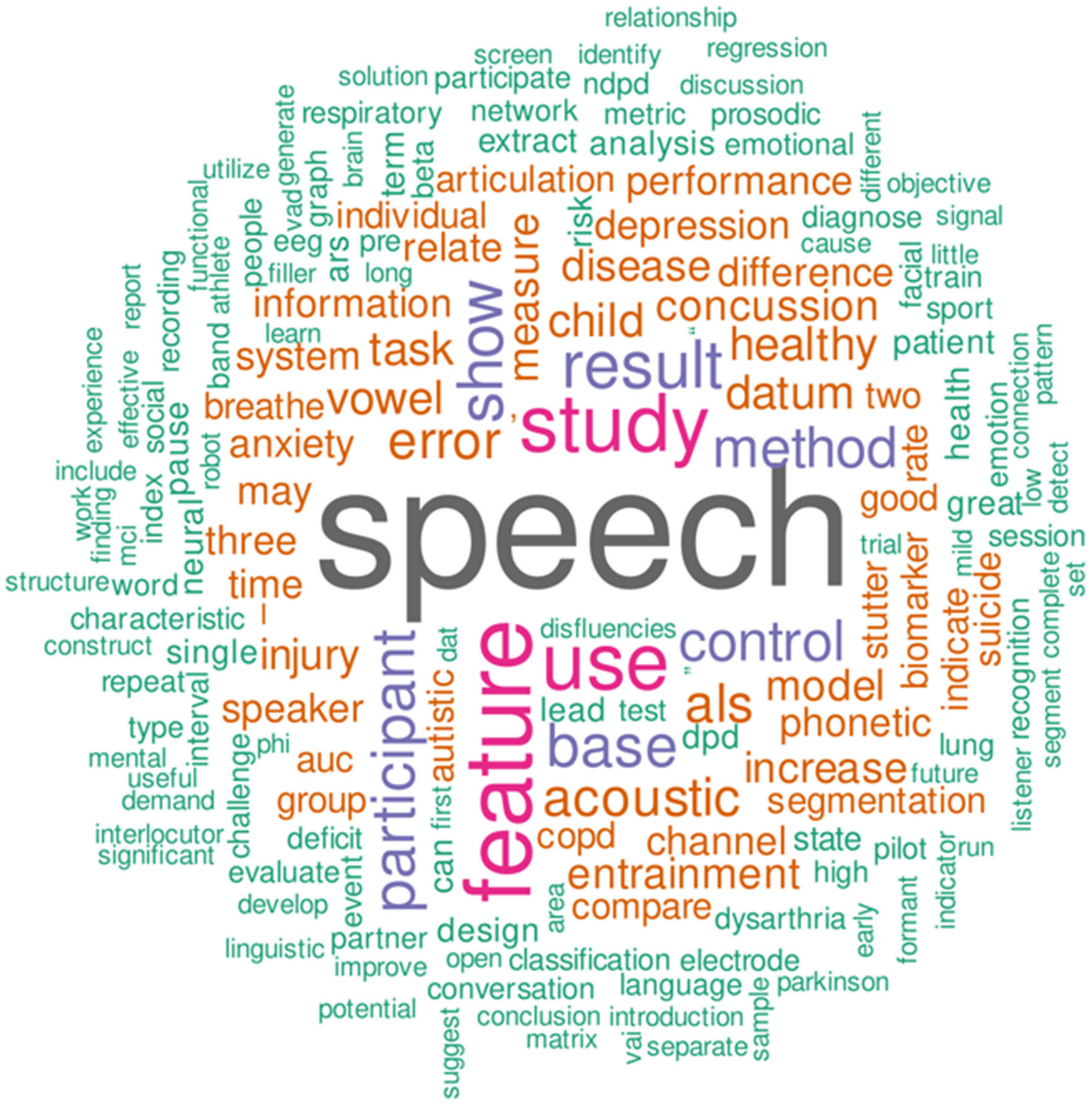


FIGURE 1
A word cloud which represents the 200 most frequent words found in the abstracts of the articles included in this special issue. The words were converted to lower case, English stop words were removed, and lemmatization was performed using the R libraries "tm" (Feinerer et al., 2008; Feinerer and Hornik, 2024) and "textstem" (Rinker, 2018). The resulting lemmas were visualized using the R library "wordcloud" (Fellows, 2018).

Neurodegenerative diseases

Dash et al. explore the use of magnetoencephalography (MEG) to identify neural biomarkers for Amyotrophic Lateral Sclerosis (ALS). By analyzing neuromagnetic patterns during speech tasks, their study identifies distinct beta band activity as a potential diagnostic marker, achieving high accuracy in single-trial classifications. Roland et al. focus on detecting early speech biomarkers of dysarthria in Parkinson's disease (PD) through vowel articulation analysis. Their use of vowel triangle areas (tVSA) and vowel articulation index (VAI) effectively distinguishes between dysarthric and non-dysarthric PD patients, highlighting the potential of speech analysis for early detection and differentiation in neurodegenerative diseases. These studies underscore the potential of advanced neural and acoustic analyses in identifying early, subtle markers of neurodegeneration.

Neurodevelopmental disorders

Hong et al. demonstrate that phonetic entrainment, where people adjust their speech to match their partner's phonetic features, is challenging for individuals with Autism Spectrum Disorder (ASD). Using a social robot to control speech variability during conversations, the study found autistic children matched their typically developing (TD) peers in vowel formants and mean fundamental frequency (f0) but struggled with f0 range entrainment. This highlights the potential of human-robot interactions for assessing phonetic entrainment in autistic children.

Cognitive impairments and dementia

Oh et al. focused on the differentiation of cognitive impairments and various forms of dementia through speech analysis. They investigate whether prosodic features can distinguish between Alzheimer's type dementia (DAT), vascular dementia (VaD), mild cognitive impairment (MCI), and healthy cognition. By identifying key features such as pitch, amplitude, rate, and syllable, they demonstrate the feasibility of using acoustic measures as diagnostic tools for cognitive conditions. This approach is complemented by listener perceptions of emotional prosody, which further validate the acoustic findings. These insights into speech characteristics offer a non-invasive and potentially scalable method for early diagnosis and differentiation of cognitive impairments.

Psychological and emotional disorders

Speech analysis also extends its utility to the realm of psychological and emotional disorders. Cohen et al. evaluate a multimodal dialog system (MDS) for characterizing mental states in individuals with depression, anxiety, and suicide risk. By integrating speech, language, and facial movement biomarkers, their system offers a comprehensive approach to remote patient monitoring. The ability to analyze multimodal data not only improves classification performance but also provides a scalable solution for ongoing mental health assessment. Chao et al. introduce a novel ResGAT emotion recognition framework, which combines residual networks and graph attention networks, to enhance emotion recognition from EEG data. This method effectively captures spatial and connection information, significantly improving the accuracy of emotion recognition. These studies highlight the potential of speech and multimodal analysis in identifying and monitoring psychological and emotional states, paving the way for more effective mental health interventions.

Speech and respiratory health

The link between speech and respiratory health is another critical area of exploration. Zeng et al. investigate how speech breathing can be linked to lung function in chronic respiratory diseases. Their study uses articulation tasks to challenge and quantify speech articulation and breathlessness. The increase in pause ratios over successive runs provides quantifiable evidence of respiratory demand, suggesting that speech tasks can effectively assess respiratory health. This approach offers a non-invasive method for monitoring chronic respiratory conditions, potentially leading to better disease management.

Speech and concussions

Speech analysis also shows potential in assessing neurological impacts from mild head injuries. Patel et al. analyze speech error rates in athletes post-concussion, revealing significant increases in pauses and time fillers. This study demonstrates that even mild head injuries can result in detectable speech changes, suggesting that speech analysis could serve as a diagnostic tool for concussions. The ability to identify subtle speech errors provides an additional layer of assessment for sports-related injuries, contributing to more comprehensive care for athletes.

Speech disorders

Finally, the application of speech analysis to detect and manage speech disorders is exemplified by Barrett et al.'s study on automatic recognition of stutters (ARS). By comparing event-based and interval-based segmentation methods, their research shows that event-based segmentation more effectively preserves stutter boundaries and types, leading to better ARS performance. This study emphasizes the importance of segmentation techniques in speech analysis and suggests that refined methods and larger datasets could further improve ARS systems. The findings point to the potential of automated speech analysis in supporting interventions for speech disorders, enhancing the ability to monitor and manage conditions like stuttering.

Conclusion

The studies presented in this Research Topic illustrate the potential of speech analysis as biomarkers for a range of neuromuscular and psychological disorders. The innovative methodologies and findings underscore the importance of further research in this field. By leveraging advanced acoustic, neural, and multimodal analyses, as well as machine learning and automatic speech recognition algorithms, researchers can enhance diagnostic accuracy and patient care, paving the way for early intervention and personalized treatment strategies. The preliminary nature of the findings of some studies calls for more research involving larger subject groups and patient populations with various diseases to validate the differential power of speech-based biomarkers across different conditions.

Author contributions

RW: Writing – review & editing, Writing – original draft. KT: Writing – review & editing, Writing – original draft, Visualization. SC: Writing – review & editing, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by NSF-National Science Foundation (SenSE) Award No. 2037266.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Feinerer, I., and Hornik, K. (2024). *tm: Text Mining Package.* R package version 0.7-13. Available online at: https://CRAN.R-project.org/package=tm

Feinerer, I., Hornik, K., and Meyer, D. (2008). Text mining infrastructure in R. *J. Stat. Softw.* 25, 1–54. doi: 10.18637/jss.v025.i05

Fellows, I. (2018). *wordcloud: Word Clouds.* R package version 2.6. Available online at: https://CRAN.R-project.org/package=wordcloud

Rinker, T. W. (2018). *textstem: Tools for Stemming and Lemmatizing Text Version 0.1.4.* Buffalo, NY. Available online at: http://github.com/trinker/textstem

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 07 March 2023
DOI 10.3389/fpsyg.2023.1135441

OPEN ACCESS

EDITED BY
Ratree Wayland,
University of Florida,
United States

REVIEWED BY
Edoardo Nicolò Aiello,
IRCCS Istituto Auxologico Italiano,
Italy
Anthony P. Salvatore,
University of Louisiana at Lafayette,
United States

*CORRESPONDENCE
Sona Patel
✉ Sona.Patel@shu.edu

SPECIALTY SECTION
This article was submitted to
Language Sciences,
a section of the journal
Frontiers in Psychology

RECEIVED 31 December 2022
ACCEPTED 13 February 2023
PUBLISHED 07 March 2023

CITATION
Patel S, Grabowski C, Dayalu V and
Testa AJ (2023) Speech error rates after a
sports-related concussion.
Front. Psychol. 14:1135441.
doi: 10.3389/fpsyg.2023.1135441

Speech error rates after a sports-related concussion

Sona Patel[1,2]*, Caryn Grabowski[1], Vikram Dayalu[1] and
Anthony J. Testa[3]

[1]Department of Speech-Language Pathology, Seton Hall University, Nutley, NJ, United States,
[2]Department of Medical Sciences, Hackensack Meridian School of Medicine, Nutley, NJ, United States,
[3]Center for Sports Medicine, Seton Hall University, South Orange, NJ, United States

Background: Alterations in speech have long been identified as indicators of various neurologic conditions including traumatic brain injury, neurodegenerative diseases, and stroke. The extent to which speech errors occur in milder brain injuries, such as sports-related concussions, is unknown. The present study examined speech error rates in student athletes after a sports-related concussion compared to pre-injury speech performance in order to determine the presence and relevant characteristics of changes in speech production in this less easily detected neurologic condition.

Methods: A within-subjects pre/post-injury design was used. A total of 359 Division I student athletes participated in pre-season baseline speech testing. Of these, 27 athletes (18–22years) who sustained a concussion also participated in speech testing in the days immediately following diagnosis of concussion. Picture description tasks were utilized to prompt connected speech samples. These samples were recorded and then transcribed for identification of errors and disfluencies. These were coded by two trained raters using a 6-category system that included 14 types of error metrics.

Results: Repeated measures analysis of variance was used to compare the difference in error rates at baseline and post-concussion. Results revealed significant increases in the speech error categories of pauses and time fillers (interjections/fillers). Additionally, regression analysis showed that a different pattern of errors and disfluencies occur after a sports-related concussion (primarily time fillers) compared to pre-injury (primarily pauses).

Conclusion: Results demonstrate that speech error rates increase following even mild head injuries, in particular, sports-related concussion. Furthermore, the speech error patterns driving this increase in speech errors, rate of pauses and interjections, are distinct features of this neurological injury, which is in contrast with more severe injuries that are marked by articulation errors and an overall reduction in verbal output. Future studies should consider speech as a diagnostic tool for concussion.

KEYWORDS

sports-related concussion, concussion, mild traumatic brain injury, speech, fluency, disfluencies, speech error rate, pausing

1. Introduction

Sports-related concussion (SRC) occurs as a result of impact to the head or neck during competitive or recreational athletic activities (Powell et al., 2021). SRC is a specific classification of concussion or mild traumatic brain injury (mTBI), broad terms used to define milder forms of traumatic brain injury (TBI) which result from insult in a closed-head nature due to linear and/or rotational forces to the head or neck (McCrory et al., 2013). SRCs are reported to occur at a rate of 4.13 per 10,000 athlete exposure (Chandran et al., 2022). SRCs are typically characterized by a range of physical and cognitive symptoms, often initially determined by self-report of symptoms (McCrory et al., 2017). There are clinical factors that differentiate SRC as a less substantial injury than mTBI, such as an abbreviated recovery period, typically less than 10 days, as compared to mTBI, where recovery can take 2–4 weeks (King, 2019). In contrast with more severe brain injuries which cause visible abnormalities detectable through neuroimaging, the nature and relative severity of damage in SRCs make detection difficult through currently available methods of neuroimaging. Traumatic brain injury is typically classified in terms of severity using the Glasgow Coma Scale and other standardized scales. A concussion can be similarly graded according to scales such as the Nelson grading system (Nelson et al., 1984) or the Colorado Grading System (Ommaya, 1985), however, 90% of concussions do not involve loss of consciousness and under 24 h of amnesia, resulting in a grading of 2 or less on the Nelson and 1 on the Colorado (Cantu, 2006). Detection and diagnosis therefore rely on a series of assessments spanning a host of domains including measures of physical symptoms (e.g., balance, visual disturbances, headache, fatigue) as well as some basic cognitive functions (e.g., concentration, memory; Broglio et al., 2014; Echemendia et al., 2017). Variance across domains can occur and must be appraised in aggregate and interpreted by the clinical provider to determine the presence of injury, often in relation to broad normative data. To date, no practically available diagnostic marker for concussion exists.

One possible diagnostic marker for concussion is speech errors, characterized by deviations in timing, articulatory precision, and fluency (Darley et al., 1969). The neurologic underpinnings of speech production cover an expansive range of structures and related functions within the brain, involving "feed-forward" and "feedback" pathways that drive the conversion of cognitive-linguistic thought to motor planning and ultimately to speech-motor movements (Murdoch, 2001; Wildgruber et al., 2001; Guenther, 2006; Hickok, 2012). This complex neural circuitry constantly monitors and updates speech output through internal and external feedback loops, optimizing accuracy of production with minimal speech errors in neurologically healthy speakers (Fox Tree, 1995). On the other hand, brain injury and advanced diseases can impact cognitive and sensorimotor components of the speech production process and result in a substantial increase in the number of errors when speaking.

Speech errors are widely accepted as hallmark sequelae of neurotrauma in conditions such as stroke, brain injury, amyotrophic lateral sclerosis (ALS), Parkinson's disease (PD), and multiple sclerosis (Yorkston, 1996; Hartelius et al., 2000; Tomik and Guiloff, 2010; Moro-Velazquez et al., 2021). These conditions can result in alterations in acoustic properties (Holmes et al., 2000; Rusz et al., 2011), articulatory precision (Karlsson and Hartelius, 2019; Karlsson et al., 2020), and measures of timing (Juste et al., 2018). The patterns of speech errors that occur can often predict the specific neurological condition, holding potential for use as biomarkers of various conditions of the central and peripheral nervous system. For example, various parameters of speech production have been identified as markers for individuals with focal cerebrovascular accidents, commonly referred to as "stroke." Individuals with right hemisphere stroke typically exhibit significant reductions in fundamental frequency range, which is especially apparent when expressing emotional tones, mainly "joy" and "anger" (Ross and Monnot, 2008; Guranski and Podemski, 2015; Patel et al., 2018). In addition, the prosodic quality of stress, comprised of multiple acoustic factors including pitch, intensity, vowel quality, and duration (Chrabaszcz et al., 2014), have been found to indicate cortical hemispheric effects. Balan and Gandour (1999) identified limitations in the ability of individuals with right hemisphere stroke to shift or adjust stress to the same degree as health controls. In addition, Vergis et al. (2014) identified significant difference in speech rate and vowel duration among individuals with left hemisphere stroke resulting in apraxia of speech and aphasia when compared to those with aphasia alone as well as healthy controls.

The specific neurological impacts of TBI vary based on the location and nature of the injury, including possible etiologies of hematoma, hemorrhage, and diffuse axonal injury (Mesfin and Taylor, 2017). Severe and moderate TBIs result in symptoms of motor speech impairment, such as dysarthria (Goozée et al., 2001; Solomon et al., 2001; Wang et al., 2005; McAuliffe et al., 2010; Kuruvilla et al., 2012) and occasionally apraxia of speech (Yadegari et al., 2014). Common characteristics include a slower articulation rate, smaller proportion of phonation time relative to sample duration, and larger total pause time (Wang et al., 2005). Other research in severe TBIs using analysis of passage readings has identified deficits in rate, resonance, and precision of consonants/overall intelligibility, variations in pitch and general stress patterns, as well as changes in phrase length among other aspects of speech production compared to healthy controls (Theodoros et al., 1994). Recent research with severely injured young children (6–10 years) suggests that during conversations there are decreases in pitch variation, the number of unique phonemes spoken, pause lengths, and increased variability of articulation rate (Noufi et al., 2019). All such findings indicate that speech deficits are a strong indicator of the presence and severity of TBI.

Deviations in various elements of speech production have also been identified as possible markers of injury in advanced stages of Alzheimer's disease, a disease process resulting from deviations in neural cellular health and integrity associated with abnormal protein deposits and metabolic processing ultimately resulting in diffuse failure of brain health and function (Mohandas et al., 2009). Speech characteristics of individuals with Alzheimer's disease include temporal changes, such as reduced rate of speech (phoneme or syllable production), increased pause or hesitation ratio, increased instances of repetitions, and increased frequency of within and between phrase pauses (Hoffmann et al., 2010; Fraser et al., 2016; Pistono et al., 2016; Slegers et al., 2018). Even in milder or earlier stages of neurodegeneration, such as early-stage Alzheimer's disease or mild cognitive impairment, differences in speech characteristics exist when compared to neurologically healthy controls. Analysis of connected speech samples in individuals with mild cognitive impairment has revealed alterations in articulation rate with and without hesitations, silent pauses, hesitation ratio, length of utterances, and pause per utterance when compared to healthy controls (Tóth et al., 2018).

Despite the scaled parallel in physical and cognitive symptoms commonly identified in concussion and more severe head injuries, consideration of the impacts of concussion on speech production has been limited. Changes in speech are typically not captured on commonly used symptom inventories for milder injuries and sports-related sideline assessments (Schatz et al., 2006; Asken et al., 2020). However, recent early evidence has shown significant alterations in rate of speech (Salvatore et al., 2019), acoustic features (Daudet et al., 2017), articulatory precision (Chong et al., 2021) and fluency (Robertson and Diaz, 2020; Rose et al., 2021; Toldi and Jones, 2021) in concussion. These preliminary findings suggest that further examination of speech changes in milder head injuries is necessary in a larger sample in order to establish the specific pattern of speech changes associated with concussion.

The goal of this study was to identify the speech changes that occur following an SRC using a comprehensive system for coding errors. Because error rates (disfluencies, misarticulations, speech errors) in typical speech production are low, small deviations from normal that might occur after a concussion may not be noticeable or identified as disordered because the errors do not interfere with functional communication, even though these errors may be systematically or consistently occurring. To investigate whether small deviations in speech fluency or patterns of speech errors exist, the present study analyzed speech samples of student athletes obtained in the days immediately following a concussion and compared these samples to their individual baseline recordings obtained prior to injury. A picture description task was used to obtain a more ecologically valid assessment of speech errors and disfluencies present. We expected student athletes with SRC to demonstrate an overall increase in the total number of speech errors compared to their individual pre-injury levels. Further, we anticipated observable patterns of errors that resemble those of more severe head injuries, albeit reduced in frequency.

2. Materials and methods

2.1. Participants

From 2018 through 2021, consenting Division I student athletes at Seton Hall University ($n = 359$) underwent speech testing concurrent with baseline testing that is completed annually as standard of care by Sports Medicine. All participants were proficient in English in lines with academic demands. All participants reported no history of vision, hearing, speech, or language issues, neurological disorders, or diagnosed psychiatric disorders (e.g., anxiety, depression, bipolar disorder). Additionally, it was confirmed at intake that participants were not experiencing upper or lower respiratory infections or other conditions that would impact speech and voice quality at the time of testing. Of the individuals tested, 27 athletes (11 males, 16 females; mean age: 18.3 years, range of 18–25 years) were determined to have a concussion by a Certified Athletic Trainer from Seton Hall University's Sports Medicine. All of the athletes diagnosed with SRC in this study had 0 min of loss of consciousness and under no reported amnesia. Injured participants represented eight sports teams at the University (see Table 1). All injured participants were initially evaluated as per the Sports Medicine protocol and were referred for testing once presence of concussion was determined. In some cases, due to latency of symptom onset or evolving presentation (such as headaches, light or sound sensitivity, sleep disturbance, among others), confirmation of the presence of concussion occurred up to 36 h after injury (Ruff et al., 2009). Participants with concussion then underwent post-injury speech testing, matching baseline testing procedures. All participants provided informed consent in accordance with the Hackensack Meridian Health Institutional Review Board on behalf of Seton Hall University.

2.2. Procedures

This study used a pre-test/post-test design where the same speech and language tasks were performed by participants before and after injury. Each test session was completed in a quiet study room reserved for student-athletes in under 20 min. As a part of baseline testing, all participants completed an intake questionnaire at the time of consent that included questions pertaining to demographic information and relevant medical history. Injured participants were tested in the days after being diagnosed with a concussion (mean = 2.83 days; range 0–6 days; see Table 1). Table 1 also provides the Standardized Assessment of Concussion (SAC) score post-injury as an indicator of concussion (out of 30; McCrea et al., 1998). Testing included a variety of speech elicitation tasks ranging in complexity and duration. Speech was recorded using an AKG head-worn microphone (HARMON International, Stamford, CT), which was routed through an Apollo audio interface with preamplifier (Universal Audio, Inc., Scotts Valley, CA) that was connected to a laptop computer dedicated for speech data collection. Audition software (Adobe, San José, CA) was used to record and store speech signals as.wav files onto the computer. Here we examined the audio files collected from one of the testing tasks, specifically the standard picture description task where participants were presented with a visual stimulus featuring a scene with multiple elements to elicit verbal output (e.g., "The Cookie Theft"; Goodglass, 1983; Shimada et al., 1998). Participants were instructed by the experimenter to "Take a look at this picture and explain to me what is happening. Tell me everything you can about the picture."

2.3. Speech error coding

To prepare the sound files for analysis (27 pre-injury or baseline samples, 27 post-concussion samples), extraneous speech that indicated acknowledgement of the task (e.g., "Okay") or the end of one's description (e.g., "That's about it") was removed. All sound files were transcribed in order to compute the number of syllables per sample. Next, error analysis was performed by two coders, who listened to the sound files to identify speech disfluencies and errors. Speech disfluencies and errors were classified as 1 of 14 types based on a combination of coding procedures commonly utilized in fluency and speech analysis (Lutz and Mallard, 1986; St. Louis et al., 1991; Ambrose and Yairi, 1999; Shriberg, 2001; Roberts et al., 2009; Sawyer and Yairi, 2010; Duffy, 2019) and marked on the transcript. Both coders were trained to identify and code 14 error types based on the specific definitions noted in Table 2. These 14 error types were also collapsed into 6 major categories based on shared features: pauses, revision/incomplete utterances, repetitions, articulation errors, time fillers, and prolongations. For example, all errors featuring repeated speech output were grouped into one larger category of "repetition" errors; all sound-level errors were grouped into the larger category of

TABLE 1 Demographic information of participants who sustained a concussion, including age, sex, sport, time of testing post-injury (days), and scores on the Standardized Assessment of Concussion (SAC).

Subject	Age	Sex	Sport	Days post-injury	SAC post-injury
s21	22	m	Men's Basketball	2	25
s38	19	f	Women's Soccer	1	28
s40	20	f	Women's Soccer	6	27
s45	23	m	Men's Soccer	2	26
s49	19	m	Men's Soccer	2	28
s61	18	m	Men's Soccer	<24 h	25
s63	18	f	Women's Soccer	4	25
s64	21	f	Women's Soccer	2	27
s73	21	m	Men's Soccer	4	28
s78	22	m	Men's Soccer	2	27
s108	19	f	Softball	3	29
s102	18	f	Women's Basketball	4	19
s103	22	f	Women's Basketball	1	26
s104	18	f	Women's Golf	2	23
s116	22	f	Women's Basketball	2	29
s141	20	f	Women's Soccer	4	24
s144	20	m	Baseball	5	29
s219	18	m	Baseball	3	26
s227	20	f	Women's Basketball	3	26
s231	18	f	Women's Basketball	3	28
s233	24	m	Men's Soccer	2	27
s241	20	f	Women's Soccer	6	27
s244	25	m	Men's Soccer	2	22
s248	18	m	Baseball	2	23
s270	19	f	Softball	1	29
s315	20	f	Volleyball	3	28
s321	18	f	Volleyball	3	29

"articulation errors." Both coders converged on the location and type of each error. The coders were blinded to the subject and condition when coding the speech samples. Reliability was assessed on approximately 15% of subjects (Corey and Cuddapah, 2008). Inter-rater reliability was calculated as the Pearson's correlation between raters for each error category. Inter-rater reliability across error categories was acceptable (greater than.81; McHugh, 2012): total errors = 0.98, articulation = n/a (no errors across samples tested), prolongation = 0.86, pause = 0.94, time fillers = 0.98, revision/incomplete = 0.92, and repetition = 1.0.

2.4. Statistical analyses

Error rates were computed for each speech error type of each sample in order to normalize error totals to the amount of speech produced (number of errors divided by the number of syllables). Since this study sought to examine changes in the within-subjects factor of time (baseline, concussion), a repeated measures analysis was required. Kolmogorov-Smirnoff tests of normality were significant ($p < 0.05$) for all parameters except fillers and pauses and the larger categories of time fillers, total dysfluency, and number of syllables at baseline. Results were similar after concussion, in addition to a lack of significance ($p > 0.05$) for interjections, indicating deviations from normality for most parameters. Examination of the skewness and kurtosis values revealed larger values than the standard error for either the baseline or concussion data for each parameter, indicating that assumptions of homoscedasticity also appear to have not been fully met. Thus, non-parametric Friedman tests were performed in SPSS (IBM SPSS Statistics v.28, Chicago, IL) on the error rates for the number of syllables produced, the total error rate, each of the six major error categories, and the 14 individual error types.

Next, stepwise regressions with bidirectional selection were performed on the 14 error types separately at baseline and after concussion to determine the extent that these variables best captured the overall error rate. Bidirectional selection involves a mixture of the forward and backward procedures in which the variable that explained the most variance in the total error rate was entered into the model first (entry criteria: probably of $f = 0.05$), followed by the variable that explained most of the residual variance, resulting in a set of variables with

TABLE 2 Coding criteria for speech errors and disfluencies within six major error categories.

Error/Disfluency category	Definition
Articulation	*Sound-level errors in articulation that include distortions, additions, omissions, substitutions*
Substitution	Any sound substitution
Distortion	Any sound-level distortion
Addition	A sound that is added
Omission	A sound that is omitted from a word
Pause	*Pauses greater than 250 ms*
Prolongation	*Sounds or syllables extended in duration more than 250 ms*
Repetition	*Any utterance (sound, word, phrase) that is repeated*
Part-word	Repetition of one or more phonemes within a word
Single-syllable whole-word	Repetition of a single syllable word
Multisyllable whole-word	Repetition of a word with two or more syllables
Phrase	Repetition of a phrase, i.e., a connected string of words
Revision/Incomplete	*A change or correction of an utterance(s) that did not convey a complete thought*
Revision	Modifications to output at a syllable, word, or phrase level
Incomplete segment	Utterance terminated abruptly or does not convey a complete thought
Time Fillers	*Extraneous sounds, words, or phrases that do not contribute to the meaning of the utterance*
Interjection	Words/phrases that are syntactically appropriate but do not add to the intended message (e.g., "So you know…," "I guess")
Filler	Sounds or "non-words" that add no meaning to the intended message (e.g., "um" and "uhh")

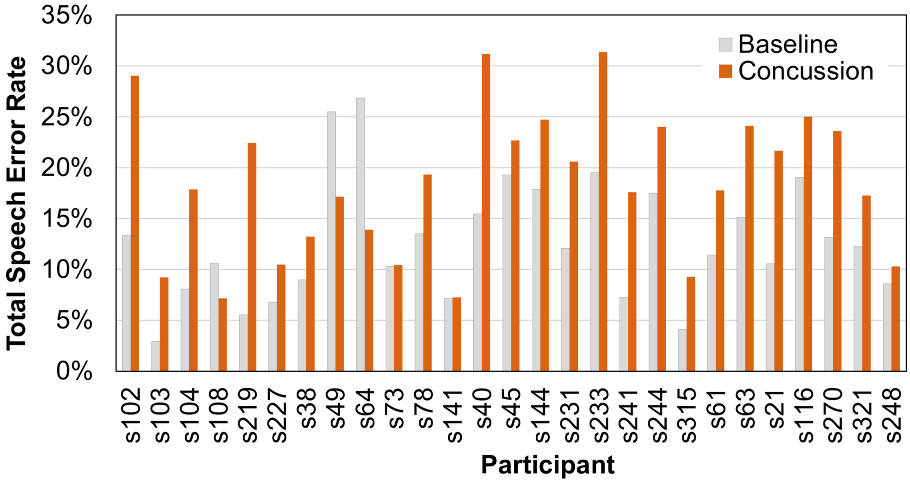


FIGURE 1
Total speech error rates (percent) for individual participants before and after a sports-related concussion.

the largest regression coefficients for inclusion in the model (Snyder, 1991). Such procedures can be advantageous for identifying the primary contributors when the number of parameters is small, thus resulting in a model with the smallest number of variables (Lewis, 2007). At each step, the predictors that were no longer significant were removed (removal criteria: probability of $f = 0.1$). Variables that accounted for more than 3% of the variance in the total error rate are reported.

3. Results

Results showed no significant difference in the average number of syllables produced after a concussion (mean or $M = 99.4$; standard deviation or $SD = 43.4$) compared to baseline ($M = 97.7$; $SD = 41.1$) at the $\alpha = 0.05$ level ($X^2 = 0.333$, $p = 0.564$). Nevertheless, individual differences in the number of syllables produced by each person existed. Therefore, the number of errors within each error type were normalized to the number of syllables, resulting in an error rate for each error type. The total error rate was significantly different between baseline and concussion samples ($X^2 = 16.333$, $p < 0.001$). The percentage of speech errors increased after sustaining a concussion ($M = 18.5\%$; $SD = 0.07$) compared to baseline ($M = 12.7\%$; $SD = 0.06$). Individual pre- and post-injury error scores are shown for each participant in Figure 1.

Next, changes in the number of errors and disfluencies in the six error categories were examined (see Figure 2). Friedman tests of the error rates showed significant differences ($\alpha = 0.05$) in the pause category ($X^2 = 10.704$, $p = 0.001$) and the time filler category ($X^2 = 19.593$, $p < 0.001$), with higher error rates occurring after a concussion (pauses: $M_{pre} = 6.8\%$, $M_{post} = 9.7\%$; time fillers: $M_{pre} = 2.9\%$, $M_{post} = 5.7\%$). No significant changes were found in articulation errors ($X^2 = 2.778$, $p = 0.096$), prolongations ($X^2 = 1.190$, $p = 0.275$), repetitions ($X^2 = 0.111$, $p = 0.739$), or revisions/incompletes ($X^2 = 0.25$, $p = 0.617$).

As the time filler, articulation error, repetition, and revision/incomplete categories consisted of multiple parameters, additional Friedman tests were performed to examine changes in particular error types. Results showed significant differences in the rate of interjections after a concussion ($p < 0.05$). The change in fillers approached significance ($p = 0.072$). Results are shown in Table 3.

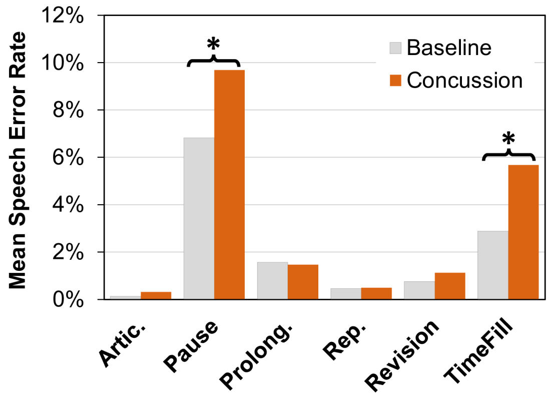


FIGURE 2
Mean speech error rate (percent) for articulation errors (Artic.), pauses, prolongations, (Prolong.), repetitions (Rep.), revisions, and time fillers (TimeFill) at baseline and after concussion. A significant difference between conditions (α=0.05 level) is indicated by an asterisk (*).

Stepwise regressions of the 14 error types at baseline showed that the total error rate was primarily driven by pauses [$R^2 = 0.628$, $\Delta F(1,25) = 42.162$, $p < 0.001$]. Prolongations accounted for an additional 28.1% of the variance [$R^2 = 0.909$, $\Delta F(1,24) = 74.260$, $p < 0.001$] followed by time fillers, which accounted for an additional 3.1% of the variance [$R^2 = 0.940$, $\Delta F(1,23) = 12.130$, $p = 0.002$]. In contrast, the total error rate after a concussion was primarily driven by time fillers [$R^2 = 0.707$, $\Delta F(1,25) = 60.252$, $p < 0.001$]. Pauses accounted for an additional 18.4% of the variance [$R^2 = 0.891$, $\Delta F(1,24) = 40.419$, $p < 0.001$] followed by prolongations, which accounted for an additional 5.3% of the variance [$R^2 = 0.944$, $\Delta F(1,23) = 21.882$, $p < 0.001$]. Addition of a fourth variable accounted for less than 3% of additional variance. Despite significant p-values for additional parameters, overfitting of models can produce misleading results. We decided to exclude variables that were contributing 3% of the variance to avoid over-fitting and simply report on the major factors contributing to the model. Such procedures have been used in prior work (Patel and Shrivastav, 2011).

4. Discussion

Changes in the characteristics of speech production including the presence of errors or deviations from typical are known to occur across various neurological conditions, including moderate and severe brain injury. However, it is not known whether the patterns of speech changes that occur in milder forms of brain injury such as SRC

TABLE 3 Results of Friedman tests of speech error/disfluency rates at the α=0.05 level for baseline compared to concussion.

Error/Disfluency category	df	X^2	p
Articulation	1	2.778	0.096
Substitution	1	–	–
Distortion	1	2.667	0.102
Addition	1	0	1.000
Omission	1	2.000	0.157
Pause	1	10.704	0.001*
Prolongation	1	1.190	0.275
Repetition	1	0.111	0.739
Part-word	1	1.286	0.257
Single-syllable whole-word	1	2.667	0.102
Multisyllable whole-word	1	1.000	0.317
Phrase	1	0.333	0.564
Revision/Incomplete	1	0.25	0.617
Revision	1	1.471	0.225
Incomplete segment	1	0.333	0.564
Time Fillers	1	19.593	<0.001*
Interjection	1	22.154	<0.001*
Filler	1	3.240	0.072

*There were no occurrences of sound substitutions for any subject.

resemble more severe forms of brain injury or whether distinct errors patterns reflective of SRC exist. In this study we examined speech error patterns in Division I college athletes within 6 days following a SRC, with a prediction that the total number of speech errors and dysfluencies would increase after a concussion compared to individual pre-injury levels. The availability of individual baseline measures is uncommon in brain injury and is advantageous as it allows more sensitive identification of error trends. As predicted, within-subject comparisons demonstrated a significant increase in the speech error rate after an SRC. To the best of our knowledge, this is the first study to demonstrate increases in speech error rates using a comprehensive system for capturing errors across domains of articulation, fluency, and timing in a large sample.

Our second prediction was that error types in SRC would be similar to those in more severe TBIs, specifically in the manifestation of articulation errors, reduced verbal output, and increased frequency of pausing (Power et al., 2020). In order to evaluate this prediction, errors were coded based on a classification system comprised of six major categories representing a total of 14 error types. Two of the six error categories, namely, number of pauses and time fillers, increased significantly after a concussion. The "pause" category captured any period of silence greater than 250 ms. Previous research has shown an increase in pausing, particularly pause duration, in TBI and even in health individuals under conditions of increased cognitive demand (Wang et al., 2005; Khawaja et al., 2008; Noufi et al., 2019). In line with these findings, the results of the present study in SRC showed that the number of pauses increased in milder head injuries. One other study by Banks et al. (2021) showed a similar pattern of results using a different task, namely that the time interval between syllables in a diadochokinetic speech task (repeated syllable production) was longer than the time interval in healthy controls. Despite the observed increase in the number of pauses in the present study, the number of syllables produced after a concussion was not significantly different from pre-injury baselines. In other words, the increased number of pauses contributed to a lengthening of the overall duration of the speech sample without a reduction in the total verbal output. These findings are in contrast with TBI research (more severe injuries than SRC) that shows a decrease in utterance length (Stubbs et al., 2018).

Although the total number of syllables did not reduce after an SRC, the verbal output might have been reduced in overall complexity, as indicated by a significant increase in the number of "time fillers" in the present study. The "time fillers" category examined in the present study included two error types that capture additional sounds, words or phrases that do not contribute to the sentence structure or meaning, namely interjections and fillers (see Table 1). Fillers differ from interjections in that they are extraneous sounds or non-words (Corley and Stewart, 2008), while interjections are extraneous words or phrases. In the present study, the number of interjections significantly increased but not the number of fillers, although they were trending. In other words, fillers occurred frequently prior to injury and continued to increase after injury. Both error types functionally serve to maintain continuity of connected speech production while accommodating increased demands on planning intended speech (Clark and Fox Tree, 2002).

The distribution of errors was also examined in athletes before any head injury to determine whether the pattern of errors changed after SRC. Results from the regression analysis of the total speech error rate at baseline by the six error categories showed that pre-injury errors primarily consist of pauses, followed by prolongations. Pausing is a behavior that allows time for linguistic ideation, motor planning, and execution for coherent and fluent speech production and a certain number of pauses are expected to occur when speaking. Prolongations slow down the rate of speaking allowing for thinking while still creating continuity/connectivity in verbal output as time fillers do, but in a subtle, less disruptive manner, that can in many cases can be perceived as typical alterations in stress patterning that occur in discourse. In contrast, errors after an SRC were primarily time fillers, followed by pauses. This suggests that individuals with SRC use time fillers (particularly interjections) to allow for seamless transition of thoughts while speaking more frequently than silent periods (pausing). This finding showcases that individuals with SRC may manifest a unique set of compensatory mechanisms to deal with the underlying neural insult. Further, in comparison to speech error data from more severe brain injuries, the use of time fillers is a unique communicative e pattern that may be available only to individuals with milder concussions.

The increased number of pauses and time fillers in individuals with SRC suggests inefficiencies in the planning of linguistic content, which are rooted in the cognitive domains of attention, memory and higher order executive functions (King et al., 2006; Crawford et al., 2007; Murray, 2012; Obermeyer et al., 2020). Concussion is known to impact the areas of cognition associated with the planning of speech output, specifically executive functions, attention, and memory (Covassin and Elbin, 2010; Kaltiainen et al., 2019). Incidentally, these cognitive linguistic functions primarily occur in cortical regions where axonal sheering and other trauma occur in SRC (Shaw, 2002). It is therefore likely that the increased rate of pauses and time fillers identified in this study is an indication of underlying cognitive dysfunction.

The relationship of these speech error categories and cognitive linguistic function is seen in typical, non-injured adults, where the number of speech errors increases with higher processing, cognitive load, and cognitive ability (Bortfeld et al., 2001; Shriberg, 2001; Engelhardt et al., 2013). The relationship of cognitive impairment and speech errors has also been established across various clinical populations. Both Power et al. (2020) and Smith et al. (2018) have demonstrated that the time fillers category (interjections and fillers) is associated with cognitive deficits in adults with Parkinson's disease. Other works have demonstrated higher-level relationships between speech output and cognitive-linguistic function, where individuals with left hemisphere stroke experience impairments in accessing the lexical-semantic network resulting in long pauses and decreased speech fluency (Yee et al., 2008; Lerman et al., 2020). In Alzheimer's disease, studies note decreases in quantity of verbal output, decreases in richness of content, and increases in semantic errors as hallmark changes representative of the disease (Kavé and Levy, 2003; de Lira et al., 2014; Slegers et al., 2018). Milder forms of neurological decline such as mild cognitive impairment and early dementia have also shown impacts on verbal fluency and speech output, demonstrating differences in linguistic properties compared to healthy controls (Beltrami et al., 2018). The sum of findings across healthy individuals and those with various neurological conditions demonstrates that speech changes are linked to cognitive processes.

In the present study, the error categories of revision/incomplete, repetition, and prolongation did not show significant changes in SRC. The categories of revision/incomplete, repetition, and prolongation have been reported error types in TBI, although the incidence is extremely low and often connected with acquired stuttering disorders (Jokel et al., 2007). Some studies report stuttering-like behaviors, including speech hesitations, brief blocks, rapid repetitions, and occasional prolongations after TBI, in addition to interjections, silent pauses, broken words, revisions and starters (Lundie et al., 2014; Roth et al., 2015). These error categories, more stuttering-like in nature, may therefore be associated with more severe injuries or concomitant conditions (e.g., post-traumatic stress disorder) not typically present in SRC (Lundgren et al., 2010; Norman et al., 2018).

Finally, the error category of articulation also failed to show significant differences between baseline and concussion conditions. Articulation errors were predicted to contribute significantly to the error patterns in SRC as it is a common issue in more severe forms of neurotrauma and neurologic disease. Articulation errors are highly prevalent in TBI, as dysarthria, or speech dysfunction due to changes in muscle strength, range-of-motion, and coordination, occurs in up to 60% of individuals with TBI in acute phases of recovery (Yorkston, 1996). Most forms of dysarthria associated with articulatory imprecision and related errors result from insult to subcortical brain regions or peripheral nerve damage (Duffy, 2019). In the case of SRC it is therefore likely that mild cortical level trauma associated with axonal sheering does not yield impacts to the neuromuscular substrates of speech production that would amount to causing errors of articulation. This further supports the notion that speech errors found in SRC, pauses and time fillers, are associated with cognitive-linguistic dysfunction rooted in insult to cortical regions of the brain.

One limitation of the current study is the inclusion of the mildest of mild cases. In order to provide the best care for student athletes, the health team identified all possible cases of concussion that might have occurred. In other words, anyone who had sustained an impact to the head underwent sideline testing for symptoms of concussion. Speech evaluations were performed on all such cases, which may have resulted in a few referrals where symptoms of concussion were minimal. In the future, these will be controlled by setting a minimum symptom score as part of the inclusion criteria. In addition, future work should examine other factors that may influence changes in disfluency, including cognitive-linguistic function, severity of injury, etc.

5. Conclusion

The purpose of this study was to determine whether changes in speech error patterns exist in the days following a sports-related concussion compared to pre-season baseline measures. Our findings suggest the presence of increased speech errors in SRC. Specifically, significantly increased rates of pauses and time fillers were observed. Therefore, speech errors serve as a measurable marker of SRC that is not typically considered in current methods of clinical evaluation. The error patterns in the present study differ from the patterns of speech changes reported in the literature for other types of neurologic disorders including severe TBI. Thus, speech changes may serve to indicate the presence of concussion or milder forms of TBI. Further work must be done to understand the relationship of speech errors in SRC to higher-order cognitive functions as well as other symptom measures of SRC.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by Hackensack University Medical Center Institutional Review Board on behalf of Seton Hall University. The patients/participants provided their written informed consent to participate in this study.

Author contributions

SP and AT contributed to the conception and design of this study. SP, AT, and CG contributed to the data collection. CG and SP wrote the first draft of the manuscript. VD and SP contributed to the design of the analysis. SP and CG performed the speech analysis. SP performed the statistical analysis. AT and VD provided feedback on different drafts of this manuscript. All authors contributed to the article and approved the submitted version.

Funding

Research reported in this publication was partially supported by a New Jersey Commission on Brain Injury Research Pilot Grant (CBIR22PIL018) and an Undergraduate Research Award from Seton Hall University.

Acknowledgments

The authors would like to thank Ji Sook Ahn and the athletics trainers Caroline Brennan, Deja Craig-Lucien, Mercedes Cunningham, Catherine Lass, Kaitlin Kelly, Gabriel Scher, and Nicholas Schulman for their collective help and support in coordinating student athlete testing, and the student athletes for their time and efforts in participation in this study. A preprint of this article was previously made available (Patel et al., 2021).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ambrose, N. G., and Yairi, E. (1999). Normative disfluency data for early childhood stuttering. *J. Speech Lang. Hear. Res.* 42, 895–909. doi: 10.1044/jslhr.4204.895

Asken, B. M., Houck, Z. M., Bauer, R. M., and Clugston, J. R. (2020). SCAT5 vs. SCAT3 symptom reporting differences and convergent validity in collegiate athletes. *Arch. Clin. Neuropsychol.* 35, 291–301. doi: 10.1093/arclin/acz007

Balan, A., and Gandour, J. (1999). Effect of sentence length on the production of linguistic stress by left- and right-hemisphere-damaged patients. *Brain Lang.* 67, 73–94. doi: 10.1006/brln.1998.2035

Banks, R. E., Beal, D. S., and Hunter, E. J. (2021). Sports related concussion impacts speech rate and muscle physiology. *Brain Inj.* 35, 1275–1283. doi: 10.1080/02699052.2021.1972150

Beltrami, D., Gagliardi, G., Rossini Favretti, R., Ghidoni, E., Tamburini, F., and Calzà, L. (2018). Speech analysis by natural language processing techniques: a possible tool for very early detection of cognitive decline? *Front. Aging Neurosci.* 10:369. doi: 10.3389/fnagi.2018.00369

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. (2001). Disfluency rates in conversation: effects of age. *Lang. Speech* 44, 123–147. doi: 10.1177/00238309010440020101

Broglio, S. P., Cantu, R. C., Gioia, G. A., Guskiewicz, K. M., Kutcher, J., Palm, M., et al. (2014). National Athletic Trainers' association position statement: management of sport concussion. *J. Athl. Train.* 49, 245–265. doi: 10.4085/1062-6050-49.1.07

Cantu, R. (2006). "Concussion classification: ongoing controversy" in *Foundations of Sport-Related Brain Injuries* (Berlin: Springer), 87–110.

Chandran, A., Boltz, A. J., Morris, S. N., Robison, H. J., Nedimyer, A. K., Collins, C. L., et al. (2022). Epidemiology of concussions in National Collegiate Athletic Association (NCAA) sports: 2014/15-2018/19. *Am. J. Sports Med.* 50, 526–536. doi: 10.1177/03635465211060340

Chong, C., Zhang, J., Li, J., Wu, T., Dumkrieger, G., Nikolova, S., et al. (2021). Altered speech patterns in subjects with post-traumatic headache due to mild traumatic brain injury. *In Rev.*, Preprint. doi: 10.21203/rs.3.rs-603443/v1

Chrabaszcz, A., Winn, M., Lin, C. Y., and Idsardi, W. J. (2014). Acoustic cues to perception of word stress by English, mandarin, and Russian speakers. *J. Speech Lang. Hear. Res.* 57, 1468–1479. doi: 10.1044/2014_JSLHR-L-13-0279

Clark, H. H., and Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition* 84, 73–111. doi: 10.1016/S0010-0277(02)00017-3

Corey, D. M., and Cuddapah, V. A. (2008). Delayed auditory feedback effects during reading and conversation tasks: gender differences in fluent adults. *J. Fluen. Disord.* 33, 291–305. doi: 10.1016/j.jfludis.2008.12.001

Corley, M., and Stewart, O. (2008). Hesitation disfluencies in spontaneous speech: the meaning of *um*. *Lang. Linguistics Compass* 2, 589–602. doi: 10.1111/j.1749-818X.2008.00068.x

Covassin, T., and Elbin, R. J. (2010). The cognitive effects and decrements following concussion. *Open Access J. Sports Med.* 1, 55–61. doi: 10.2147/oajsm.s6919

Crawford, M. A., Knight, R. G., and Alsop, B. L. (2007). Speed of word retrieval in postconcussion syndrome. *J. Int. Neuropsychol. Soc.* 13, 178–182. doi: 10.1017/S135561770707021X

Darley, F. L., Aronson, A. E., and Brown, J. R. (1969). Clusters of deviant speech dimensions in the Dysarthrias. *J. Speech Hear. Res.* 12, 462–496. doi: 10.1044/jshr.1203.462

Daudet, L., Yadav, N., Perez, M., Poellabauer, C., Schneider, S., and Huebner, A. (2017). Portable mTBI assessment using temporal and frequency analysis of speech. *IEEE J. Biomed. Health Inform.* 21, 496–506. doi: 10.1109/JBHI.2016.2633509

de Lira, J. O., Minett, T. S. C., Bertolucci, P. H. F., and Ortiz, K. Z. (2014). Analysis of word number and content in discourse of patients with mild to moderate Alzheimer's disease. *Dement. Neuropsychol.* 8, 260–265. doi: 10.1590/S1980-57642014DN83000010

Duffy, J. R. (2019). *Motor Speech Disorders E-book: Substrates, Differential Diagnosis, and Management*. London: Elsevier Health Sciences.

Echemendia, R. J., Meeuwisse, W., McCrory, P., Davis, G. A., Putukian, M., Leddy, J., et al. (2017). The sport concussion assessment tool 5th edition (SCAT5): background and rationale. *Br. J. Sports Med.* 51, 848–850. doi: 10.1136/bjsports-2017-097506

Engelhardt, P., Nigg, J., and Ferreira, F. (2013). Is the fluency of language outputs related to individual differences in intelligence and executive function? *Acta Psychol.* 144, 424–432. doi: 10.1016/j.actpsy.2013.08.002

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *J. Mem. Lang.* 34, 709–738. doi: 10.1006/jmla.1995.1032

Fraser, K. C., Meltzer, J. A., and Rudzicz, F. (2016). Linguistic features identify Alzheimer's disease in narrative speech. *J. Alzheimers Dis.* 49, 407–422. doi: 10.3233/JAD-150520

Goodglass, H. (1983). *Goodglass and Kaplan, the Assessment of Aphasia and Related Disorders. Stimulus Cards*, *2nd Edn*, Lea & Febiger, Philadelphia.

Goozée, J. V., Murdoch, B. E., and Theodoros, D. G. (2001). Physiological assessment of tongue function in dysarthria following traumatic brain injury. *Logoped. Phoniatr. Vocol.* 26, 51–65. doi: 10.1080/140154301753207421

Guenther, F. H. (2006). Cortical interactions underlying the production of speech sounds. *J. Commun. Disord.* 39, 350–365. doi: 10.1016/j.jcomdis.2006.06.013

Guranski, K., and Podemski, R. (2015). Emotional prosody expression in acoustic analysis in patients with right hemisphere ischemic stroke. *Neurol. Neurochir. Pol.* 49, 113–120. doi: 10.1016/j.pjnns.2015.03.004

Hartelius, L., Runmarker, B., and Andersen, O. (2000). Prevalence and characteristics of dysarthria in a multiple-sclerosis incidence cohort: relation to neurological data. *Folia Phoniatr. Logop.* 52, 160–177. doi: 10.1159/000021531

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nat. Rev. Neurosci.* 13, 135–145. doi: 10.1038/nrn3158

Hoffmann, I., Nemeth, D., Dye, C. D., Pákáski, M., Irinyi, T., and Kálmán, J. (2010). Temporal parameters of spontaneous speech in Alzheimer's disease. *Int. J. Speech Lang. Pathol.* 12, 29–34. doi: 10.3109/17549500903137256

Holmes, R. J., Oates, J. M., Phyland, D. J., and Hughes, A. J. (2000). Voice characteristics in the progression of Parkinson's disease. *Int. J. Lang. Commun. Disord.* 35, 407–418. doi: 10.1080/136828200410654

Jokel, R., Nil, L. D., and Sharpe, K. (2007). Speech disfluencies in adults with neurogenic stuttering associated with stroke and traumatic brain injury. *J. Med. Speech Lang. Pathol.* 15, 243–262.

Juste, F. S., Sassi, F. C., Costa, J. B., and de Andrade, C. R. F. (2018). Frequency of speech disruptions in Parkinson's disease and developmental stuttering: a comparison among speech tasks. *PLoS One* 13:e0199054. doi: 10.1371/journal.pone.0199054

Kaltiainen, H., Liljeström, M., Helle, L., Salo, A., Hietanen, M., Renvall, H., et al. (2019). Mild traumatic brain injury affects cognitive processing and modifies oscillatory brain activity during attentional tasks. *J. Neurotrauma* 36, 2222–2232. doi: 10.1089/neu.2018.6306

Karlsson, F., and Hartelius, L. (2019). How well does diadochokinetic task performance predict articulatory imprecision? Differentiating individuals with Parkinson's disease from control subjects. *FPL* 71, 251–260. doi: 10.1159/000498851

Karlsson, F., Schalling, E., Laakso, K., Johansson, K., and Hartelius, L. (2020). Assessment of speech impairment in patients with Parkinson's disease from acoustic quantifications of oral diadochokinetic sequences. *J. Acoust. Soc. Am.* 147, 839–851. doi: 10.1121/10.0000581

Kavé, G., and Levy, Y. (2003). Morphology in picture descriptions provided by persons with Alzheimer's disease. *J. Speech Lang. Hear. Res.* 46, 341–352. doi: 10.1044/1092-4388(2003/027)

Khawaja, M. A., Ruiz, N., and Chen, F. (2008). Think before you talk: an empirical study of relationship between speech pauses and cognitive load. In: Proceedings of the 20th Australasian conference on computer-human interaction: Designing for habitus and habitat, OZCHI'08, Association for Computing Machinery. New York, NY, 335–338.

King, N. S. (2019). 'Mild traumatic brain injury' and 'sport-related concussion': different languages and mixed messages? *Brain Inj.* 33, 1556–1563. doi: 10.1080/02699052.2019.1655794

King, K. A., Hough, M. S., Walker, M. M., Rastatter, M., and Holbert, D. (2006). Mild traumatic brain injury: effects on naming in word retrieval and discourse. *Brain Inj.* 20, 725–732. doi: 10.1080/02699050600743824

Kuruvilla, M., Murdoch, B., and Goozee, J. (2012). A kinematic investigation of speaking rate changes after traumatic brain injury. *J. Med. Speech Lang. Pathol.* 20, 9–18.

Lerman, A., Goral, M., Edmonds, L. A., and Obler, L. K. (2020). Measuring treatment outcome in severe Wernicke's aphasia. *Aphasiology* 34, 1487–1505. doi: 10.1080/02687038.2020.1787729

Lewis, M. (2007). Stepwise versus hierarchical regression: pros and cons. Paper presented at the annual meeting of the southwest educational research association, Feb, San Antonio, TX.

Lundgren, K., Helm-Estabrooks, N., and Klein, R. (2010). Stuttering following acquired brain damage: a review of the literature. *J. Neurolinguistics* 23, 447–454. doi: 10.1016/j.jneuroling.2009.08.008

Lundie, M., Erasmus, Z., Zsilavecz, U., and Van der Linde, J. (2014). Compilation of a preliminary checklist for the differential diagnosis of neurogenic stuttering. South African. *J. Commun. Disord.* 61:10. doi: 10.4102/sajcd.v61i1.64

Lutz, K. C., and Mallard, A. R. (1986). Disfluencies and rate of speech in young adult nonstutterers. *J. Fluen. Disord.* 11, 307–316. doi: 10.1016/0094-730X(86)90018-5

McAuliffe, M. J., Carpenter, S., and Moran, C. (2010). Speech intelligibility and perceptions of communication effectiveness by speakers with dysarthria following traumatic brain injury and their communication partners. *Brain Inj.* 24, 1408–1415. doi: 10.3109/02699052.2010.511590

McCrea, M., Kelly, J. P., Randolph, C., Kluge, J., Bartolic, E., Finn, G., et al. (1998). Standardized assessment of concussion (SAC): on-site mental status evaluation of the athlete. *J. Head Trauma Rehabil.* 13, 27–35. doi: 10.1097/00001199-199804000-00005

McCrory, P., Meeuwisse, W., Dvorak, J., Aubry, M., Bailes, J., Broglio, S., et al. (2017). Consensus statement on concussion in sport—the 5th international conference on concussion in sport held in Berlin, October 2016. *Br. J. Sports Med.* 51, 838–847. doi: 10.1136/bjsports-2017-097699

McCrory, P., Meeuwisse, W. H., Echemendia, R. J., Iverson, G. L., Dvořák, J., and Kutcher, J. S. (2013). What is the lowest threshold to make a diagnosis of concussion? *Br. J. Sports Med.* 47, 268–271. doi: 10.1136/bjsports-2013-092247

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med.* 22, 276–282.

Mesfin, F., and Taylor, R. (2017). *Diffuse Axonal Injury (DAI)*. StatPearls Publishing, Treasure Island (FL)

Mohandas, E., Rajmohan, V., and Raghunath, B. (2009). Neurobiology of Alzheimer's disease. *Indian J. Psychiatry* 51, 55–61. doi: 10.4103/0019-5545.44908

Moro-Velazquez, L., Gomez-Garcia, J. A., Arias-Londoño, J. D., Dehak, N., and Godino-Llorente, J. I. (2021). Advances in Parkinson's disease detection and assessment using voice and speech: a review of the articulatory and phonatory aspects. *Biomed. Signal Process. Control* 66:102418. doi: 10.1016/j.bspc.2021.102418

Murdoch, B. E. (2001). Subcortical brain mechanisms in speech and language. *FPL* 53, 233–251. doi: 10.1159/000052679

Murray, L. L. (2012). Attention and other cognitive deficits in aphasia: presence and relation to language and communication measures. *Am. J. Speech Lang. Pathol.* 21, S51–S64. doi: 10.1044/1058-0360(2012/11-0067)

Nelson, W. E., Jane, J. A., and Deck, J. H. (1984). Minor head injury in sports: a new system of classification and management. *Phys. Sportsmed.* 12, 103–107. doi: 10.1080/00913847.1984.11701798

Norman, R. S., Jaramillo, C. A., Eapen, B. C., Amuan, M. E., and Pugh, M. J. (2018). Acquired stuttering in veterans of the wars in Iraq and Afghanistan: the role of traumatic brain injury, post-traumatic stress disorder, and medications. *Mil. Med.* 183, e526–e534. doi: 10.1093/milmed/usy067

Noufi, C., Lammert, A., Mehta, D., Williamson, J., Ciccarelli, G., Sturim, D., et al. (2019). Vocal biomarker assessment following Pediatric traumatic brain injury: a retrospective cohort study. *Inter Speech*, 3895–3899. doi: 10.21437/Interspeech.2019-1200

Obermeyer, J., Schlesinger, J., and Martin, N. (2020). Evaluating the contribution of executive functions to language tasks in cognitively demanding contexts. *Am. J. Speech Lang. Pathol.* 29, 463–473. doi: 10.1044/2019_AJSLP-CAC48-18-0216

Ommaya, A. K. (1985). "Biomechanics of head injury: experimental aspects" in *Biomedics of Trauma*. eds. A. M. Nahum and J. Melvin (New York: Appleton & Lange), 245–269.

Patel, S., Grabowski, C., Dayalu, V., Cunningham, M., and Testa, A. J. (2021). Fluency changes due toSports-related concussion. *MedRxiv* 2021:19.21263791. doi: 10.1101/2021.09.19.21263791

Patel, S., Oishi, K., Wright, A., Sutherland-Foggio, H., Saxena, S., Sheppard, S. M., et al. (2018). Right hemisphere regions critical for expression of emotion through prosody. *Front. Neurol.* 9:224. doi: 10.3389/fneur.2018.00224

Patel, S., and Shrivastav, R. (2011). A preliminary model of emotional prosody using multidimensional scaling. In INTERSPEECH-2011, 2957–2960.

Pistono, A., Jucla, M., Barbeau, E. J., Saint-Aubert, L., Lemesle, B., Calvet, B., et al. (2016). Pauses during autobiographical discourse reflect episodic memory processes in early Alzheimer's disease. *J. Alzheimers Dis.* 50, 687–698. doi: 10.3233/JAD-150408

Powell, D., Stuart, S., and Godfrey, A. (2021). Sports related concussion: an emerging era in digital sports technology. Npj digit. *NPJ Digital Med.* 4, 1–8. doi: 10.1038/s41746-021-00538-w

Power, E., Weir, S., Richardson, J., Fromm, D., Forbes, M., MacWhinney, B., et al. (2020). Patterns of narrative discourse in early recovery following severe traumatic brain injury. *Brain Inj.* 34, 98–109. doi: 10.1080/02699052.2019.1682192

Roberts, P. M., Meltzer, A., and Wilding, J. (2009). Disfluencies in non-stuttering adults across sample lengths and topics. *J. Commun. Disord.* 42, 414–427. doi: 10.1016/j.jcomdis.2009.06.001

Robertson, S. C., and Diaz, K. (2020). Case report of acquired stuttering after soccer-related concussion: functional magnetic resonance imaging as a prognostic tool. *World Neurosurg.* 142, 401–403. doi: 10.1016/j.wneu.2020.06.233

Rose, S. C., Weldy, D. L., Zhukivska, S., and Pommering, T. L. (2021). Acquired stuttering after pediatric concussion. *Acta Neurol. Belg.* 122, 545–546. doi: 10.1007/s13760-021-01653-x

Ross, E. D., and Monnot, M. (2008). Neurology of affective prosody and its functional-anatomic organization in right hemisphere. *Brain Lang.* 104, 51–74. doi: 10.1016/j.bandl.2007.04.007

Roth, C. R., Cornis-Pop, M., and Beach, W. A. (2015). Examination of validity in spoken language evaluations: adult onset stuttering following mild traumatic brain injury. *NeuroRehabilitation* 36, 415–426. doi: 10.3233/NRE-151230

Ruff, R. M., Iverson, G. L., Barth, J. T., Bush, S. S., Broshek, D. K., Policy, N. A. N., et al. (2009). Recommendations for diagnosing a mild traumatic brain injury: a National Academy of neuropsychology education paper. *Arch. Clin. Neuropsychol.* 24, 3–10. doi: 10.1093/arclin/acp006

Rusz, J., Cmejla, R., Ruzickova, H., and Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *J. Acoust. Soc. Am.* 129, 350–367. doi: 10.1121/1.3514381

Salvatore, A. P., Cannito, M. P., Hewitt, J., Dolan, L. D., King, G., and Brassil, H. E. (2019). Motor speech and motor limb status in athletes following a concussion. *Clin. Arch. Commun. Disord.* 4, 214–222. doi: 10.21849/cacd.2019.00150

Sawyer, J., and Yairi, E. (2010). Characteristics of disfluency clusters over time in preschool children who stutter. *J. Speech Lang. Hear. Res.* 53, 1191–1205. doi: 10.1044/1092-4388(2010/09-0067)

Schatz, P., Pardini, J., Lovell, M., Collins, M., and Podell, K. (2006). Sensitivity and specificity of the ImPACT test battery for concussion in athletes. *Arch. Clin. Neuropsychol.* 21, 91–99. doi: 10.1016/j.acn.2005.08.001

Shaw, N. A. (2002). The neurophysiology of concussion. *Prog. Neurobiol.* 67, 281–344. doi: 10.1016/S0301-0082(02)00018-7

Shimada, M., Meguro, K., Yamazaki, H., Horikawa, A., Hayasaka, C., Yamaguchi, S., et al. (1998). Impaired verbal description ability assessed by the picture description task in Alzheimer's disease. *Arch. Gerontol. Geriatr.* 27, 57–65. doi: 10.1016/s0167-4943(98)00099-5

Shriberg, E. (2001). To 'errrr' is human: ecology and acoustics of speech disfluencies. *J. Int. Phon. Assoc.* 31, 153–169. doi: 10.1017/S0025100301001128

Slegers, A., Filiou, R.-P., Montembeault, M., and Brambati, S. M. (2018). Connected speech features from picture description in Alzheimer's disease: a systematic review. *J. Alzheimers Dis.* 65, 519–542. doi: 10.3233/JAD-170881

Smith, K. M., Ash, S., Xie, S. X., and Grossman, M. (2018). Evaluation of linguistic markers of word-finding difficulty and cognition in Parkinson's disease. *J. Speech Lang. Hear. Res.* 61, 1691–1699. doi: 10.1044/2018_JSLHR-L-17-0304

Snyder, P. (1991). "Three reasons why stepwise regression methods should not be used by researchers" in *Advances in Social Science Methodology*. ed. B. Thompson, vol. *1* (Greenwich, CT: JAI Press), 99–105.

Solomon, N. P., McKee, A. S., and Sandra, G.-B. (2001). Intensive voice treatment and respiration treatment for hypokinetic-spastic dysarthria after traumatic brain injury. *Am. J. Speech Lang. Pathol.* 10, 51–64. doi: 10.1044/1058-0360(2001/008)

St. Louis, K. O., Murray, C. D., and Ashworth, M. S. (1991). Coexisting communication disorders in a random sample of school-aged stutterers. *J. Fluen. Disord.* 16, 13–23. doi: 10.1016/0094-730X(91)90032-8

Stubbs, E., Togher, L., Kenny, B., Fromm, D., Forbes, M., MacWhinney, B., et al. (2018). Procedural discourse performance in adults with severe traumatic brain injury at 3 and 6 months post injury. *Brain Inj.* 32, 167–181. doi: 10.1080/02699052.2017.1291989

Theodoros, D. G., Murdoch, B. E., and Chenery, H. J. (1994). Perceptual speech characteristics of dysarthric speakers following severe closed head injury. *Brain Inj.* 8, 101–124. doi: 10.3109/02699059409150963

Toldi, J., and Jones, J. (2021). A case of acute stuttering resulting after a sports-related concussion. *Curr. Sports Med. Rep.* 20, 10–12. doi: 10.1249/JSR.0000000000000795

Tomik, B., and Guiloff, R. J. (2010). Dysarthria in amyotrophic lateral sclerosis: a review. *Amyotroph. Lateral Scler.* 11, 4–15. doi: 10.3109/17482960802379004

Tóth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., et al. (2018). A speech recognition-based solution for the automatic detection of mild cognitive impairment from spontaneous speech. *Curr. Alzheimer Res.* 15, 130–138. doi: 10.2174/1567205014666171121114930

Vergis, M. K., Ballard, K. J., Duffy, J. R., McNeil, M. R., Scholl, D., and Layfield, C. (2014). An acoustic measure of lexical stress differentiates aphasia and aphasia plus apraxia of speech after stroke. *Aphasiology* 28, 554–575. doi: 10.1080/02687038.2014.889275

Wang, Y.-T., Kent, R. D., Duffy, J. R., and Thomas, J. E. (2005). Dysarthria associated with traumatic brain injury: speaking rate and emphatic stress. *J. Commun. Disord.* 38, 231–260. doi: 10.1016/j.jcomdis.2004.12.001

Wildgruber, D., Ackermann, H., and Grodd, W. (2001). Differential contributions of motor cortex, basal ganglia, and cerebellum to speech motor control: effects of syllable repetition rate evaluated by fMRI. *NeuroImage* 13, 101–109. doi: 10.1006/nimg.2000.0672

Yadegari, F., Azimian, M., Rahgozar, M., and Shekarchi, B. (2014). Brain areas impaired in oral and verbal apraxic patients. *Iranian J. Neurol.* 13, 77–82. PMID: 25295150

Yee, E., Blumstein, S., and Sedivy, J. C. (2008). Lexical-semantic activation in Broca's and Wernicke's aphasia: evidence from eye movements. *J. Cogn. Neurosci.* 20, 592–612. doi: 10.1162/jocn.2008.20056

Yorkston, K. M. (1996). Treatment efficacy. *J. Speech Lang. Hear. Res.* 39, S46–S57. doi: 10.1044/jshr.3905.s46

TYPE Original Research
PUBLISHED 19 June 2023
DOI 10.3389/fpsyg.2023.1128976

Phonetic entrainment in L2 human-robot interaction: an investigation of children with and without autism spectrum disorder

Yitian Hong[1], Si Chen[1,2,3,4]*, Fang Zhou[1], Angel Chan[1,2,4] and Tempo Tang[1]

[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China, [2]Department of Chinese and Bilingual Studies, Research Centre for Language, Cognition, and Neuroscience, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China, [3]Research Institute for Smart Ageing, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China, [4]The HK PolyU-PekingU Research Centre on Chinese Linguistics, The Hong Kong Polytechnic University, Hong Kong, Hong Kong SAR, China

Phonetic entrainment is a phenomenon in which people adjust their phonetic features to approach those of their conversation partner. Individuals with Autism Spectrum Disorder (ASD) have been reported to show some deficits in entrainment during their interactions with human interlocutors, though deficits in terms of significant differences from typically developing (TD) controls were not always registered. One reason related to the inconsistencies of whether deficits are detected or not in autistic individuals is that the conversation partner's speech could hardly be controlled, and both the participants and the partners might be adjusting their phonetic features. The variabilities in the speech of conversation partners and various social traits exhibited might make the phonetic entrainment (if any) of the participants less detectable. In this study, we attempted to reduce the variability of the interlocutors by employing a social robot and having it do a goal-directed conversation task with children with and without ASD. Fourteen autistic children and 12 TD children participated the current study in their second language English. Results showed that autistic children showed comparable vowel formants and mean fundamental frequency (*f0*) entrainment as their TD peers, but they did not entrain their *f0* range as the TD group did. These findings suggest that autistic children were capable of exhibiting phonetic entrainment behaviors similar to TD children in vowel formants and *f0*, particularly in a less complex situation where the speech features and social traits of the interlocutor were controlled. Furthermore, the utilization of a social robot may have increased the interest of these children in phonetic entrainment. On the other hand, entrainment of *f0* range was more challenging for these autistic children even in a more controlled situation. This study demonstrates the viability and potential of using human-robot interactions as a novel method to evaluate abilities and deficits in phonetic entrainment in autistic children.

KEYWORDS

phonetic entrainment, autism, children, robot, controlled speech, conversation task

EDITED BY
Stefanie Peykarjou,
Heidelberg University, Germany

REVIEWED BY
Julia Hirschberg,
Columbia University, United States
Emilia I. Barakova,
Eindhoven University of Technology,
Netherlands

*CORRESPONDENCE
Si Chen
✉ sarah.chen@polyu.edu.hk

RECEIVED 21 December 2022
ACCEPTED 19 May 2023
PUBLISHED 19 June 2023

CITATION
Hong Y, Chen S, Zhou F, Chan A and Tang T (2023) Phonetic entrainment in L2 human-robot interaction: an investigation of children with and without autism spectrum disorder.
Front. Psychol. 14:1128976.
doi: 10.3389/fpsyg.2023.1128976

1. Introduction

During conversations, the interlocutors from a typical population coordinate with each other in verbal and non-verbal ways. These cooperative behaviors—where individuals adjust their behaviors to match closely mirror their conversation partners—are referred to as entrainment (also called "convergence," "alignment," or "accommodation" in some studies). For example, it has been found that interlocutors who are strangers to one another use head nodding and eye gaze coordination to signal mutual understanding (Cassell et al., 2007). On the other hand, entrainment in speech is more subtle and complicated. Studies working on phonetic entrainment have adapted diverse cooperative tasks and involved a wide variety of speech features. For example, a series of English and Mandarin corpus studies revealed similar *f0*, intensity, and speech rate between interlocutors when they were playing computer games that required communication (Levitan and Hirschberg, 2011; Xia et al., 2014; Levitan et al., 2015). In addition, children as young as 9 years old were found to converge in mean *f0* in "spot-the-difference" games in Lehnert-LeHouillier et al. (2020). Hogstrom et al. (2018) also reported convergence of phoneme duration from children aged from 12 to 18 in a cooperative map searching task.

In essence, phonetic entrainment is the product of the connection between perception and production (Coles-Harris, 2017). The process of phonetic entrainment requires the ability to detect the acoustic cues of the interlocutor(s) and adjust one's own production accordingly (Phillips-Silver et al., 2010; Wynn et al., 2018). From this perspective, deficits in speech prosody might cause failure in phonetic entrainment. Atypical prosodic production—such as wider *f0* range (Nadig and Shaw, 2012) and longer turn-taking gaps (Ochi et al., 2019)—was found in individuals with Autism Spectrum Disorder (ASD). Studies on phonetic entrainment of autistic children showed mixed results. Hogstrom et al. (2018) found that TD children converged in their phoneme duration in the post-interaction period while autistic children showed a trend of divergence. However, some studies reported a tendency of similar unchanged phonetic adjustment between autistic and TD children, for example, similar unchanged adjustment in speech rate (Wynn et al., 2018) and *f0* range (Lehnert-LeHouillier et al., 2020).

One reason for the undetected phonetic entrainment in children might be due to the fact that the required prosodic skills have not developed into an adult-like level (Wynn et al., 2018). Another reason is that both participants and conversation partners have the potential to adjust their phonetic features at the same time, which makes it harder to examine phonetic convergence from one side. Additionally, variation in conversation partners, such as their various social traits, might make the phonetic entrainment (if any) of the participants more varied and with less detectable patterns. Furthermore, as previous studies have indicated, the age range of 7–12 is a critical period for children's development of rhythm recognition (Upitis, 1987). Therefore, if the speech of their interlocutors can be controlled with no phonetic adjustment and no variations throughout the experiments, we might be able to detect phonetic entrainment patterns in autistic children. This possibility has not been available until the application of social robots.

In this study, we use a social robot as a conversation partner to investigate whether phonetic entrainment can be found among children with and without ASD in a conversation task with a better controlled interlocutor. During the experiment, the acoustic features and social traits (reflected in facial expressions and the manner of interactions) of the robot remained consistent. Children's conversations with the robot were recorded and compared with their baseline production and post-interaction production. We target at the entrainment of fundamental frequency (*f0*), and formant frequencies (F1, F2). *F0* refers to the vibration of vocal folds (Yavas, 2011). The perceptual correlate of *f0* is pitch, which reveals signals of sound identity and information about meaning (McPherson and McDermott, 2018). The variation of *f0* is an important part of speech prosody manipulation. By examining mean *f0* and *f0* range, we can understand more about their adjustment of pitch during interaction. Formant frequencies relates to vocal tract configuration, reflecting the tongue position when the speaker articulates the vowels (Yavas, 2011). The investigation of first and second vowel formant improves our understanding about vowel space area manipulation during conversations (Pettinato et al., 2016).

2. Background

2.1. Entrainment in the broad sense

Humans show an in-born tendency to coordinate with outside stimuli (Phillips-Silver et al., 2010). For example, humans tend to clap hands or shake heads along with the rhythm of a song when they are exposed to it. Infants as young as 5 months old have shown coordinated body movement with music (Ilari, 2015). Such coordination is called entrainment.

Social entrainment occurs when the outside stimulus comes from another human (Phillips-Silver et al., 2010). During the social interaction, social entrainment demonstrates social functions important in facilitating social communication. By entraining in the time domain (e.g., entrainment of turn-taking gaps), it improves mutual understanding between the interlocutors, helps build consensus and establish positive connections (Borrie and Liss, 2014). It fulfills the function of sustaining the emotional and social relationship between interlocutors (Borrie and Liss, 2014). Social entrainment also increases the interlocutors' enjoyment of the communication and facilitates the development of the social relationship. In the study of Chartrand and Bargh (1999), they asked the interlocutor to intentionally mimic the gestures of the participants and asked the participants to rate the experience of social interaction after the experiment. They found that when interlocutors mimicked participants' gestures, the participants rated the experience as smoother and the interlocutor as friendlier compared to the control group (where the interlocutors did not mimic any gestures). In regard to phonetic entrainment, Borrie and Delfino (2017) found that interlocutor dyads who showed a match of vocal fry frequency tended to find their communication more enjoyable. In a corpus study, Lee et al. (2010) demonstrated that couples with positive emotions during conversations showed *f0* related entrainment as compared to those with negative emotions. Furthermore, social entrainment increases communication efficiency. It improves information transfer and enhances mutual agreement and sympathy (Gill, 2012). In the same study, Borrie and Delfino (2017) found that participants' degree of entrainment on their frequency of vocal fry was also positively correlated with their efficiency in doing a cooperative task. Similarly, Nenkova et al. (2008) found that entrainment of high frequency lexicons led to higher scores in cooperative games. More specifically, Levitan et al. (2011) found that entraining backchannel cues decreased turn-taking gaps and interruptions and improved task complication efficiency.

Since phonetic entrainment might be correlated with social rapport and social communication efficiency, deficits in entrainment could be associated with poor social skills. Therefore, populations with communication disorders are more likely to have deficits in entrainment. Autism Spectrum Disorder (ASD) is one group of disorder correlated with communication and social interaction difficulties. According to American Psychiatric Association (2022), individuals with ASD demonstrate three core characteristics: atypical social communication, restricted interests, and repetitive behaviors. Empirical evidence has shown that autistic populations did present certain degrees of deficits in social entrainment. Previous studies have found that autistic individuals did not show comparable non-verbal social entrainment relative to their TD peers. Nakano et al. (2011) found that autistic adults failed to entrain their eyeblink with the speakers. The eyeblink entrainment occurred at conversation pause, forming an important part of conversation coordination. The disruption of eyeblink entrainment might affect autistic individuals' social interactions. Other than the eyeblink, autistic individuals also demonstrated incomparable facial muscle movement when mimicking others' emotions, which was suggested to affect their social reciprocity with conversation partners (Mathersul et al., 2013). Similarly, Yoshimura et al. (2015) reported reduced times of facial expression synchrony of autistic individuals as compared to the TD population. They found that individuals with higher degree of social dysfunction tended to show lower frequency of facial expression synchrony. In a different study working with autistic children, Helt et al. (2010) reported delays in development in yawning mimicry. They suggested that autistic children's delayed acquisition of social behavior mimicry might be due to the lack of social interest in interaction, and in turn, they have fewer social experiences compared to their TD peers.

These behaviors are categorized as contextual and socially meaningful entraining behavior, distinguishable from simple automatic mimicry (Nakano et al., 2011). The breakdown of such behaviors could potentially be associated with unpleasant and ineffective social communication. On the other hand, the model of social entrainment might provide a new perspective for understanding autistic populations' social behaviors.

2.2. Speech features and phonetic entrainment of autistic children

Unlike non-verbal entrainment, entrainment in phonetic features is a more fine-grained process, where the interlocutors detect and perceive the phonetic features (e.g., speech rate, fundamental frequency (*f0*), vowel formant) of their conversation partners and adjust their own phonetic features in speech production accordingly. This process involves the processes of phonetic perception and production. Atypicality in any step of this process might lead to deficits in phonetic entrainment. The autistic population has long been found to show different speech features from the TD population, such as vowel formants and *f0* range. Although the reasons behind their atypical speech features remain unclear, the empirical studies working on gaining a better understanding of their speech features might provide some hints on their phonetic entrainment.

Studies on vowels mainly reported exaggerated vowel formants produced by autistic children. Lyakso et al. (2016) found larger vowel formant triangles in autistic children when compared to their TD peers. Mohanta and Mittal (2022) reported higher vowel formants for autistic children than TD children, which was interpreted as atypicality in vowel production mechanism. However, their production tended to show less dispersion. Kissine and Geelhand (2019) and Kissine et al. (2021) reported lower variabilities of vowel formants in autistic children, compared to their *f0*. They proposed a possibility that autistic individuals tended to pay more attention to the precision of vowel pronunciation and thus might overact the target articulatory manners, leading to exaggerated vowel formants, while TD individuals spoke in a more leisure style.

In regard to speech prosody, discrepancies exist between the findings from production and perception. Nadig and Shaw (2012) found a significantly larger *f0* range in the autistic group than TD group, but no significant difference in the mean *f0*. The larger pitch range of the autistic population, although acoustically abnormal, was not perceived as a signal of odd speech by TD listeners (Nadig and Shaw, 2012). Similarly, Patel et al. (2020) found larger *f0* excursion in utterance-final position, but it did not serve as a marker of autism to non-clinical listeners. However, in contrast to this, Shriberg et al. (2001) reported that over half of the autistic participants were rated as exhibiting unusual prosody while only about 6% in TD participants were rated the same. This finding was associated with differences in the mean *f0* and *f0* range between these two groups. The mixed findings of perceptual differences of their prosody indicated that it was difficult for listeners to interpret the prosodic cues of autistic individuals. This might be due to the fact that autistic population did not use prosody functionally in communication (Nadig and Shaw, 2012). They tended to use a limited repertoire of prosody repetitively, which may be related to one of their core features—restricted and repetitive behaviors (Green and Tobin, 2009). On the other hand, the exaggerated style of prosody (higher *f0* and larger *f0* range) is similar to infant-directed speech, which is suggested to be a signal of inability to outgrow from motherese, indicating their undeveloped control of prosody (Sharda et al., 2010). These two indications point to a possibility that autistic individuals are less flexible in adjusting prosodic features in communication relative to their TD peers. Therefore, it is suspected that their entrainment in prosodic cues might not be as comparable as their TD peers.

Some studies have revealed lack of entrainment in a variety of phonetic features from autistic adults in their first language. For instance, no flexible adjustment of speech volume by autistic adults was reported in Ochi et al. (2019). Autistic adults were also found to lack speech rate entrainment in a quasi-conversation experiment as opposed to their TD peers (Wynn et al., 2018). However, in terms of studies of entrainment from autistic children, the results were inconsistent. In the study of Wynn et al. (2018), although they found significant differences of speech rate convergence between autistic and TD adults, autistic children and TD children's speech rate did not show significant differences. Hogstrom et al. (2018) compared the phoneme duration of keywords before and after a conversation task with a TD interlocutor. They reported that autistic children tended to diverge in phoneme duration from the interlocutor after the task, compared with the pre-task production, while TD children showed convergence.

However, they also found that neither autistic children dyads nor TD children dyads demonstrated *f0* adjustment. Lehnert-LeHouillier et al. (2020) reported no significant difference of *f0* range entrainment between autistic and TD teens. These findings suggest that we need to understand more about the conditions under which autistic children show or do not show TD-like phonetic entrainment, before one can better evaluate whether phonetic entrainment can serve as a linguistic biomarker for differentiating and TD children. Moreover, specifying these conditions inform us about their capabilities in achieving phonetic entrainment and their deficits in this aspect.

2.3. Phonetic entrainment in second language (L2)

Previous studies carrying out conversation task between L1 and L2 speakers reported more phonetic convergence from L2 speakers than their L1 interlocutors (Hwang et al., 2015). Similarly, in word shadowing task, L2 speakers are found with more phonetic convergence than L1 speakers (Lewandowski and Nygaard, 2018; Gnevsheva et al., 2021). They argued that larger phonetic differences between L1 and L2 speech allow L2 speakers to have more space for entrainment (Lewandowski and Nygaard, 2018). It can also be explained by a mediated priming effect with the intention of producing more native-like speech (Hwang et al., 2015), namely the more prestigious variety (Gnevsheva et al., 2021) and increasing communication efficiency. It remains unknown whether non-native autistic speakers demonstrated a similar pattern of L2 phonetic entrainment.

Although previous studies have reported a delay of autistic individuals' L1 development, particularly in discourse and pragmatic functions (Kelley et al., 2006), there are studies reporting that their L2 was relatively unaffected (Práinsson, 2012; Agostini and Best, 2015). These studies mainly involved autistic subjects who did not suffer from intellectual impairment and whose language abilities were comparable with their TD peers in general. For example, the case study in Práinsson (2012) reported that the autistic subjects showed a good command of pragmatics, discourse prosody, and syntax of second language and even surpassed their TD peers. Agostini and Best (2015) found that the second language grammatical development of young autistic children was comparable and even faster than their TD peers. Because studies focusing on autistic individuals' second language are scarce, and no study examined phonetic entrainment of their L2, this study attempts to provide some innovative empirical evidence of autistic children's phonetic entrainment in their L2 to further our understanding of their second language acquisition.

2.4. Benefits of using a social robot as a conversation partner

As compared to entrainment in non-social contexts, such as entrainment with musical rhythm, social entrainment is special because it is a mutual process where both individuals adjust their behaviors to approximate each other's. This special condition brings uncertainty and might be the reason why previous findings on the phonetic entrainment of autistic population tended to be inconsistent. Lehnert-LeHouillier et al. (2020) found that the atypical entrainment behavior of autistic youth, evidenced by a manipulation of difference between conversation dyads, was in fact the result of adjustment from their conversation partner. Therefore, the current study uses a social robot as a conversation partner to investigate the phonetic entrainment of autistic children in comparison with their TD peers. A social robot has the advantages of controlled speech with no phonetic entrainment and consistent social complexities, which might facilitate the detection of children's phonetic adjustment.

Social robots have been used previously in therapy and research on autism in a longitudinal study, Robins et al. (2005) found that autistic children's social skills were improved with the help of a humanoid robot. Dautenhahn and Werry (2004) found that autistic children showed more engagement in activities with robots and learned how to take turns and imitate the robot. Similar findings were reported by Barakova et al. (2015) where a robot-present scenario led to more social initiations of autistic children. Stanton et al. (2008) also found that autistic children were able to treat social robots as a social category and produce more words than playing with a non-verbal robot. In addition, some studies working on robotic voice have reported that autistic children exhibit a special preference to mechanic voices rather than human voices (Kuhl et al., 2005).

These attempts of using social robots to assist autistic populations reveal benefits, such as reducing the social pressure of autistic individuals and attracting their attention. Compared to human beings, social robots have fewer social complexities, e.g., more controlled facial expressions. They are more predictable due to their consistent voice and gestures (Marchi et al., 2014). They provide a structured interaction environment for autistic individuals to converse and learn (Kumazaki et al., 2020). These advantages of social robot might resolve the uncertainty of phonetic entrainment in human-human interactions. By designing an experiment of human-robot interaction, we aim to examine autistic children's phonetic entrainment in a more controlled context.

2.5. Research questions and predictions

As reviewed above, autistic individuals might have problems in manipulating phonetic features in conversations. The inconsistency of interlocutors increases their difficulties in phonetic entrainment and also makes the phonetic manipulation of autistic individuals less detectable. The controlled nature of a robot provides a controlled conversation environment which might facilitate phonetic entrainment and its detection of autistic individuals. Moreover, as convergence on more speech features toward words recorded naturally than words generalized in synthetic voice has been reported (Gessinger et al., 2021), more natural speech used in the current study might trigger more entrainment than previous child-robot interaction studies (see Section 3.2.2. for more details about the sound used in the robot). Therefore, our main research question is: do autistic children and TD children show comparable phonetic entrainment when interacting with a social robot?

We expect that autistic children may show phonetic entrainment in a more controlled phonetic and social environment, but their performance may still be different from TD children. Specifically, we will examine vowel formants and fundamental frequency in the speech production of a group of autistic children and compare their production with their TD peers to identify whether they would show

TD-like phonetic entrainment. We predict that autistic children are more likely to entrain vowel formant toward the standardized vowel target, consistently produced by the robot. On the other hand, we predict that their deficits of prosody will still affect their entrainment even when they interact with a controlled interlocutor. Therefore, they are predicted to show problems in phonetic entrainment of *f0*-related parameters (mean *f0* and *f0* range in the study).

3. Methods

Because phonetic entrainment is supposed to occur in both segmental level and prosodic level, the main task should be able to elicit natural conversational speech, and also yield enough repetitions for word-level acoustic analysis. We did not consider Map Tasks (Anderson et al., 1991), where one interlocutor found a route in the picture following the instruction of the other. Because the conversation dyads do not receive equal amount of information in the task, they have different pre-defined roles (i.e., a giver and a follower) and very uneven amount of production, which is not suitable for investigating phonetic entrainment. Therefore, we finally decided on a "spot the difference" game (van Engen et al., 2010). The main task was between the child and the robot, during which the participant interacted with a robot to find the differences between four pairs of pictures. The robot and the child participant would refer to pictures with slight differences in these four pairs. The robot asked questions regarding the color, number, and behavior of the objects in the pictures, guiding the child to notice the differences and to elicit keywords from him or her (see Section 3.2 for more details).

3.1. Participants

Fourteen L2 English-speaking autistic children and 12 age-matched typically developing (TD) children were recruited in Hong Kong. The autistic children received a clinical diagnosis of ASD from clinical settings in Hong Kong according to information provided by their parents. Both autistic and TD children had nonverbal IQ above 80 as assessed by the Raven's Standard Progressive Matrices Test (Raven, 2003). These children acquire Cantonese as their first and home language, and English and Mandarin as their second languages at school. Since this study focused on L2 English, their spoken English was assessed by Comprehensive Assessment of Spoken Language (CASL; Carrow-Woolfolk, 2017) and autistic children showed moderate English language proficiency. There were no reported hearing impairments nor neurological disorders for all participants. As previous study has found that musical experience might affect perception of phonetic details (Tsang et al., 2018), the musical training experience of two groups was controlled to be comparable. Their chronological age, duration of musical training, IQ standard score, CASL standard score, age of English acquisition, and their English proficiency (out of 5 as the maximum score) reported by their parents are shown in Table 1. One TD child (t10) did not take the Raven Test or the CASL test, and she showed no sign of abnormality according to the observation of the experimenter. Parents of the participants signed a written consent form, which was approved by the Departmental Research Committee of the Hong Kong Polytechnic University, and the participants were reimbursed for participation.

3.2. Materials and procedures

3.2.1. Pictures and keywords

The task materials were adapted from pictures designed in DiapixUK tasks (Baker and Hazan, 2011). They are 12 pairs of cartoon pictures specially designed for "spot-the-difference" game in English. The pictures included three themes. Each theme has four pairs of pictures, sharing similar vocabulary and depicting the same keywords. The picture set depicting the farm theme was selected. There were originally 12 differences (depicting 12 different keywords) per pair in their design. This design has been used in studies with native speakers as young as 8 years old (Pettinato et al., 2016; Tuomainen et al., 2022). Given the condition of our participants (children with Autism Spectrum Disorder often have issues with executive function), we revised the pictures to reduce the number of differences to five, to reduce the level of task complexity. The differences related to either a change of the item (e.g., an apple and a pear in picture A vs. two pears in picture B; an empty sack in picture A vs. a full sack in picture B; white sheep in picture A vs. gray sheep in picture B) or an item that was missing in one picture (e.g., a bush with flowers in picture A vs. a bush without flowers in picture B). To increase visual saliency, the areas associated with the differences between a pair of pictures were circled and numbered in the pictures (see Figure 1 for a sample picture pair). The keywords related to the differences will be used for analyzing phonetic entrainment in segmental level while the conversational speech produced during interaction will be used for investigating prosodic entrainment.

3.2.2. Robot and experimental setup

The robot we used in this study as a conversation partner is social robot Furhat (Al Moubayed et al., 2012). Furhat robot has a physical body with a neck and a movable head with a light-projected face. Its speech production was pre-scripted to be triggered by corresponding keywords. The robot's speech was generated using Amazon Polly neural TTS system. Compared to usual robotic speech, their speech showed more naturalness in dialog due to shorter response time and higher articulation accuracy (Amazon Polly Developer Guide, 2023). In particular, we selected the voice of an American English male named Matthew which was produced by the neural TTS system rather than standard system. The neural system used a sequence-to-sequence method to generate "the most natural and human-like" sounds with rather higher quality (Amazon Polly Developer Guide, 2023, p. 1). The volume of the speech was set consistently for all the children.

The experiment took place in a soundproof booth, and the robot was placed on a table about 85 cm away from the participant. The child participant sat facing the robot, and the seat height was adjusted to make sure that each child was at the robot's eye level. The picture to elicit speech interaction and a microphone Blue Snowball connected to the robot were placed on a table in between the participant and the robot. The robot used the microphone to receive speech from the participant so as to trigger its corresponding response upon perceiving certain keywords. The speech recordings were done at a 44.1 kHz sampling rate with 16-bit resolution by another microphone, an Azden ECZ-990 microphone, connected with audacity in the

TABLE 1 Means (and standard deviations) of chronological age, IQ standard score, CASL standard score, duration of musical training (months), age of English acquisition, and English proficiency score (5 points each) across the two groups of children.

Group	Number (male)	Chronological age	IQ score	CASL score	AoA	Musical training	Listening	Writing	Reading	Speaking
ASD	14 (9)	9.5 (1.16)	102.29 (14.79)	66.43 (21.25)	2.8 (1.65)	18 (21.05)	3.5 (0.76)	2.9 (1.07)	3.4 (1.01)	2.9 (1.07)
TD	12 (8)	9.1 (1.16)	103.2 (14.03)	88.8 (22.76)	2.6 (1.44)	22 (19.54)	4.1 (0.51)	3.5 (0.90)	3.8 (0.87)	3.5 (1.09)

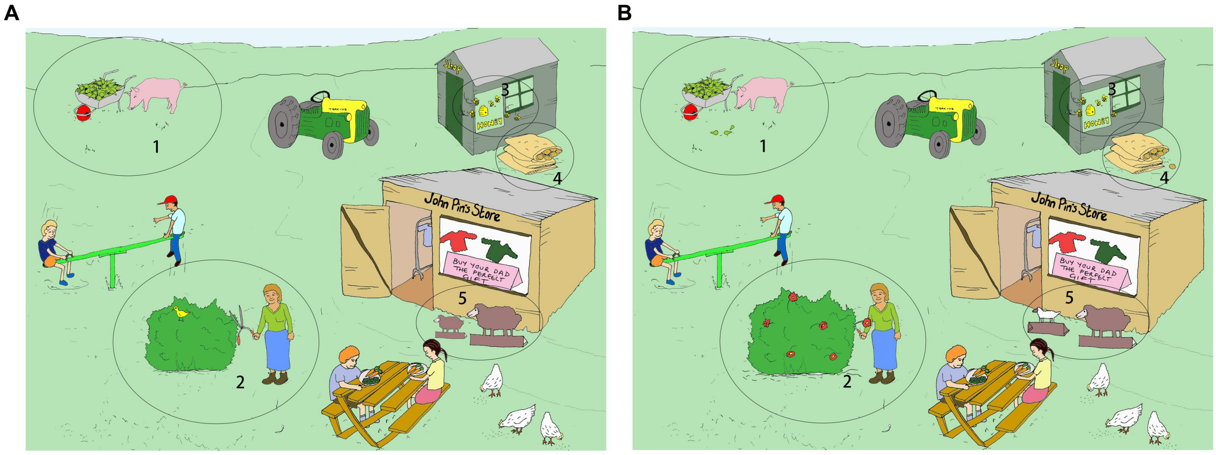


FIGURE 1
One of the four picture pairs for experiment. The child held picture **(A)**. The robot held picture **(B)**. Image source: https://doi.org/10.5281/zenodo.3703202. Reproduced under the terms of Creative Commons Attribution 4.0.

computer. This recording microphone was placed on another table by the left of the participant. The experimental set-up is demonstrated in Figure 2.

3.3. Procedures

Before interacting with the robot, the child recorded five keywords as baseline production. They were shown pictures of keywords one by one on the screen and asked to say what they could see in the picture in English. Each keyword was produced in singular and plural forms, twice in isolation and once in a carrier sentence: I can see the KEYWORD(s) in the picture. Each keyword was elicited in 2 forms * (2 isolation + 1 carrier) = 6 repetitions. After baseline production, the child watched a video introducing how to play the 'spot the difference' game presented in their first language Cantonese to ensure they understood the task expectation well. Then, one pair of pictures depicting themes different from the experimental test items was given to the child for practice. To allow adequate time to let the child become familiarized with the task procedures, each child was given 5 min to try to determine the differences between the two pictures by themselves.

The interaction with the social robot started with a "say-hello" session. The robot greeted the child to familiarize the child with the robot's voice. The "say-hello" session triggered four turns of interactions between the child and the robot. The experimenter double checked with the child to confirm their readiness before starting the interaction tasks. There are four pairs of pictures to look for differences (four tasks). These tasks were launched by the experimenter one at a time. The task order was randomized. Each task lasted for 10–15 min. The child was allowed to take a break between the tasks.

After the child finished all the tasks, the experimenter asked the child to record the keywords again following the same procedure as the baseline speech production. The keywords produced before, during and after the interaction will be compared.

3.4. Data analysis

3.4.1. Data extraction and normalization

The vowel portions of the five keywords produced in each recording by the child and the robot were segmented manually by a trained phonetician using Praat (Boersma and Weenink, 2018). The first formant (F1) and second formant (F2) values were extracted at the midpoint of each vowel portion. We adapted the praat script from Stanley and Lipani (2019) to extract the vowel formants automatically.

In order to investigate the adjustment of *f0* parameters in more details, we segmented the child's production into multiple inter-pause units (IPU). IPU is defined following Levitan and Hirschberg (2011) as a chunk of utterances with pauses in certain duration from one single speaker in one turn, with the adaptation that we adjusted the pause duration from 50 ms to 180 ms, based on previous studies showing that the articulation rate of children (as in our study) in spontaneous speech is significantly slower than adults (as in Jacewicz et al., 2010; Levitan and Hirschberg, 2011). This number was derived empirically from the maximum length of Voice Onset Time of all the

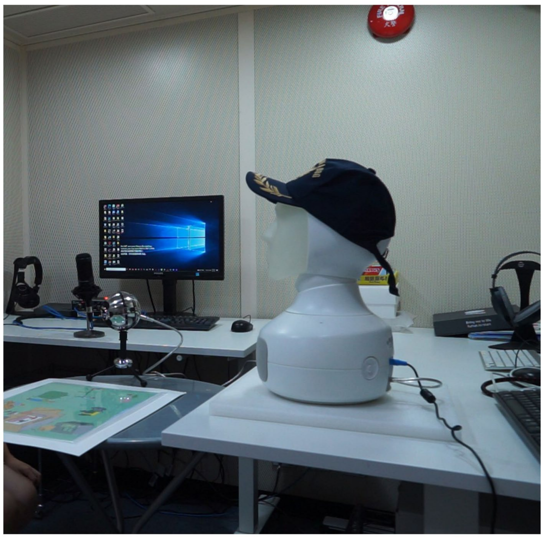


FIGURE 2
Experimental setting. Image source: https://doi.org/10.5281/zenodo.3703202. Reproduced under the terms of Creative Commons Attribution 4.0.

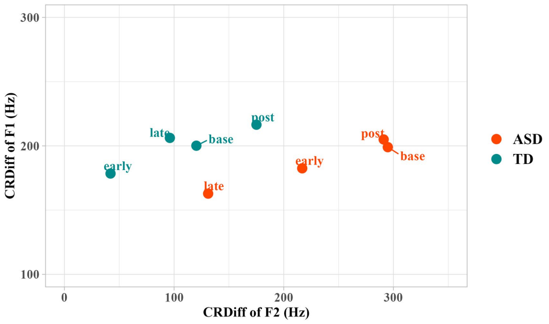


FIGURE 3
Difference of F1 and F2 between the children and the robot in different time periods. CRDiff=Children's Production—Robot's Production.

recordings. Mean *f0*, maximum *f0*, and minimum *f0* were extracted in each IPU by Praat (Boersma and Weenink, 2018). The *f0* range was calculated as the distance between the minimum *f0* and maximum *f0* in each IPU. We applied the log z-score normalization as in Zhu (2005) to *f0* values.

3.4.2. Statistical analyses

The measurements of phonetic entrainment were to evaluate the similarity of acoustic cues between interlocutors. Regarding the three target parameters (i.e., vowel formant, log mean *f0*, log *f0* range), we compared the differences between the robot and each child across baseline, early production (the first two tasks), late production (the last two tasks), and post-task production. Since the robot's production was controlled to be consistent throughout the experiment, the differences across time would be contributed by the child.

We first calculated the distance in each parameter (i.e., F1, F2, log mean *f0*, log *f0* range) between the child's production and the robot's production. The absolute values of the robot's production were subtracted from the corresponding values of the child's production, yielding CRDiff (CRDiff = children's baseline/early production/late production/post-task production—robot's production). Linear mixed effects models were then fitted using the "lmerTest" package (Kuznetsova et al., 2017) in R (R Core Team, 2016) to determine whether CRDiffs in vowel formant, log mean *f0*, and log *f0* range were significantly affected by group (autistic vs. TD children) and time period (base vs. early vs. late vs. post). The "effectsize" package (Ben-Shachar et al., 2020) was used to report the standardized coefficient (β') and confidential intervals of the optimal models.

4. Results

4.1. Vowel formant entrainment

To investigate whether vowel formant adjustment was influenced by subject group and time period, first, a linear mixed effects model was fitted with the CRDiff value as the response variable, the time period and group as fixed effects, and subject and keyword as random effects. The fixed effects and their interaction terms were tested using likelihood ratio tests by adding each variable one at a time for a comparison until the optimal model was chosen.

Regarding RCDiff in F1, only (Time) Period showed a significant effect (Df = 3, $p = 0.01^{**}$). Neither adding Group nor the two-way interaction of Period and Group significantly improved the model. According to Figure 3, both groups of children reduced RCDiff of F1 in the early period. Marginally significant differences were found in comparison between the early and baseline periods ($t = -1.68$, $p = 0.09$; $\beta' = -0.06$, 95%, CI [−0.13, 0.01]). Autistic children further reduced the RCDiff in the late production, but TD children did not, as indicated by an increase of RCDiff in late period. No significant difference was registered between post-task production and baseline, suggesting that the entrainment only occurred during the interaction.

The adjustment of RCDiff in F2 was more evident. Statistical modeling showed that, by adding Period as a fixed effect, the model significantly improved ($p < 0.001^{***}$). By adding Group and Group * Period interaction, the model improved with marginal significance ($p = 0.07$). Early ($t = -3.74$; $p < 0.001^{***}$; $\beta' = -0.22$, 95% CI [−0.34, −0.11]) and late ($t = -5.02$; $p < 0.001^{***}$; $\beta' = -0.30$, 95% CI [−0.42, −0.18]) production showed significant reduction of RCDiff of F2 compared to the baseline, suggesting that both groups of children significantly converged toward the vowel formant of the robot during interaction in terms of F2. We performed *post-hoc* tests using the "emmeans package" (Lenth et al., 2018) to further interpret the Group * Period interaction, and used the estimated marginal means difference (EMMdiff) measures to report the effect size by the "eff_size" function of this package. The *post-hoc* analysis showed a significant reduction in RCDiff of F2 in both early ($t = 3.74$, $p < 0.01^{**}$; $\Delta = 0.24$, 95% CI [0.10, 0.39]) and late ($t = 5.02$, $p < 0.001^{***}$; $\Delta = 0.33$, 95% CI [0.18, 0.48]) periods in the autistic group as compared to the baseline. The TD group showed a trend of reduction in RCDiff of F2, but the reduction did not reach significance. In addition, significant increases of CRDiff (i.e., indicating increasing divergence from the robot) in the post-production relative to early periods ($t = -3.65$, $p < 0.01^{**}$; $\Delta = -0.24$, 95% CI [−0.38, −0.09]) as well as in the post-production relative to late periods ($t = -4.95$, $p < 0.001^{***}$; $\Delta = -0.32$, 95% CI [−0.47, −0.18]) were registered for the autistic group, while TD group showed a significant increase of CRDiff in the

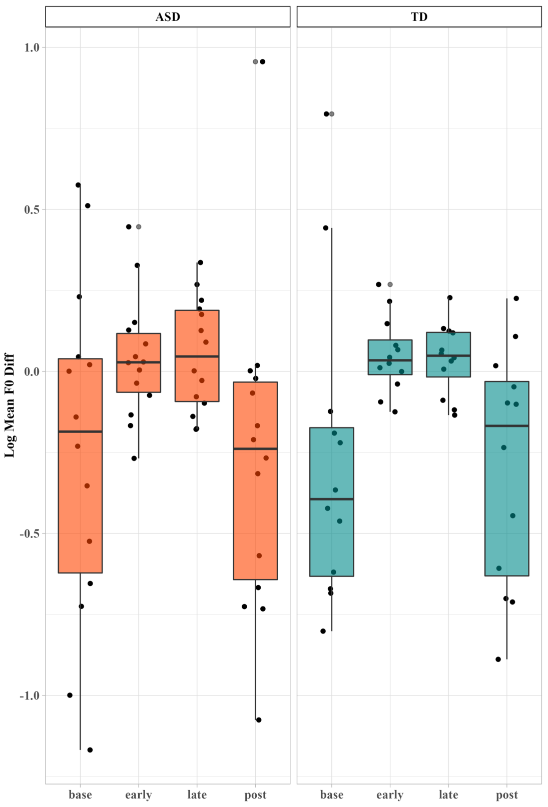


FIGURE 4
Log mean *f0* difference between the children and the robot across different time periods. Each dot represents a child's production.

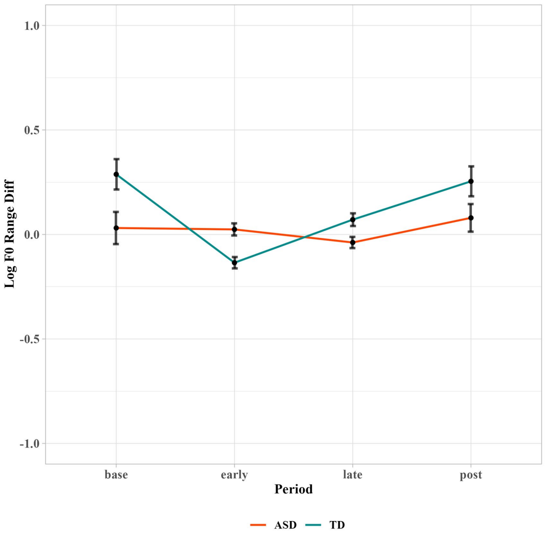


FIGURE 5
Log *f0* range difference between the children and robot across different time periods.

post-production relative to early periods ($t = -3.07$, $p = 0.04^{*}$; $\Delta = -0.22$, 95% CI [−0.38, −0.06]), suggesting that the convergence toward the robot occurred specifically during the interaction with the robot interlocutor rather than an adjustment as a result of time/practice with this speech production activity. These results indicate that F2 entrainment occurred more prominently during the early period and started to reduce in the late period for TD children. In contrast, F2 entrainment occurred more prominently in late period for autistic children. Similar to F1, no significant difference between the post-production and the baseline was registered for CRDiff of F2.

In summary, these autistic children entrained in a more gradual way. The degree of entrainment was larger in the late production than the early production in the autistic group. As for TD children, they entrained more prominently in early period and less prominently in late period. The entrainment did not persist in post-task production for either group.

4.2. Prosody entrainment

4.2.1. Mean *f0*

A linear mixed effects model was fitted to test the fixed effect of Period (i.e., early and late), Group and their interaction on CRDiff of log mean *f0* with subject as a random effect. We performed the same modeling procedure as used to analyze vowel formant. Only the fixed effect of Period reached significance ($p < 0.001^{***}$). As we can see from Figure 4, both groups of children reduced the difference of mean *f0* when interacting with the robot (early: $\beta' = 0.30$, 95% CI [0.23, 0.38]; late: $\beta' = 0.33$, 95% CI [0.25, 0.40]), and the differences increased in the post-interaction period ($\beta' = -0.01$, 95% CI [−0.11, 0.09]). No difference between the early and late periods was found, indicating that they entrained as soon as interacting with the robot and that the entrainment remained throughout the tasks.

4.2.2. *F0* range

Regarding the log *f0* range, the linear mixed effects model improved significantly by adding Period ($p < 0.01^{**}$) and the two-way interaction of Period and Group ($p < 0.01^{**}$) as fixed effects. *Post-hoc* analyses to interpret the significant interaction of Period and Group showed that the contribution mainly came from the TD group. The TD group reduced the difference in *f0* range significantly in early period ($t = 4.9$, $p < 0.001^{***}$; $\Delta = 0.28$, 95% CI [0.17, 0.39]) as compared to the baseline. They further adjusted *f0* range difference in late period as compared to early period ($t = -5.0$, $p < 0.001^{***}$; $\Delta = -0.14$, 95% CI [−0.19, −0.08]). By contrast, autistic children did not show much entrainment in terms of *f0* range, as shown in Figure 5. The differences in the *f0* range between the robot and children remained similar when during interactions, suggesting that the participants' *f0* range were not affected by the interaction. The group difference reached significance in early period ($t = 3.98$, $p < 0.01^{*****}$; $\Delta = 0.11$, 95% CI [0.05, 0.16]), suggesting that at the baseline level, the two groups did not significantly differ in *f0* range difference, but as soon as the TD children started to interact with the robot, their entrainment enlarged the group difference.

As we can see from Figure 5, autistic children showed a similar *f0* range with the robot throughout the time periods, even at baseline prior to interacting with the robot. In order to further our understanding of their *f0* range entrainment, we calculated the standard deviation over each subject's mean *f0* range in each period, as shown in Table 2. Autistic children showed larger standard deviation in baseline, early, and late periods than TD children. We also noticed a slight fluctuation of f0 range difference from early period to late period for autistic children. We then calculated the number of autistic children showing a reduction of mean *f0* range difference from early to late periods (i.e., more entrainment in late period than early period). Five out of fourteen autistic children exhibited a reduction of differences in late periods while seven out of twelve TD children showed a reduction. This indicated that there were indeed a few autistic children showing phonetic entrainment of *f0* range during interaction. The large individual variation suggested that the reasons behind their lack of *f0* range entrainment were complicated. It is challenging for some autistic children to entrain *f0* range, but not others.

4.2.3. Summary

To summarize, both autistic and TD children exhibited a reduction of the mean *f0* differences between them and the robot during their interaction. Regarding *f0* range, our results showed that TD children exhibited reduction of the *f0* range differences from the robot when interacting with the robot, while autistic children showed more individual differences in the phonetic entrainment of *f0* range and did not exhibit adjustment of *f0* range differences from the robot as a group.

5. Discussion

We present the first empirical study using a social robot as an interlocutor to investigate whether and how children with and without ASD showed phonetic entrainment in conversations. Since having a social robot interlocutor with speech features and social traits controlled may facilitate phonetic entrainment and its detection in autistic individuals, we expect autistic children may show phonetic entrainment in a more controlled phonetic and social environment, but they may still be different from TD children.

Our study aimed to conduct a more comprehensive investigation examining phonetic entrainment both in vowels and prosody. Specifically, we examined vowel formants and fundamental frequency in the speech production of a group of autistic children and compared these measurements with their TD peers to identify whether or not they would show TD-like phonetic entrainment behaviors. Consistent with our predictions, though autistic children showed some phonetic entrainment, they still exhibited some deficits. Autistic children showed comparable vowel formant entrainment as TD children. Both groups entrained more on F2 than F1. Regarding prosody entrainment, autistic children also showed comparable mean *f0* entrainment as their TD peers. However, while their TD peers showed *f0* range entrainment, the group of autistic children did not exhibit significant convergence toward the interlocutor in terms of *f0* range adjustment, suggesting that entrainment of *f0* range was more challenging and vulnerable for these autistic children even in a more controlled situation.

The fact that autistic children produced vowels in a more extreme way has been documented. In the baseline and post-interaction production, autistic children did produce vowels with larger F2 values, consistent with the results reported by Mohanta and Mittal (2022). These previous findings were interpreted by the authors as attributable to the atypical oral and pharyngeal constriction in autistic individuals when they produced vowels. Nevertheless, our study demonstrated that this atypical mechanism of vowel production did not affect entrainment of vowel formants with an interlocutor producing more controlled speech. Our findings provide support for the claim by Kissine and Geelhand (2019) that autistic population might attend more to the precision of pronunciation. The observed extreme vowel production of autistic population might be due to their overact of articulatory gesture to approach a more precise pronunciation. In our study, the robot produced standard pronunciation of English vowels in a consistent manner, which might be preferred by autistic children and thus triggered their entrainment.

In addition to the findings of vowel formants, autistic children entrained their mean *f0* comparably to their TD peers. This result was inconsistent with some previous studies, where neither autistic nor TD children showed prosody entrainment (Hogstrom et al., 2018; Wynn et al., 2018). As most studies computed the differences between conversation partners to indicate phonetic entrainment, it is very likely that their findings about entrainment or lack of entrainment was actually driven by the adjustment of their interlocutors. In addition, the exaggerated production of the autistic population might trigger atypical judgment of their interlocutors, leading to these interlocutors' adjustment being more unpredictable. They might entrain to compensate for the larger difference between themselves and the autistic individuals, or they might manipulate their phonetic features away from autistic population because of their atypical production. In our study, the interlocutor (i.e., the social robot) did not adjust its phonetic features no matter whom the robot was talking to. Any manipulation of differences between the dyads came from the child. The consistency of the robot interlocutor made the entrainment more detectable. Another possible reason for this discrepancy in our findings and previous findings could be that social robots were more attractive to children. Previous studies have shown that phonetic entrainment is not merely an automatic imitation process but is mediated by social factors (Coles-Harris, 2017). According to Communication Accommodation Theory (CAT; Giles and Ogay, 2007), positive perception of a conversation partner would reduce the social distance between the individual and the interlocutor and motivate an individual to show entrainment. The attractiveness of a social robot might have reduced its social distance with children and motivated them to entrain phonetically. Yet, an anonymous reviewer pointed out another possibility that the entrainment of prosody might be motivated by a desire for being better understood by the robot. Previous studies have shown that autistic children could

TABLE 2 Standard deviation of *f0* range difference between children and robot in each time period.

Group	Baseline	Early interaction	Late interaction	Post-interaction
ASD	45.97	25.76	28.10	22.15
TD	24.50	22.82	19.32	24.68

differentiate a human voice and a robotic voice (Stanton et al., 2008). But the manipulation of mean *f0* from the robotic voice did not significantly affect children's performance in a learning task (Molenaar et al., 2021). It is possible that participants entrained to the robot to make themselves understood better. Future studies using both a more human-like voice with natural prosody and a robotic voice without natural prosody may help differentiate the underlying reasons for entrainment in speech prosody. If participants entrain to both a robotic voice without natural prosody, it is likely that they are attempting to build a relationship as lack of natural prosody will not lead to better understandability.

In fact, studies have reported that autistic children showed more interest in interacting with social robots than human beings (Dautenhahn and Werry, 2004; Barakova et al., 2015). Autistic individuals have also been shown to have less interest in human speech voice (Yu and Wang, 2021) and be less able to orient their attention to the human sounds than their TD peers (Čeponienė et al., 2003). It is likely that social robot speaks with a controlled and consistent voice, which may aid their perception and facilitate their phonetic entrainment. On the other hand, autistic individuals have been found to experience multiple difficulties in processing social information such as emotion evaluation (Embregts and Van Nieuwenhuijzen, 2009) and voice identification (Lerner et al., 2013). Previous studies have reported that social robots were usually treated as a human-like category (Eyssel and Kuchenbrandt, 2012; Cohn et al., 2020) and that people tended to compare them with human-beings and evaluate them in a social way—for example, evaluating the 'membership' of a robot from the cues of its gender and age (Eyssel and Kuchenbrandt, 2012). In spite of this, the social features of robots are far simpler than humans. Their social complexities demonstrated in interaction, such as facial expressions, social responses, are more limited and controllable. The relatively consistent social information in social robots can reduce the processing load for autistic children. The controlled voice and consistent social information together could have contributed to a more tractable and structured conversational environment, making it more predictable for autistic individuals and easier for them to demonstrate phonetic entrainment.

Apart from documenting comparable phonetic entrainment between the two groups in vowel formants and mean *f0* in this phonetically and socially controlled communication environment, the current study also documented that these autistic children did not show significant *f0* range entrainment as what the TD children exhibited, even with a partner of more controlled speech and social traits. Recall that we also reported larger individual variations of *f0* range differences in the autistic group relative to TD group. It can be inferred that their entrainment behaviors in terms of *f0* range may show more variation compared to the TD group. But as a group, their *f0* range entrainment is not as robust as the TD group. This is in line with findings by Lehnert-LeHouillier et al. (2020) that a few autistic children showing phonetic entrainment, but their statistical results indicated that the autistic children, as a group, did not show comparable entrainment with the TD group. This is also consistent with the emerging literature suggesting a high heterogeneity within the autistic population (Schadenberg et al., 2020). Future research is needed to examine factors that may predict why some autistic individuals are better than the others in phonetic entrainment.

One thing that needs to be noted is that our study examined second language entrainment. Our findings are in line with previous findings of neurotypical L2 speakers, who tended to entrain toward the more prestigious variety when interacting with native speakers (Gnevsheva et al., 2021). In our study, the social robot spoke standard American English, which was more likely to trigger more robust phonetic entrainment from our children participants who spoke English as second language. A few studies have reported signs of relatively intact L2 in autistic populations (Práinsson, 2012; Agostini and Best, 2015), but no study examined phonetic entrainment of their L2 so far. We cannot rule out the possibility that the lack of phonetic entrainment found in previous studies is due to their problems of linguistic entrainment skills in their L1, whereas their L2 entrainment skill might remain relatively intact.

The present study has some limitations, which might need to be addressed in future research. Due to poor speech recognition in Cantonese by the robot, we did not examine phonetic entrainment of the participants' first language (i.e., Cantonese), which may provide some more direct evidence for phonetic entrainment in human-robot interaction than the evidence in the current study on their L2. It remains to be explored with a Cantonese-speaking social robot in future studies. Moreover, we observed larger individual variation of *f0* range entrainment by autistic children; yet, as the sample size is relatively small, we did not further investigate the contributing factors of their variation. Future work can include more participants and examine the interaction of severity of autism symptoms and phonetic entrainment. In addition, more age groups can be included to obtain a more comprehensive developmental trajectory of children's phonetic entrainment.

6. Conclusion

To conclude, we present the first study investigating phonetic entrainment in autistic children when they interacted with a social robot. The new evidence suggested that these autistic children could entrain phonetically similarly to their TD peers when the interlocutor was controlled in both phonetic and social features. On the other hand, these autistic children did not show entrainment in *f0* range as a group compared to their TD peers, suggesting that phonetic entrainment in *f0* range could be more challenging and vulnerable in autistic children. This study deepens our understanding of autistic children's conversation behaviors and has implications in designing trainings for autistic children using social robots.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by Departmental Research Committee of the Hong Kong Polytechnic University. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

Author contributions

SC and YH contributed to the conceptualization and experimental design of the study. FZ and TT helped with recruiting participants and data collection. YH ran the statistical analysis under the guidance of SC and wrote the first draft of the manuscript. SC and AC revised the manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by Department of Chinese and Bilingual Studies, Faculty of Humanities, the Hong Kong Polytechnic University [UAK7,1-ZVRT; 1-ZE0D; 1-W08C, 1-TA41]. It is also partly supported by the grant from Standing Committee on Language Education and Research (SCOLAR), Education Bureau, HKSAR government [K-ZB2P] and RGC direct allocation grant [A-PB1B].

Acknowledgments

The authors acknowledge the contributions of Lok Mang Cheung, Shui Shan Yee, Wai Ling Mak, and Yuen Yu Fung to the data collection process. The authors are particularly grateful to the children who participated in this research as well as their parents for allowing them to participate.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1128976/full#supplementary-material

References

Agostini, T. G., and Best, C. T. (2015). Exploring processability theory-based hypotheses in the second language acquisition of a child with autism spectrum disorder. *Grammatical Development in Second Languages: Exploring the Boundaries of Processability Theory*. Italy: European Second Language Association, 275–290 (Assessed December 19, 2022).

Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. (2012). Furhat: a Back-projected human-like robot head for multiparty human-machine interaction. *Cogn. Behav. Syst. 7403*, 114–130). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-34584-5_9

Amazon Polly Developer Guide (2023). Available at: https://docs.aws.amazon.com/polly/latest/dg/what-is.html (Accessed May 1, 2023).

American Psychiatric Association (2022). *Diagnostic and statistical manual of mental disorders*, *5*. Virginia: American Psychiatric Association.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Lang. Speech* 34, 351–366. doi: 10.1177/002383099103400404

Baker, R., and Hazan, V. (2011). DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behav. Res. Methods* 43, 761–770. doi: 10.3758/s13428-011-0075-y

Barakova, E. I., Bajracharya, P., Willemsen, M., Lourens, T., and Huskens, B. (2015). Long-term LEGO therapy with humanoid robot for children with ASD. *Expert. Syst.* 32, 698–709. doi: 10.1111/exsy.12098

Ben-Shachar, M. S., Lüdecke, D., and Makowski, D. (2020). Effectsize: estimation of effect size indices and standardized parameters. *J. Open Source Softw.* 5:2815. doi: 10.21105/joss.02815

Boersma, P., and Weenink, D. (2018). Praat: Doing phonetics by computer [computer program]. Available at: http://www.praat.org/ (Assessed December 19, 2022).

Borrie, S. A., and Delfino, C. R. (2017). Conversational entrainment of vocal fry in young adult female American English speakers. *J. Voice* 31, 513–e25. doi: 10.1016/j.jvoice.2016.12.005

Borrie, S. A., and Liss, J. M. (2014). Rhythm as a coordinating device: entrainment with disordered speech. *J. Speech Lang. Hear. Res.* 57, 815–824. doi: 10.1044/2014_JSLHR-S-13-0149

Carrow-Woolfolk, E. (2017). *CASL: Comprehensive assessment of spoken language* Torrance, CA: American Guidance Services.

Cassell, J., Gill, A., and Tepper, P. (2007). "Coordination in conversation and rapport" in *Proceedings of the workshop on embodied language processing* (Prague: Czech Republic), 41–50.

Čeponienė, R., Lepistö, T., Shestakova, A., Vanhala, R., Alku, P., Näätänen, R., et al. (2003). Speech–sound-selective auditory impairment in children with autism: they can perceive but do not attend. *Proc. Natl. Acad. Sci.* 100, 5567–5572. doi: 10.1073/pnas.0835631100

Chartrand, T. L., and Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *J. Pers. Soc. Psychol.* 76:893. doi: 10.1037/0022-3514.76.6.893

Cohn, M., Sarian, M., Predeck, K., and Zellou, G. (2020). Individual variation in language attitudes toward voice-AI: the role of listeners' autistic-like traits. *Proc. Interspeech*, 1813–1817. doi: 10.21437/Interspeech.2020-1339

Coles-Harris, E. H. (2017). Perspectives on the motivations for phonetic convergence. *Lang. Linguist. Compass* 11:e12268. doi: 10.1111/lnc3.12268

Dautenhahn, K., and Werry, I. (2004). Towards interactive robots in autism therapy: background, motivation and challenges. *Pragmat. Cogn.* 12, 1–35. doi: 10.1075/pc.12.1.03dau

Embregts, P. J. C. M., and Van Nieuwenhuijzen, M. (2009). Social information processing in boys with autistic spectrum disorder and mild to borderline intellectual disabilities. *J. Intellect. Disabil. Res.* 53, 922–931. doi: 10.1111/j.1365-2788.2009.01204.x

Eyssel, F., and Kuchenbrandt, D. (2012). Social categorization of social robots: anthropomorphism as a function of robot group membership. *Br. J. Soc. Psychol.* 51, 724–731. doi: 10.1111/j.1365-2788.2009.01204.x

Gessinger, I., Raveh, E., Steiner, I., and Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: behavior of groups and individuals in speech shadowing. *Speech Comm.* 127, 43–63. doi: 10.1016/j.specom.2020.12.004

Giles, H., and Ogay, T. (2007). "Communication accommodation theory" in *Explaining communication: Contemporary theories and exemplars*. eds. B. B. Whaley and W. Samter (Mahwah, NJ: Lawrence Erlbaum), 293–310.

Gill, S. P. (2012). Rhythmic synchrony and mediated interaction: towards a framework of rhythm in embodied interaction. *AI Soc.* 27, 111–127. doi: 10.1007/s00146-011-0362-2

Gnevsheva, K., Szakay, A., and Jansen, S. (2021). Phonetic convergence across dialect boundaries in first and second language speakers. *J. Phon.* 89:101110. doi: 10.1016/j.wocn.2021.101110

Green, H., and Tobin, Y. (2009). Prosodic analysis is difficult… but worth it: a study in high functioning autism. *Int. J. Speech Lang. Pathol.* 11, 308–315. doi: 10.1080/17549500903003060

Helt, M. S., Eigsti, I. M., Snyder, P. J., and Fein, D. A. (2010). Contagious yawning in autistic and typical development. *Child Dev.* 81, 1620–1631. doi: 10.1111/j.1467-8624.2010.01495.x

Hogstrom, A., Theodore, R. M., Canfield, A., Castelluccio, B., Green, J., Irvine, C., et al. (2018). "Reduced phonemic convergence in autism Spectrum disorder" in *Proceedings of the 40th annual conference of the cognitive science society* (Madison), 1797–1802.

Hwang, J., Brennan, S. E., and Huffman, M. K. (2015). Phonetic adaptation in non-native spoken dialogue: effects of priming and audience design. *J. Mem. Lang.* 81, 72–90. doi: 10.1016/j.jml.2015.01.001

Ilari, B. (2015). Rhythmic engagement with music in early childhood: a replication and extension. *J. Res. Music. Educ.* 62, 332–343. doi: 10.1177/0022429414555984

Jacewicz, E., Fox, R. A., and Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *J. Acoust. Soc. Am.* 128, 839–850. doi: 10.1121/1.3459842

Kelley, E., Paul, J. J., Fein, D., and Naigles, L. R. (2006). Residual language deficits in optimal outcome children with a history of autism. *J. Autism Dev. Disord.* 36, 807–828. doi: 10.1007/s10803-006-0111-4

Kissine, M., and Geelhand, P. (2019). Brief report: acoustic evidence for increased articulatory stability in the speech of adults with autism spectrum disorder. *J. Autism Dev. Disord.* 49, 2572–2580. doi: 10.1007/s10803-019-03905-5

Kissine, M., Geelhand, P., Philippart De Foy, M., Harmegnies, B., and Deliens, G. (2021). Phonetic inflexibility in autistic adults. *Autism Res.* 14, 1186–1196. doi: 10.1002/aur.2477

Kuhl, P. K., Coffey-Corina, S., Padden, D., and Dawson, G. (2005). Links between social and linguistic processing of speech in preschool children with autism: behavioral and electrophysiological measures. *Dev. Sci.* 8, F1–F12. doi: 10.1111/j.1467-7687.2004.00384.x

Kumazaki, H., Muramatsu, T., Yoshikawa, Y., Matsumoto, Y., Ishiguro, H., Kikuchi, M., et al. (2020). Optimal robot for intervention for individuals with autism spectrum disorders. *Psychiatry Clin. Neurosci.* 74, 581–586. doi: 10.1111/pcn.13132

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). Lmertest package: tests in linear mixed effects models. *J. Stat. Softw.* 82, 1–26. doi: 10.18637/jss.v082.i13

Lee, C. C., Black, M., Katsamanis, A., Lammert, A. C., Baucom, B. R., Christensen, A., et al. (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples. *Proc. Interspeech*, 793–796. doi: 10.21437/Interspeech

Lehnert-LeHouillier, H., Terrazas, S., and Sandoval, S. (2020). Prosodic entrainment in conversations of verbal children and teens on the autism spectrum. *Front. Psychol.* 11:582221. doi: 10.3389/fpsyg.2020.582221

Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2018). Emmeans: estimated marginal means, aka least-squares means. R package version 1, 3.

Lerner, M. D., McPartland, J. C., and Morris, J. P. (2013). Multimodal emotion processing in autism spectrum disorders: an event-related potential study. *Dev. Cogn. Neurosci.* 3, 11–21. doi: 10.1016/j.dcn.2012.08.005

Levitan, R., Beňuš, Š., Gravano, A., and Hirschberg, J. (2015). Acoustic-prosodic entrainment in Slovak, Spanish, English and Chinese: a cross-linguistic comparison. In *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, 325–334.

Levitan, R., Gravano, A., and Hirschberg, J. B. (2011). Entrainment in speech preceding backchannels. *Proceedings of ACL/HLT* 2011. doi: 10.7916/D89Z9DCS

Levitan, R., and Hirschberg, J. B. (2011). Measuring acoustic-prossodic entrainment with respect to multiple levels and dimensions. *Proc. Interspeech* 2011, 3081–3084. doi: 10.7916/D8V12D8F

Lewandowski, E. M., and Nygaard, L. C. (2018). Vocal alignment to native and non-native speakers of English. *J. Acoust. Soc. Am.* 144:620 (2018). doi: 10.1121/1.5038567

Lyakso, E., Frolova, O., and Grigorev, A. (2016). "A comparison of acoustic features of speech of typically developing children and children with autism spectrum disorders," in *International conference on speech and computer* (Cham: Springer), 43–50.

Marchi, E., Ringeval, F., and Schuller, B. (2014). "Voice-enabled assistive robots for handling autism spectrum conditions: an examination of the role of prosody," in *Speech and automata in health care* (Berlin, Germany: De Gruyter).

Mathersul, D., McDonald, S., and Rushby, J. A. (2013). Automatic facial responses to briefly presented emotional stimuli in autism spectrum disorder. *Biol. Psychol.* 94, 397–407. doi: 10.1016/j.biopsycho.2013.08.004

McPherson, M. J., and McDermott, J. H. (2018). Diversity in pitch perception revealed by task dependence. *Nat. Hum. Behav.* 2, 52–66. doi: 10.1038/s41562-017-0261-8

Mohanta, A., and Mittal, V. K. (2022). Analysis and classification of speech sounds of children with autism spectrum disorder using acoustic features. *Comput. Speech Lang.* 72:101287. doi: 10.1016/j.csl.2021.101287

Molenaar, B., Soliño Fernández, B., Polimeno, A., Barakova, E., and Chen, A. (2021). Pitch it right: using prosodic entrainment to improve robot-assisted foreign language learning in school-aged children. *Multim. Technol. Interact.* 5:76. doi: 10.3390/mti5120076

Nadig, A., and Shaw, H. (2012). Acoustic and perceptual measurement of expressive prosody in high-functioning autism: increased pitch range and what it means to listeners. *J. Autism Dev. Disord.* 42, 499–511. doi: 10.1007/s10803-011-1264-3

Nakano, T., Kato, N., and Kitazawa, S. (2011). Lack of eyeblink entrainments in autism spectrum disorders. *Neuropsychologia* 49, 2784–2790. doi: 10.1016/j.neuropsychologia.2011.06.007

Nenkova, A., Gravano, A., and Hirschberg, J. (2008). "High-frequency word entrainment in spoken dialogue" in *Proceedings of the 46th annual meeting of the Association for Computational Linguistics on human language technologies*, 169–172. doi: 10.7916/D8TM7KF6

Ochi, K., Ono, N., Owada, K., Kojima, M., Kuroda, M., Sagayama, S., et al. (2019). Quantification of speech and synchrony in the conversation of adults with autism spectrum disorder. *PLoS One* 14:e0225377. doi: 10.1371/journal.pone.0225377

Patel, S. P., Nayar, K., Martin, G. E., Franich, K., Crawford, S., Diehl, J. J., et al. (2020). An acoustic characterization of prosodic differences in autism spectrum disorder and first-degree relatives. *J. Autism Dev. Disord.* 50, 3032–3045. doi: 10.1007/s10803-020-04392-9

Pettinato, M., Tuomainen, O., Granlund, S., and Hazan, V. (2016). Vowel space area in later childhood and adolescence: effects of age, sex and ease of communication. *J. Phon.* 54, 1–14. doi: 10.1016/j.wocn.2015.07.002

Phillips-Silver, J., Aktipis, C. A., and Bryant, G. A. (2010). The ecology of entrainment: foundations of coordinated rhythmic movement. *Music. Percept.* 28, 3–14. doi: 10.1525/mp.2010.28.1.3

Práinsson, K. Ó. (2012). *Second language acquisition and autism*, BA essay University of Iceland, Iceland.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Raven, J. (2003). "Raven progressive matrices" in *Handbook of nonverbal assessment* (Boston, MA: Springer), 223–237.

Robins, B., Dautenhahn, K., Boekhorst, R. T., and Billard, A. (2005). Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills? *Univ. Access Inf. Soc.* 4, 105–120. doi: 10.1007/s10209-005-0116-3

Schadenberg, B. R., Reidsma, D., Heylen, D. K., and Evers, V. (2020). Differences in spontaneous interactions of autistic children in an interaction with an adult and humanoid robot. *Front. Robot. AI* 7:28. doi: 10.3389/frobt.2020.00028

Sharda, M., Subhadra, T. P., Sahay, S., Nagaraja, C., Singh, L., Mishra, R., et al. (2010). Sounds of melody—pitch patterns of speech in autism. *Neurosci. Lett.* 478, 42–45. doi: 10.1016/j.neulet.2010.04.066

Shriberg, L. D., Paul, R., McSweeny, J. L., Klin, A., Cohen, D. J., and Volkmar, F. R. (2001). Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. *J. Speech Lang. Hear. Res.* 44, 1097–1115. doi: 10.1044/1092-4388(2001/087)

Stanley, J., and Lipani, L. (2019). Automatic Formant Extraction in Praat. Available at: https://joeystanley.com/downloads/191002-formant_extraction. (Accessed 1 January 2022).

Stanton, C. M., Kahn, P. H., Severson, R. L., Ruckert, J. H., and Gill, B. T. (2008). "Robotic animals might aid in the social development of children with autism" in *In 2008 3rd ACM/IEEE international conference on human-robot interaction (HRI), IEEE*, 271–278. doi: 10.1145/1349822.1349858

Tsang, G. J. Y., Dana, E. L., Farbood, M. M., and Levi, S. V. (2018). Musical training and the perception of phonetic detail in a shadowing task. *J. Acoust. Soc. Am.* 143:1922. doi: 10.1121/1.5036272

Tuomainen, O., Taschenberger, L., Rosen, S., and Hazan, V. (2022). Speech modifications in interactive speech: effects of age, sex and noise type. *Philos. Trans. R. Soc. B* 377:20200398. doi: 10.1098/rstb.2020.0398

Upitis, R. (1987). Children's understanding of rhythm: the relationship between development and music training. *Psychomusicol. J. Res. Music Cogn.* 7:41. doi: 10.1037/h0094187

van Engen, K. J., Baese-Berk, M., Baker, R. E., Choi, A., Kim, M., and Bradlow, A. R. (2010). The wildcat corpus of native-and foreign-accented English: communicative efficiency across conversational dyads with varying language alignment profiles. *Lang. Speech* 53, 510–540. doi: 10.1177/00238309103724

Wynn, C. J., Borrie, S. A., and Sellers, T. P. (2018). Speech rate entrainment in children and adults with and without autism spectrum disorder. *Am. J. Speech Lang. Pathol.* 27, 965–974. doi: 10.1044/2018_AJSLP-17-0134

Xia, Z., Levitan, R., and Hirschberg, J. B. (2014). Prosodic entrainment in mandarin and English: a cross-linguistic comparison. *In Proceedings of the International Conference on Speech Prosody* 2014, 65–69. doi: 10.21437/SpeechProsody.2014-1

Yavas, M. S. (2011). *Applied English phonology*. *2nd* Edn. United Kingdom: Wiley-Blackwell.

Yoshimura, S., Sato, W., Uono, S., and Toichi, M. (2015). Impaired overt facial mimicry in response to dynamic facial expressions in high-functioning autism spectrum disorders. *J. Autism Dev. Disord.* 45, 1318–1328. doi: 10.1007/s10803-014-2291-7

Yu, L., and Wang, S. (2021). Aberrant auditory system and its developmental implications for autism. *Sci. China Life Sci.* 64, 861–878. doi: 10.1007/s11427-020-1863-6

Zhu, X. (2005). 上海声调实验录 *[an experimental study in Shanghai tones]*. Shanghai: Shanghai Educational Publishing House.

TYPE Original Research
PUBLISHED 22 June 2023
DOI 10.3389/fpsyg.2023.1129406

Analysis of emotional prosody as a tool for differential diagnosis of cognitive impairments: a pilot research

Chorong Oh[1]*, Richard Morris[2], Xianhui Wang[3] and Morgan S. Raskin[2]

[1]School of Rehabilitation and Communication Sciences, Ohio University, Athens, OH, United States, [2]School of Communication Science and Disorders, Florida State University, Tallahassee, FL, United States, [3]School of Medicine, University of California Irvine, Irvine, CA, United States

Check for updates

OPEN ACCESS

EDITED BY
Ratree Wayland,
University of Florida, United States

REVIEWED BY
Emilie Brotherhood,
University College London, United Kingdom
Bin Huang,
BrainCheck Inc., United States

*CORRESPONDENCE
Chorong Oh
✉ ohc@ohio.edu

RECEIVED 21 December 2022
ACCEPTED 26 May 2023
PUBLISHED 22 June 2023

CITATION
Oh C, Morris R, Wang X and Raskin MS (2023) Analysis of emotional prosody as a tool for differential diagnosis of cognitive impairments: a pilot research.
Front. Psychol. 14:1129406.
doi: 10.3389/fpsyg.2023.1129406

Introduction: This pilot research was designed to investigate if prosodic features from running spontaneous speech could differentiate dementia of the Alzheimer's type (DAT), vascular dementia (VaD), mild cognitive impairment (MCI), and healthy cognition. The study included acoustic measurements of prosodic features (Study 1) and listeners' perception of emotional prosody differences (Study 2).

Methods: For Study 1, prerecorded speech samples describing the *Cookie Theft* picture from 10 individuals with DAT, 5 with VaD, 9 with MCI, and 10 neurologically healthy controls (NHC) were obtained from the DementiaBank. The descriptive narratives by each participant were separated into utterances. These utterances were measured on 22 acoustic features *via* the Praat software and analyzed statistically using the principal component analysis (PCA), regression, and Mahalanobis distance measures.

Results: The analyses on acoustic data revealed a set of five factors and four salient features (i.e., pitch, amplitude, rate, and syllable) that discriminate the four groups. For Study 2, a group of 28 listeners served as judges of emotions expressed by the speakers. After a set of training and practice sessions, they were instructed to indicate the emotions they heard. Regression measures were used to analyze the perceptual data. The perceptual data indicated that the factor underlying pitch measures had the greatest strength for the listeners to separate the groups.

Discussion: The present pilot work showed that using acoustic measures of prosodic features may be a functional method for differentiating among DAT, VaD, MCI, and NHC. Future studies with data collected under a controlled environment using better stimuli are warranted.

KEYWORDS
dementia, mild cognitive impairment, emotion, prosody, acoustic analysis, listener perception, diagnosis

Introduction

Currently, diagnosis of cognitive impairments relies heavily on invasive (e.g., lumbar puncture) and/or expensive (e.g., neuroimaging panel) biomarker tests (López-de-Ipiña et al., 2015). The results of biomarker tests, primarily obtained using invasive lumbar punctures, depend significantly on the patient's physical health and age, which decreases the efficacy of the method (Maclin et al., 2019). Expensive neuroimaging lacks definitive characteristics with significant diagnostic value (Filippi et al., 2012) which decreases the diagnostic accuracy, and many patients experience claustrophobia, discomfort, or behavioral problems during the imaging sessions and cannot tolerate them (Bonifacio and Zamboni, 2016). These issues lead to the decreased diagnostic accuracy and eventually the overall costs for dementia care increase not only because of the high cost and invasive nature of the exams but also because of the extensive clinical testing that often takes place while individuals seek opinions from multiple providers regarding the source of their symptoms before ultimately reaching a provider in a facility that has access to these diagnostic exams. The extended time increases both personal and monetary costs associated with dementia diagnosis, which subsequently increases financial burden on people with cognitive impairment, families, and society and also delays the initiation of proper care.

Speech and language production requires coordination among highly complicated and calibrated brain systems, including but not limited to Broca's and Wernicke's areas. When the coordination is not accomplished properly due to a brain disease or accident, it may yield significant changes in the person's speech and/or language functions. People with cognitive impairment such as dementia demonstrate various speech and language deficits. While linguistic deficits such as word finding difficulty and agrammatism are well documented and have been used to identify early-stage cognitive declines (e.g., Lundholm Fors et al., 2018; Calzá et al., 2021), data on speech deficits in people with different types of cognitive impairment are limited. It should also be noted that speech and language deficits are not clearly distinguished in the dementia literature; often, language deficits are misinterpreted as speech deficits or the two terms (i.e., speech impairment and language impairment) are used interchangeably. However, the distinction between speech and language impairments is critical to understanding any impaired communication functioning and for making more accurate diagnoses and creating appropriate management plans.

The use of vocal biomarker may provide useful information for diagnosis and monitoring of different diseases/disorders as well as for phenotyping a condition (Fagherazzi et al., 2021). Among many voice features, prosody is an aspect of speech that consists of perceptible suprasegmental modulations of vocal pitch, syllable length, loudness, and pauses (Odell et al., 1991). These modulations deliver the speaker's meaning beyond the literal meaning of the utterance and give the listener clues to interpret the connotative meaning intended by the speaker (Hupp and Junger, 2013). The manipulation of prosody requires a wide range of interhemispheric cerebral networks, which are impaired in people with cognitive impairment to different extents depending on the type of condition (Lian et al., 2018; Qi et al., 2019; Cheung and Mak, 2020). For instance, the accumulation of amyloid fibrils decreases interhemispheric functional connectivity (IFC) in visual network for dementia of the Alzheimer's type (DAT) while it increases with the IFC in default mode network, central executive network, sensory motor network, and dorsal attention network for vascular dementia (VaD) (Cheung and Mak, 2020). Such differences suggest that prosody, of which manipulation is completed *via* interhemispheric connectivity, may be an effective, reliable, and low-cost method to differentiate cognitive impairment types. In particular, the emotional aspects of prosody (i.e., expression of emotion through variations of different parameters of speech) provide a method for the speaker to utter a nuanced message that can be accurately perceived by a listener and may vary systematically with the expression of emotion (Pell et al., 2009). However, the available data on emotion expression in people with different types of cognitive impairment and neurotypical speakers are sparse.

The production of the prosodic features involves movement variations in all components of the speech production mechanism (Pell et al., 2009). Thus, changes in these acoustic measures may represent changes in the motor system associated with the neurologic changes associated with the different dementia types. In a review of cognitive, psychiatric, and motor symptoms of different dementia types, Magdy and Hussein (2022) reported that motor symptoms were significant indicators for Parkinson disease related dementias (e.g., corticobasal degeneration, dementia with Lewy bodies, and multiple system atrophy), normal pressure hydrocephalus, frontotemporal dementias and the posterior cortical atrophy variant of DAT. People with mild cognitive impairment (MCI) and DAT exhibit motor issues for complex tasks that can distinguish them from neurologically healthy controls (NHC) (Kluger et al., 1997). Although early-stage VaD and MCI can have similar cognitive symptoms, people with early-stage VaD do not tend to have motor symptoms (Kandasamy et al., 2020). The specific patterns of the motor issues relative to speech production for people with MCI and DAT have not been specifically described. Quite possibly, these motor issues may differ among the dementia types. Thus, the prosodic patterns for expressing emotion may provide a means to explore differences among cognitive impairment types.

Acoustic measurements that comprise prosody, such as fundamental frequency (f0), amplitude measured in dB level, and speech rate have been associated with the vocal expressions of emotions (Scherer, 2003), and several authors have reported evidence for emotion-specific patterns of acoustic cues (Banse and Scherer, 1996; Juslin and Laukka, 2003; Hamnmerschmidt and Jurgens, 2007). Mean f0 tends to be high (with a fast speech rate) for happiness, fear, and anger, and low for sadness (with a slow speech rate). F0 variability tends to be wide for happiness and anger but narrow for fear and sadness (Juslin and Laukka, 2003). Listeners exhibit approximately 60% accuracy for recognizing emotion from voice samples, although some emotions with more distinctive acoustic profiles (such as sadness and anger) may be easier for raters to identify than others (Johnstone and Scherer, 2000). However, this issue is complicated as the acoustic features of "emotional" prosody are not clear, given that there is no consensus on how acoustic features are manipulated to express different "emotions" (c.f., Bulut and Narayanan, 2008). For example, it is unclear how the frequency, amplitude, duration, and/or spectrum measures change when a person is in a state of emotional arousal, compared to when s/he is not (Patel et al., 2011). Without this discussion, the investigations into emotional prosody cannot be complete.

The present investigation, thus, was designed to provide preliminary evidence of unique prosodic production profiles of people with three types of cognitive impairment: DAT, VaD, and MCI (Study 1). Specifically, it was aimed to clarify how prosodic features differ acoustically across people with DAT, VaD, MCI, and healthy cognition and to determine whether the patterns of prosodic features can be used to differentially diagnose DAT, VaD, and MCI. One important concern of this study was whether these prosodic features could be associated with the expression of emotion. Accordingly, the categorization of perceived acoustic features into emotional versus non-emotional, or neutral, prosody was also carried out (Study 2). It was hypothesized that (1) the types of cognitive impairments will

TABLE 1 Speaker demographics.

Group	Mean years of age[a] (σ)	Sex (men, women)	Mean years of education (σ)	Mean MMSE[b] score (σ)
NHC	63.00 (9.24)	2, 8	14.9 (2.56)	29.3 (1.16)
DAT	69.36 (5.90)	4, 6	13.45 (3.47)	17.91 (5.54)
VaD	72.6 (6.12)	2, 3	11.2 (2.71)	15.4 (1.74)
MCI	63.11 (11.22)	5, 5	14.78 (3.42)	27.89 (1.45)

[a]Standard deviation.
[b]Mini mental state exam (score range: 0–30).

be associated with different prosodic features in comparison to neurotypical older adults and (2) unique patterns of emotion expression will be perceived for each group by neurotypical listeners. Overall, it was expected that the different prosodic features could lead to a useful tool for differential diagnosis of DAT, VaD, and MCI.

Methods

Procedures

Study 1 – Acoustic analysis of emotional prosody

Materials

For the first purpose, audio recordings of people with DAT, VaD, MCI, and NHC were obtained through DementiaBank,[1] a shared database supported by NIH-NIDCD grant R01-DC008524. The use of the secondary data was approved by Institutional Review Board at Ohio University (21-X-74). Included in this dataset were 10 people with DAT, 9 with MCI, 5 with VaD, and 10 NHC. On average, the speakers were 66.4 years old with 13.97 years of education at the time of original data collection. The one-way analysis of variance (ANOVA) revealed that across the four groups, the level of education ($F(3, 31) = 1.791$, $p = 0.169$) and age ($F(3, 31) = 2.094$, $p = 0.121$) of participants were not significantly different but the difference in Mini Mental State Exam (MMSE) scores were significant ($F(3, 31) = 34.761$, $p < 0.001$). Detailed characteristics of the speakers can be found in Table 1. Among the speech samples available to DementiaBank members, those describing the *Cookie Theft* picture in English from the *Pitt* (Becker et al., 1994) corpus were used based on previous research showing that the *Cookie Theft* picture description task, from the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001), provides a rich context in which mental state language and the cognitive processes associated with this language can be investigated (Cummings, 2019). It has been used to determine atypical emotional prosodic features of different clinical populations: Villain et al. (2016) found that stroke survivors described the picture using atypical emotional prosodic patterns, which is indicative of post-stroke depression. Wright et al. (2018) also reported atypical emotional prosody when describing the picture in right hemisphere stroke survivors and Patel et al. (2018) provided MRI images supporting the atypical prosodic patterns in this population. In individuals with dementia, Nevler et al. (2017) found that the Cookie Theft picture description task evoked emotional responses in people with behavioral variant frontotemporal dementia. Similarly, Haider et al. (2020) demonstrated that when using the *Cookie Theft* picture description task with a focus on emotional prosody, the accuracy of detecting Alzheimer's disease was 63.42%, which is comparable to when using the Berlin Database of Emotional Speech.

Acoustic analysis

The audio recordings and accompanying transcripts were downloaded and saved. The transcripts were compared to the audio files and amended as needed. Most amendments consisted of adding repetitions and filled pauses. The audio files were then parsed into utterances by the first and second authors of the current research independently, considering pauses and connectivity. After the independent work, the two researchers compared their evaluations and disagreements were resolved *via* discussions, until they reached 100% agreement. This parsing process resulted in a final outcome of 365 utterances including 108 utterances in the DAT, 75 in the MCI, 49 in the VaD, and 133 in the NHC groups. The utterances were then analyzed acoustically using the Praat software (Boersma and Weenink, 2017, v. 6.1.14) via a set of timing, pitch, and amplitude measures.

For timing, the following set of measurements was made for each utterance: the duration of the complete utterance including pauses and repetitions. This measure was recorded as the speech time. Then, the pauses longer than 200 ms and filled pauses, word repetitions, and syllable repetitions were removed from the utterances and the duration of the remaining signal was measured. This measure was recorded as the articulation time. In addition, the number of syllables in the utterance and the number of repeated syllables and repeated words were recorded. Finally, the duration of the removed pauses and duration of the repeated syllables and words were recorded. The speech time was divided by the total number of syllables, repeated syllables, and repeated words to determine the speech rate in syllables per second. The articulation time was divided by the number of syllables in the utterance to determine the articulation rate.

Many of these duration, timing, and extra syllable measures have indicated differences in expressed emotions. Comparisons between neutral and emotional speech have revealed that syllable and word repetitions decrease for emotional speech (Buchanan et al., 2014). Tao et al. (2018) reported that nonlinguistic fillers have no lexical information but contain emotional information. In addition, sad and fearful emotions are produced with more pauses, in comparison to neutral speech (Sauter et al., 2010). When rates have been calculated, they carry emotional valence as speaking rate differs among happiness, anger, sadness, fear, and neutral and articulation rate is slower for negative emotions (Petrushin, 1999; Erdemir et al., 2018; Tao et al., 2018).

After completing the utterance rate measures, the waveform of the articulation time for each utterance was displayed and the voiceless segments were removed using hand-controlled cursors to mark the voiceless segments. This version of the utterance was used for the pitch, loudness, and LTAS measures.

[1] https://dementia.talkbank.org

For pitch, the following set of f0 measurements were made for each utterance: the f0 of the first stable cycle of the first voiced sound and the f0 of the last stable cycle of the final voiced sound. In addition, the following measurements were collected using the output from the Voice Report from the Pulse menu in Praat: the highest f0 in the utterance, the lowest f0 in the utterance, and the median f0. The median f0 was used to reduce the effects of possible wide upward f0 shifts on the mean f0. The minimum f0 was subtracted from the maximum f0 to determine the range of f0 used (Δf0).

Similar to the duration and timing measures, the frequency measures have been used to differentiate among emotions and f0 measures considered to be primary indicators of emotional prosody (Bulut and Narayanan, 2008; Patel et al., 2011). Fear, joy, and anger are portrayed at a higher f0 than sadness and the f0 extent differs between happiness and fear (Bachorowski, 1999; Paeschke and Sendlmeier, 2000). The initial f0 differs between anger and sadness and the final f0 differs between happiness and sadness (Paeschke and Sendlmeier, 2000; Sauter et al., 2010). Finally, the average f0 differs between happiness and sadness (Paeschke and Sendlmeier, 2000).

For the loudness of the speech in dB (SPL), the following set of measurements was made for each utterance: the SPL of the first stable cycle of the first voiced sound and the SPL of the last stable cycle of the final voiced sound. In addition, the following measurements were collected using the output from the Intensity menu in Praat: the highest SPL in the utterance, the lowest SPL in the utterance, and the average SPL.

In comparison to the previous two sets of measures, measures of the relationship between SPL and emotion have been less explored. The average SPL differs between fear and sadness (Tao et al., 2006). In addition, the extent of SPL variations differs between anger and happiness (Tao et al., 2006). Since the SPL extent is determined from the maximum and minimum SPL levels, these measures may individually mark emotional differences. Similarly, the initial and final SPL levels may mark emotional differences.

Finally, three long-term average spectral (LTAS) measurements were made using the utterances without the voiceless segments: the LTAS slope, the LTAS offset, and the LTAS alpha ratio. These were extracted using standard bandwidth settings in the Praat LTAS routines. The LTAS measures indicate the pattern of amplitude by frequency. This interaction has indicated differences in emotional prosody as LTAS differences have been reported between sadness and anger and these measures mark the strength of emotional prosodic change (Tao et al., 2006; Cole et al., 2007).

Study 2 – Listener perception of emotional prosody

Study 2 was aimed at providing data to define "emotional" prosody to be used for differential diagnosis of cognitive impairments: when do listeners perceive emotion and what acoustic features are associated with the specific emotion? Neurotypical native English users were recruited to evaluate emotions expressed in each of the utterances per the approval of Institutional Review Board at Ohio University (21-X-61). The listeners were tested for their cognitive functioning using the Montreal Cognitive Assessment (MoCA; Nasreddine et al., 2005) and only those who scored above 26 (out of 30) were allowed to participate in the emotion evaluation.

For the emotion evaluation, a perception experiment consisting of practice, screening, and main sessions, was built online with Gorilla™.[2] The practice session was offered to anchor the listeners' evaluation using pseudo examples of seven emotions (i.e., happiness, sadness, disappointment, fear, surprise, anger, and neutral), developed and validated by Pell et al. (2009). During the practice trials, each listener was asked to choose the emotion of each utterance spoken by a professional actor or actress from seven choices including the 6 emotions mentioned above and a "neutral" option. The practice session consisted of 70 trials (with each of the six emotions and "neutral" appearing 10 times in random order) and feedback was provided following each response. After the practice session, each listener was asked whether s/he was confident to proceed to the screening test, which was shorter (10 trials) but followed the same format as the practice session. If the listener was not self-assured, another round of practice using a different set of utterances would be offered. A participant was considered passing the screening when s/he correctly identified at least 7 out of the 10 utterances. Failing the screening test would lead to an extra session of practice followed by a second screening test with a different set of utterances. Those who made two successive failures in the screening test would be excluded from participation. A total of 51 listeners participated in the screening test: 13 of them did not complete the screening and 28 of the 38 who completed the screening passed the screening at the pass rate of 73.6%. On average, the listeners were 29.6 years old ($\sigma = 11.62$) with 15.67 years of education ($\sigma = 1.75$) and earned 27.9 ($\sigma = 1.30$) on MoCA. Fourteen of them were men.

These 28 listeners, who successfully passed the screening, then moved on to the main test, where they were instructed to judge the emotions expressed in the *Cooke Theft* description utterances obtained from the DementiaBank. The listeners were informed that no feedback would be provided during the test. They were also instructed to make their best judgments based on their knowledge gained through the practice and screening sessions.

Statistical analysis

All statistical analyses were done using R version 4.1.0. The acoustic measures in Study 1 were analyzed using a principal component analysis (PCA) to determine the separate factors and grouping of the acoustic measures and a regression model to determine the acoustic measures representing a unique aspect of the variance across the cognitive impairment types. The criteria used as the probability to for entering additional terms to the model was set at less than or equal to $p = 0.05$. Finally, a Mahalanobis distance measure for multivariate ANOVA to determine how well the factors discriminated among the cognitive impairment types.

The utterances used in acoustic measures were then categorized into different emotions based on the perceptual evaluations by listeners in Study 2. Specifically, the counts for all emotions were obtained for each utterance, and that utterance was labelled as the emotion with the most counts. For example, if an utterance was perceived as "Angry" by 10 listeners and "Sad" by 3 listeners, that utterance would be labelled as "Angry." Utterances classified to the same emotion were then calculated for the descriptive statistics (i.e.,

[2] https://app.gorilla.sc/admin/home

mean and standard deviation) for each acoustic measure. A logistic regression model was constructed with the factor scores for each factor identified in the PCA for all utterances as independent variables and the emotion as the dependent variable. The emotion was coded into two classes: either neutral or emotional. The neutral class included utterances that were perceived as 'Neutral', and the emotional class included those identified as the rest 6 types of emotions.

Results

Study 1 – Acoustic analysis of emotional prosody

Factor analysis

The PCA was used to identify a small number of factors to represent relationships among sets of interrelated variables. The factor analysis of the acoustic measures revealed five factors with eigenvalues greater than 1.5. These factors and the included acoustic measures are depicted in Figure 1. The eigenvalues indicated the total variance explained by the correlated acoustic measures that comprise each factor. Two aspects of the data supported this stopping point for factors to include: first the variability accounted for dropped from 7 to 5.5% and second, the cumulative variability flattened after factor 5, as displayed in the scree plot in Figure 1. The five-factor model explained 67% of the total variance among the acoustic measures when separating the cognitive impairment types.

Acoustic measures were considered components of a factor when the factor loading was greater than 0.5 (Table 2). The first factor, labeled 'Mixed', was comprised of the following acoustic measures: the number of syllables for the speech and the articulation measures, the change in dB level, and the elapsed time for the speech and the articulation samples. The second factor was labeled 'Loudness' and included the initial, final, maximum, and minimum dB levels. The third factor was labeled 'Pitch' and was comprised of the final and maximum fundamental frequency levels as well as the difference in fundamental frequency level within each sample. The fourth factor is titled 'Rate' and included the speech and articulation rates. The fifth factor included the extra syllable count and extra syllable time and was labeled 'Syllable.' The extra syllables were repetitions and filled pauses. The difference between the acoustic measures in Factor 1 and Factor 4 was as follows: In Factor 1 the measures for speech and articulation are the number of syllables in the utterances and the elapsed time for each of those. In Factor 4, the acoustic measures are the division of the number of syllables by the elapsed time. It is noteworthy that these arithmetically related acoustic measures represented different aspects and proportions of the total variance of the differences among the four cognitive impairment groups.

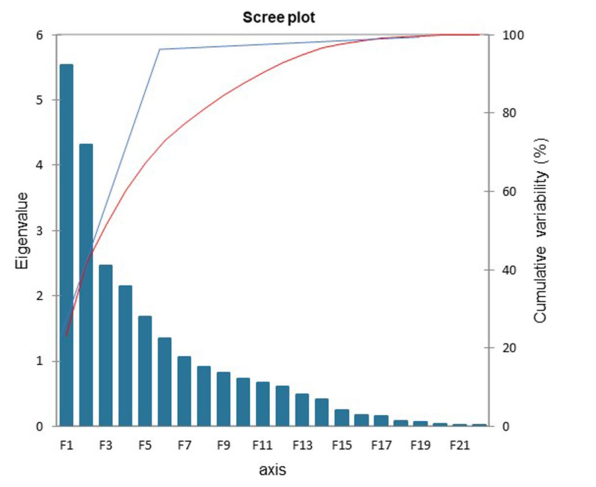


FIGURE 1
Scree plot: factors accounted for 67% of the overall variance.

Regression model

The stepwise fixed effects regression resulted in the inclusion of the following acoustic measures: the fundamental frequency at the end of the utterances, the articulation rates, the change in dB level during the utterance, and the sum of extra syllables in the utterances (Table 3). These acoustic measures were loaded onto separate factors in the factor analysis. The fixed effects regression model summary showed that all four measures were attributed to a significant amount of the total variance of the acoustic measures in relation to the cognitive impairment types. The negative Beta values for change in dB level within the utterance and articulation rate indicate that reductions in these two acoustic measures differentiated the cognitive impairment types. Finally, the tolerance information in the fourth model indicates that the variance explained by each of the acoustic measures was independent of the variance explained by the other acoustic measures included in the model.

Multivariate distance model

A set of Mahalanobis distance tests were completed. The Mahalanobis distance shows how far the test point is from the benchmark point. A Malahanobis distance of 1 or lower indicates that the test point is similar to the benchmark point. These measures indicate the distance between selected points in multivariate space. The Mahalanobis distance tests revealed that all of the factors exhibited relatively weak sensitivity; however, they exhibited good specificity. Although the discriminatory sensitivity was weak, the Mahalanobis distance factors had separate patterns across the cognitive impairment groups. The first, 'Mixed,' factor separated the DAT group from the other three factors ($F(21,1,020) = 4.259$, $p < 0.001$) with Mahalanobis distances that ranged from 0.423 to 1.06 which included the sum of extra syllables from the regression analysis. The second, 'Loudness,' factor distinguished the participants in the NHC and VaD groups from those in the DAT and MCI groups ($F(15,986) = 3.489$, $p < 0.001$) with Mahalanobis distances ranging from 0.231 to 0.773. 'Loudness' included the change in dB level during the utterance acoustic measurement from the regression analysis. The third, 'Pitch,' factor then separated the DAT and MCI groups from the NHC and VaD groups ($F(9,874) = 9.378$, $p < 0.001$) with Mahalanobis distances that ranged from 0.119 to 1.603 which included the final fundamental frequency measure from the regression analysis. The fourth, 'Rate,' factor included the articulation rate measure from the regression analysis which separated the DAT and NHC groups from the MCI and VaD groups ($F(6,720) = 4.579$, $p < 0.001$) with Mahalanobis distances of 0.008 to 0.438. The final, 'Syllable,' factor separated the NHC group from the three cognitive impairment groups ($F(6,720) = 5.366$, $p < 0.001$) with Mahalanobis distances that ranged from 0.109 to 0.491.

TABLE 2 Output of the principal component factor analysis including the correlation between the acoustic measures and the factors, the factor eigenvalues, and the percentage of variance explained by each factor.

Construct	Loadings	Factors				
		1	2	3	4	5
Mixed	Speech syllables	0.818				
	Speech time	0.837				
	Articulation syllables	0.802				
	Articulation time	0.849				
Loudness	dB change	**0.732**				
	dB initial		0.780			
	dB final		0.806			
	dB maximum		0.715			
	dB minimum		0.883			
Pitch	Frequency initial			0.559		
	Frequency final			**0.645**		
	Frequency maximum			0.792		
	Frequency change			0.708		
Rate	Speech rate				0.733	
	Articulation rate				**0.805**	
Syllable	Extra syllables					0.702
	Sum of extra syllables					**0.704**
	Eigenvalues	5.542	4.322	2.469	2.145	1.686
	Variance percentage	23.093	18.009	10.289	8.936	7.024

The acoustic measures included in the model created by the stepwise regression are in bold type.

TABLE 3 Results of stepwise regression including the four models, the R^2 explained, and the R^2 change for each model.

Variable	Beta (standard error)	*t*	*p*	95% CI [lower][a]	95% CI [upper]	Tolerance
Model 1 ($R^2 = 0.081$, R^2 change = 0.081)						
(Constant)		12.336	<0.001	1.287	1.776	
Frequency final	0.284 (0.001)	5.635	<0.001	0.003	0.005	1.000
Model 2 ($R^2 = 0.1451$, R^2 change = 0.064)						
(Constant)		11.365	<0.001	2.052	2.91	
Frequency final	0.277 (0.001)	5.691	<0.001	0.003	0.005	0.999
Articulation rate	−0.254 (0.041)	−5.204	<0.001	−0.294	−0.133	0.999
Model 3 ($R^2 = 0.163$, R^2 change = 0.017)						
(Constant)		10.51	<0.001	2.425	3.541	
Frequency final	0.279 (0.001)	5.772	<0.001	0.003	0.005	0.999
Articulation rate	−0.281 (0.042)	−5.694	<0.001	−0.318	−0.155	0.959
dB change	−0.135 (0.007)	−2.733	0.007	−0.034	−0.006	0.959
Model 4 ($R^2 = 0.174$, R^2 change = 0.012)						
(Constant)		10.771	<0.001	2.504	3.622	
Frequency final	0.284 (0.001)	5.898	<0.001	0.003	0.005	0.997
Articulation rate	−0.296 (0.042)	−5.981	<0.001	−0.331	−0.167	0.941
dB change	−0.164 (0.008)	−3.232	0.001	−0.039	−0.010	0.898
Sum of extra syllables	0.113 (0.139)	2.271	0.024	0.042	0.587	0.928

[a] Confidence interval.

The 'Syllable' factor did not incorporate any acoustic measures included in the model from the stepwise regression.

Study 2 – Listener perception of emotional prosody

Neurotypical listeners perceived neutral prosody in most of the utterances in all speaker groups. The NHC and MCI groups were most similar in terms of the composition of perceived emotions, while the highest number of angry utterances was identified in the DAT group and sad utterances in the VaD group. Table 4 shows the counts of responses and corresponding percentages, and Figure 2 presents the percentage of each perceived emotion.

To evaluate potential linguistic cues on listener perception of emotional prosody, the words used in the *Cookie Theft* picture description tasks were collected. Words without semantic valence such as *be* verbs and articles were excluded from the collection. As illustrated in Figure 3, the speakers across the 4 groups used similar words to describe the picture. In particular, the 10 most frequently used words constituted approximately 30.67% of NHC speech, 31.07% of MCI speech, 26.11% of DAT speech, and 37.04% of VaD speech, as described in Table 5. Given this finding, the impact of word choice on listener perception of emotion was deemed minimal.

The logistic regression revealed that Factor 3 of the PCA containing pitch measures as loadings was a significant predictor of emotional prosody. The odds of identifying emotional prosody increased by 22.6% (95% CI: [1.063, 1.418]) for using pitch measures compared to using other measures (i.e., mixed, loudness, rate, and syllable measures). Table 6 presents the outputs of the logistic regression in detail.

Discussion

An accurate diagnosis of cognitive impairment is critical to understand the person's condition, to establish care and treatment plans and to prepare for expected changes in different areas of daily living. However, the invasive nature and/or high cost of current diagnostic tools make it challenging for people experiencing cognitive impairment to get a precise diagnosis in a timely manner (López-de-Ipiña et al., 2015). Differential diagnosis is particularly important as it guides healthcare professionals and family caregivers in looking into key features and pathology of each type of dementia, so individuals living with the condition can receive the most appropriate treatments and support services that will in turn lead to the highest possible quality of life (Alzheimer's Association, n.d.). The current research was designed to address this issue by proposing a novel non-invasive and cost-efficient tool for differentiating cognitive impairment phenotypes. To achieve this goal, speech samples of people with different types of cognitive impairment (i.e., MCI, DAT, VaD) and healthy controls were analyzed acoustically for prosodic feature production (Study 1) and neurotypical listeners evaluated emotions conveyed by each utterance (Study 2).

The results of Study 1 demonstrated that acoustic features measured in this study can separate the cognitive impairment types. These features have been associated with emotional prosody (Patel et al., 2011; Pell et al., 2015). Five factors to separate cognitive impairment types were identified using the PCA and 4 of these factors were found salient for differentiating among cognitive impairment groups. Measures included in the 4 factors were the extent of dB changes, the fundamental frequency at the end of utterances, the number of extra syllables in the utterances, and the articulation rate. However, these factors and salient features provided a minimal separation among the cognitive impairment types. In Study 2, the neurotypical listeners perceived distinctive patterns in the utterances of the 4 groups. Although statistical differences were not calculated due to the imbalance of the number of utterances collected across the groups, NHC and MCI showed the most similar patterns. While listeners perceived a neutral prosodic pattern in the majority (>60%) of the utterances in NHC, MCI, and DAT, they indicated that approximately 40% of the utterances of the VaD group were neutral. Across NHC, MCI, and DAT groups, sad prosody consisted of 1 to 3% of all utterances. However, sad emotion was identified in approximately 22% of the VaD utterances. In addition, the listeners perceived that the VaD speakers expressed more utterances in fearful and surprised emotions compared to the other groups. These differences are noteworthy, despite the small number of VaD utterances.

Compiling the results of the two studies, frequency measures were found most critical for the listeners to perceive emotional prosody. This finding agrees with the results of some previous acoustic studies: Bulut and Narayanan (2008) found that the synthetic f0 modification to mean, range, and shape parameters affected the listener's perception of emotion embedded in the same utterance and Patel et al. (2018) demonstrated that voicing frequency affects the vocal expression of emotion. Although pitch was the strongest perceptual feature, amplitude and timing features also differentiated the four groups in the acoustic analyses. The manipulation of emotional prosody helps the speaker deliver the intention using non-linguistic clues and the listener interpret the intention accurately. This activity requires a wide range of interhemispheric cerebral networks, which is often impaired in people with cognitive impairment (Lian et al., 2018; Qi et al., 2019). The specific domains and severity of the impairment differ across the cognitive impairment groups and therefore, the analysis of emotional prosody can provide a low-cost and non-invasive tool to diagnose different types of cognitive impairment. Despite the strong potential of the analysis of emotional prosody, this line of study has been sparse and shown inconsistent findings. For example, some studies demonstrated that people with dementia struggle when attempting to express emotion (Horley et al., 2010; Haider et al., 2020) and the expression is completed in different ways than neurotypical speakers

TABLE 4 Listener perception of emotional prosody.

	NHC	MCI	DAT	VaD
	counts (%)	counts (%)	counts (%)	counts (%)
Neutral	86 (64.66)	51 (68.00)	67 (62.04)	15 (40.54)
Happy	12 (9.02)	6 (8.00)	11 (10.19)	2 (5.41)
Angry	10 (7.52)	4 (5.33)	14 (12.96)	3 (8.11)
Fearful	10 (7.52)	4 (5.33)	5 (4.63)	4 (10.81)
Surprised	6 (4.51)	3 (4.00)	4 (3.70)	4 (10.81)
Disappointed	5 (3.76)	3 (4.00)	5 (4.6)	1 (2.70)
Sad	4 (3.01)	4 (5.33)	2 (1.85)	8 (21.62)

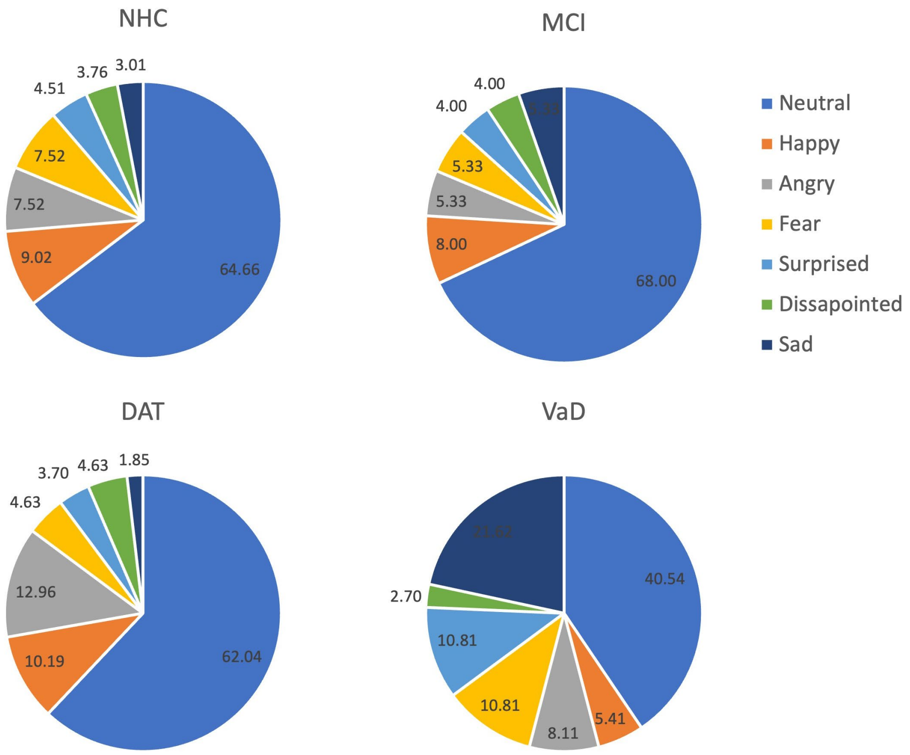


FIGURE 2
Listener perception (%) of each emotion per group.

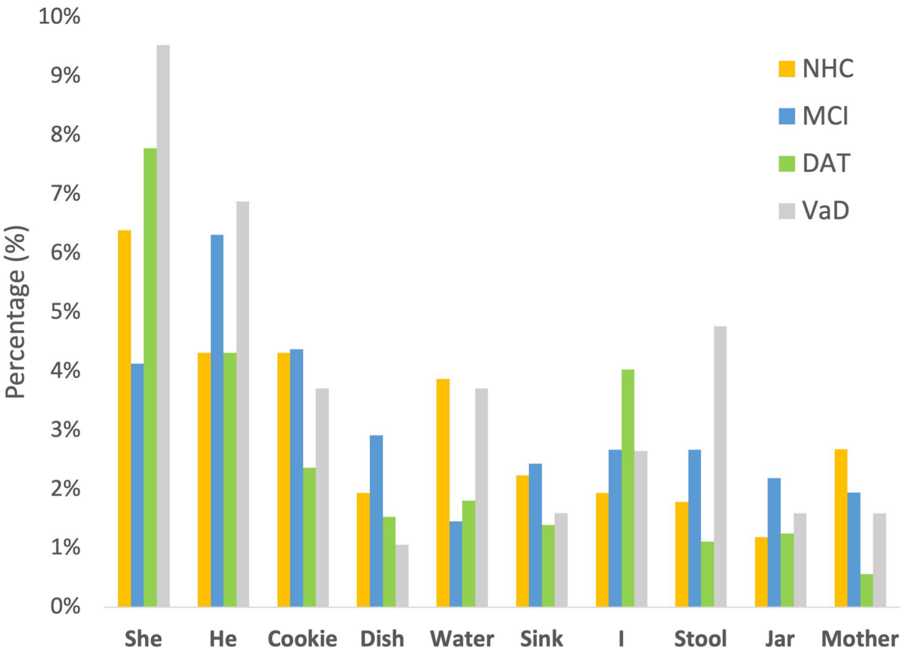


FIGURE 3
Ten common words used by the four groups of speakers.

do (e.g., Meilán et al., 2014; Nevler et al., 2017). Themistocleous et al. (2020) also found that aspects of voice quality and speech fluency of people with MCI and healthy controls differ significantly. Yang et al. (2021) showed correlations between speech features and brain atrophy among people with MCI and DAT and concluded that speech analysis may assist in MCI detection. Other researchers investigated prosody

TABLE 5 Words frequently used in the Cookie Theft picture description task.

	Group (%)			
Words	NHC	MCI	DAT	VaD
She	43 (6.39)	17 (4.13)	56 (7.78)	18 (9.52)
He	29 (4.31)	26 (6.31)	31 (4.31)	13 (6.88)
Cookie	29 (4.31)	18 (4.37)	17 (2.36)	7 (3.70)
Dish	13 (1.93)	12 (2.91)	11 (1.53)	2 (1.06)
Water	26 (3.83)	6 (1.46)	13 (1.81)	7 (3.70)
Sink	15 (2.23)	10 (2.43)	10 (1.39)	3 (1.59)
I	13 (1.93)	11 (2.67)	29 (4.03)	5 (2.65)
Stool	12 (1.78)	11 (2.67)	8 (1.11)	9 (4.76)
Jar	8 (1.19)	9 (2.18)	9 (1.25)	3 (1.59)
Mother	18 (2.67)	8 (1.94)	4 (0.56)	3 (1.59)

TABLE 6 Univariate logistic regression to differentiate emotional prosody from neutral prosody.

Variable	B	SE[a]	Z value	*p*	Exp (B)	95% CI lower	95% CI upper
Intercept	−0.505	0.112	−4.505	<0.001	0.603	0.483	0.750
Mixed	0.083	0.047	1.759	0.078	1.087	0.991	1.194
Loudness	0.024	0.054	0.447	0.655	1.025	0.922	1.141
Pitch	0.204	0.073	2.781	<0.01	1.226	1.063	1.418
Rate	0.057	0.076	0.749	0.454	1.059	0.911	1.231
Syllable	−0.109	0.086	−1.265	0.206	0.896	0.754	1.059

[a]Standard error.

production impairments in people with dementia and reported the potential of acoustic analysis of prosodic features as a dementia diagnostic tool (Kato et al., 2015, 2018; Martinc et al., 2021). However, other studies showed no differences in speech prosody between people with cognitive impairment and those who are healthy (e.g., Testa et al., 2001; Dara et al., 2013 – for spontaneous speech task only, Wright et al., 2018). Nevertheless, these studies either did not report the specific acoustic measures used or assessed a small set of acoustic measures. In addition, they did not clearly distinguish emotional prosody from linguistic prosody.

The present research provides several novel findings: First, it is the first to utilize a large set of acoustic measures that are specifically important for emotion expression to differentially diagnose 3 types of cognitive impairment. In particular, the current research is the first to include the VaD group. According to a recent systematic review (Oh et al., 2021), prosody and dementia studies included DAT and frontotemporal dementia groups only. Second, in this study, emotional prosody was clearly distinguished from linguistic prosody, supported by the neurotypical listeners' emotion evaluation. It is noteworthy that utterances of the NHC and MCI groups were perceived in a similar pattern while those of the VaD group were unique.

The current research has some limitations: First, the *Cookie Theft* picture description task may not be ideal to elicit emotional responses. Most of the utterances were perceived as neutral by the neurotypical listeners. Unlike the findings of previous studies showing the effectiveness of the *Cookie Theft* picture description task in evoking emotional responses (e.g., Villain et al., 2016; Nevler et al., 2017; Patel et al., 2018; Wright et al., 2018; Haider et al., 2020), the neurotypical listeners involved in this study as emotion raters identified neutral prosody in most of the speakers' utterances. This leads to the need to develop and validate a more appropriate procedure and/or stimuli. Second, a larger dataset including similar amount of data for each cognitive impairment and healthy group is warranted. Particularly, the listeners' perception was not statistically tested due to the different number of utterances collected for each group. Despite all the limitations, the findings of the research provide novel and functional implications that are clinically relevant. The findings demonstrate that the analysis of emotional prosody is a promising tool for differential diagnosis of cognitive impairment.

Data availability statement

The raw data supporting the conclusions of this article is available for verified members of DementiaBank (https://dementia.talkbank.org). Researchers and clinicians working with dementia who are interested in joining the consortium should read the Ground Rules and then send email to macw@cmu.edu with contact information and affiliation. Please include a brief general statement about how you envision using the data.

Ethics statement

The studies involving human participants were reviewed and approved by Ohio University Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

Author contributions

CO: conceptualization (lead), data curation (lead), formal analysis (supporting), funding acquisition (lead), investigation (lead), methodology (lead), project administration (lead), resources (lead), supervision (lead), visualization (lead), writing – original draft (lead), writing – review and editing (supporting). RM: conceptualization (supporting), data curation (supporting), formal analysis (equal), funding acquisition (supporting), investigation (supporting), methodology (supporting), writing – original draft (supporting), writing – review and editing (lead). XW: formal analysis (lead), funding acquisition (supporting), investigation (supporting), methodology (supporting), software (lead), writing – original draft (supporting), writing – review and editing (supporting). MR: investigation (supporting). All authors contributed to the article and approved the submitted version.

Funding

This research was supported by the Ohio University Research Committee grant.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alzheimer's Association. (n.d.). Differential Diagnosis. Available at: (https://www.alz.org/professionals/health-systems-clinicians/dementia-diagnosis/differential-diagnosis).

Bachorowski, J.-A. (1999). Vocal expression and perception of emotion. *Curr. Dir. Psychol. Sci.* 8, 53–57. doi: 10.1111/1467-8721.00013

Banse, R., and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* 70, 614–636. doi: 10.1037/0022-3514.70.3.614

Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., and McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Arch. Neurol.* 51, 585–594. doi: 10.1001/archneur.1994.00540180063015

Boersma, P., and Weenink, D. (2017). Praat: doing phonetics by computer [computer program]. Version 6.0.29, Available at: http://www.praat.org/ (Accessed May 24, 2017).

Bonifacio, G., and Zamboni, G. (2016). Brain imaging in dementia. *Postgrad. Med. J.* 92, 333–340. doi: 10.1136/postgradmedj-2015-133759

Buchanan, T. W., Laures-Gore, J. S., and Duff, M. C. (2014). Acute stress reduces speech fluency. *Biol Psych.* 97, 60–66. doi: 10.1016/j.biopsycho.2014.02.005

Bulut, M., and Narayanan, S. (2008). On the robustness of overall F0-only modifications to the perception of emotions in speech. *J. Acoust. Soc. Am.* 123, 4547–4558. doi: 10.1121/1.2909562

Calzá, L., Gagliardi, G., Favretti, R. R., and Tamburini, F. (2021). Linguistic features and automatic classfiers for identifying mild cognitive impairment and dementia. *Comput. Speech Lang.* 65:101113. doi: 10.1016/j.csl.2020.101113

Cheung, E. Y. W., and Mak, H. (2020). Association between Interhemispheric Functional Connectivity (IFC) and amyloid deposition in patients with different types of dementia. *Alzheimer's Dement.* 16.

Cole, J., Kim, H., Choi, H., and Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: evidence from radio news speech. *J. Phon.* 35, 180–209. doi: 10.1016/j.wocn.2006.03.004

Cummings, L. (2019). Describing the cookie theft picture. Sources of breakdown in Alzheimer's dementia. *Pragmat. Soc.* 10, 153–176. doi: 10.1075/ps.17011.cum

Dara, C., Kirsch-Darrow, L., Ochfeld, E., Slenz, J., Agranovich, A., Vasconcellos-Faria, A., et al. (2013). Impaired emotion processing from vocal and facial cues in frontotemporal dementia compared to right hemisphere stroke. *Neurocase* 19, 521–529. doi: 10.1080/13554794.2012.701641

Erdemir, A., Walden, T. A., Jefferson, C. M., Choi, D., and Jones, R. M. (2018). The effect of emotion on articulation rate in persistence and recovery of childhood stuttering. *J. Fluen. Disord.* 56, 1–17. doi: 10.1016/j.jfludis.2017.11.003

Fagherazzi, G., Fischer, A., Ismael, M., and Despotovic, V. (2021). Voice for health: the use of vocal biomarkers from research to clinical practice. *Digit. Biomark.* 5, 78–88. doi: 10.1159/000515346

Filippi, M., Agosta, F., Barkhof, F., et al. (2012). Efns Task Force: The use of neuroimaging in the diagnosis of dementia. *Eur. J. Neurol.* 19, 1487–1501. doi: 10.1111/j.1468-1331.2012.03859.x

Goodglass, H., Kaplan, E., and Barresi, B. (2001). *Boston diagnostic aphasia examination–third edition (BDAE-3)*. Philadelphia, PA: Lippincott Williams and Wilkins.

Haider, F., de la Fuente, S., and Luz, S. (2020). An assessment of paralinguistic acoustic features for detection of Alzheimer's dementia in spontaneous speech. *IEEE J. Sel. Top. Signal Process.* 14, 272–281. doi: 10.1109/jstsp.2019.2955022

Hamnmerschmidt, K., and Jurgens, U. (2007). Acoustical correlates of affective prosody. *J. Voice* 21, 531–540. doi: 10.1016/j.jvoice.2006.03.002

Horley, K., Reid, A., and Burnham, D. (2010). Emotional prosody perception and production in dementia of the Alzheimer's type. *J. Speech Lang. Hear. Res.* 53, 1132–1146. doi: 10.1044/1092-4388(2010/09-0030)

Hupp, J. M., and Junger, M. K. (2013). Beyond words: comprehension and production of pragmatic prosody in adults and children. *J. Exp. Child Psychol.* 115, 536–551. doi: 10.1016/j.jecp.2012.12.012

Johnstone, T., and Scherer, K. R. (2000). "Vocal communication of emotion" in *Handbook of emotions*. eds. M. Lewis and M. J. Haviland-Jones. *2nd* ed (New York, NY: Guilford), 220–235.

Juslin, P. N., and Laukka, P. (2003). Communication of emotions in vocal expression and music performance: different channels, same code? *Psychol. Bull.* 129, 770–814. doi: 10.1037/0033-2909.129.5.770

Kandasamy, M., Anusuyadevi, M., Aigner, K. M., Unger, M. S., Kniewallner, K. M., Bessa De Sousa, D. M., et al. (2020). TGF-β signaling: a therapeutic target to reinstate regenerative plasticity in vascular dementia? *Aging Dis.* 11, 828–850. doi: 10.14336/AD.2020.0222

Kato, S., Homma, A., and Sakuma, T. (2018). Easy screening for mild Alzheimer's disease and mild cognitive impairment from elderly speech. *Curr. Alzheimer Res.* 15, 104–110. doi: 10.2174/1567205014666171120144343

Kato, S., Homma, A., Sakuma, T., and Nakamura, M. (2015). Detection of mild Alzheimer's disease and mild cognitive impairment from elderly speech: binary discrimination using logistic regression. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2015, 5569–5572. doi: 10.1109/EMBC.2015.7319654

Kluger, A., Gianutsos, J. G., Golomb, J., Ferris, S. H., George, A. E., Franssen, E., et al. (1997). Patterns of motor impairment in normal aging, mild cognitive decline, and early Alzheimer's disease. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 52B, P28–P39. doi: 10.1093/geronb/52B.1.P28

Lian, C., Liu, M., Zhang, J., and Shen, D. (2018). Hierarchical convolutional network for joint atrophy localization and Alzheimer's disease diagnosis using structural MRI. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 880–893. doi: 10.1109/TPAMI.2018.2889096

López-de-Ipiña, K., Alonso, J. B., Solé-Casals, J., Barroso, N., Henriquez, P., Faundez-Zanuy, M., et al. (2015). On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. *Cognit. Comput.* 7, 44–55. doi: 10.1007/s12559-013-9229-9

Lundholm Fors, K., Fraser, K., and Kokkinakis, D. (2018). Automated syntactic analysis of language abilities in persons with mild and subjective cognitive impairment. *Stud. Health Technol. Inform.* 247:705–709.

Maclin, J. M., Wang, T., and Xiao, S. (2019). Biomarkers for the diagnosis of Alzheimer's disease, dementia Lewy body, frontotemporal dementia and vascular dementia. General. *Psychiatry.* 32:e10054. doi: 10.1136/gpsych-2019-100054

Magdy, R., and Hussein, L. (2022). Cognitive, psychiatric, and motor symptoms–based algorithmic approach to differentiate among various types of dementia syndromes. *J. Nerv. Ment. Dis.* 210, 129–135. doi: 10.1097/NMD.0000000000001428

Martinc, M., Haider, F., Pollak, S., and Luz, S. (2021). Temporal integration of text transcripts and acoustic features for Alzheimer's diagnosis based on spontaneous speech. *Front. Aging Neurosci.* 13:642647. doi: 10.3389/fnagi.2021.642647

Meilán, J. J. G., Martínez-Sánchez, F., Carro, J., López, D. E., Millian-Morell, L., and Arana, J. M. (2014). Speech in Alzheimer's disease: can temporal and acoustic parameters discriminate dementia? *Dement. Geriatr. Cogn. Disord.* 37, 327–334. doi: 10.1159/000356726

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal cognitive assessment, MoCA: a brief screening tool for mild cognitive impairment. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Nevler, N., Ash, S., Jester, C., Irwin, D. J., Liberman, M., and Grossman, M. (2017). Automatic measurement of prosody in behavioral variant FTD. *Neurology* 89, 650–656. doi: 10.1212/WNL.0000000000004236

Odell, K., McNeil, M. R., Rosenbek, J. C., and Hunter, L. (1991). Perceptual characteristics of vowel and prosody production in apraxic, aphasic, and dysarthric speakers. *J. Speech Hear. Res.* 34, 67–80. doi: 10.1044/jshr.3401.67

Oh, C., Morris, R. J., and Wang, S. (2021). A systematic review of expressive and receptive prosody in people with dementia. *J. Speech Lang. Hear. Res.* 64, 3803–3825. doi: 10.1044/2021_JSLHR-21-00013

Paeschke, A., and Sendlmeier, W. (2000). Prosodic characteristics of emotional speech: measurements of fundamental frequency movements. ISCA Archive. 75–80.

Patel, S., Oishi, K., Wright, A., Sutherland-Foggio, H., Saxena, S., Sheppard, S. M., et al. (2018). Right hemisphere regions critical for expression of emotion through prosody. *Front. Neurol.* 9:224. doi: 10.3389/fneur.2018.00224

Patel, S., Scherer, K. R., Björkner, E., and Sundberg, J. (2011). Mapping emotions into acoustic space: the role of voice production. *Biol. Psychol.* 87, 93–98. doi: 10.1016/j.biopsycho.2011.02.010

Pell, M. D., Paulmann, S., Dara, C., Alasseri, A., and Kotz, S. A. (2009). Factors in the recognition of vocally expressed emotions: a comparison of four languages. *J. Phon.* 37, 417–435. doi: 10.1016/j.wocn.2009.07.005

Pell, M.D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., and Rigoulot, S. (2015) Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biol. Psychol.* 111, 14–25. doi: 10.1016/j.biopsycho.2015.08.008

Petrushin, V. (1999). Emotion in speech: recognition and application to call centers. *Proc. Artificial Neural Netw. Eng.* 1, 7–10.

Qi, Z., An, Y., Zhang, M., Li, H., and Lu, J. (2019). Altered cerebro-cerebellar limbic network in AD spectrum: a resting-state fMRI study. *Front. Neural Circuits* 13:72. doi: 10.3389/fncir.2019.00072

Sauter, D. A., Eisner, F., Calder, A. J., and Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Q. J. Exp. Psychol.* 63, 2251–2272. doi: 10.1080/17470211003721642

Scherer, K. R. (2003). Vocal communication of emotion: a review of research paradigms. *Speech Commun.* 40, 227–256. doi: 10.1016/S0167-6393(02)00084-5

Tao, J., Kang, Y., and Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE Trans. Audio Speech Lang. Process.* 14, 1145–1154. doi: 10.1109/tasl.2006.876113

Tao, F., Liu, G., and Zhao, Q. (2018). An ensemble framework of voice-based Emotion Recognition System. *First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).*:2018. doi: 10.1109/aciiasia.2018.8470328

Testa, J. A., Beatty, W. W., Gleason, A. C., Orbelo, D. M., and Ross, E. D. (2001). Impaired affective prosody in AD: relationship to aphasic deficits and emotional behaviors. *Neurology* 57, 1474–1481. doi: 10.1212/WNL.57.8.1474

Themistocleous, C., Eckerström, M., and Kokkinakis, D. (2020). Voice quality and speech fluency distinguishing individuals with mild cognitive impairment from healthy controls. *PLoS One* 15:e0236009. doi: 10.1371/journal.pone.0236009

Villain, M., Cosin, C., Glize, B., Berthoz, S., Swendsen, J., Sibon, I., et al. (2016). Affective prosody and depression after stroke. *A pilot study. Stroke* 47, 2397–2400. doi: 10.1161/STROKEAHA.116.013852

Wright, A., Saxena, S., Sheppard, S. M., and Hillis, A. E. (2018). Selective impairments in components of affective prosody in neurologically impaired individuals. *Brain and Cognition.* 124, 29–36. doi: 10.1016/j.bandc.2018.04.001

Yang, Q., Xu, F., Ling, Z., Li, Y., and Fang, D. (2021). Selecting and analyzing speech features for the screening of mild cognitive impairment. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2021, 1906–1910. doi: 10.1109/EMBC46164.2021.9630752

TYPE Original Research
PUBLISHED 25 July 2023
DOI 10.3389/fnins.2023.1135850

Check for updates

OPEN ACCESS

EDITED BY
Si Chen,
Hong Kong Polytechnic University, Hong Kong SAR, China

REVIEWED BY
Yixin Zhang,
Hong Kong Polytechnic University, Hong Kong SAR, China
Guijun Chen,
Taiyuan University of Technology, China
Redha Taiar,
Université de Reims Champagne-Ardenne, France

*CORRESPONDENCE
Hao Chao
✉ chaohao1981@163.com

RECEIVED 02 January 2023
ACCEPTED 26 June 2023
PUBLISHED 25 July 2023

CITATION
Chao H, Cao Y and Liu Y (2023) Multi-channel EEG emotion recognition through residual graph attention neural network.
Front. Neurosci. 17:1135850.
doi: 10.3389/fnins.2023.1135850

Multi-channel EEG emotion recognition through residual graph attention neural network

Hao Chao*, Yiming Cao and Yongli Liu

College of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, China

In this paper, a novel EEG emotion recognition method based on residual graph attention neural network is proposed. The method constructs a three-dimensional sparse feature matrix according to the relative position of electrode channels, and inputs it into the residual network to extract high-level abstract features containing electrode spatial position information. At the same time, the adjacency matrix representing the connection relationship of electrode channels is constructed, and the time-domain features of multi-channel EEG are modeled using graph. Then, the graph attention neural network is utilized to learn the intrinsic connection relationship between EEG channels located in different brain regions from the adjacency matrix and the constructed graph structure data. Finally, the high-level abstract features extracted from the two networks are fused to judge the emotional state. The experiment is carried out on DEAP data set. The experimental results show that the spatial domain information of electrode channels and the intrinsic connection relationship between different channels contain salient information related to emotional state, and the proposed model can effectively fuse these information to improve the performance of multi-channel EEG emotion recognition.

KEYWORDS

EEG, emotion recognition, residual network, graph attention neural network, feature fusion

1 Introduction

Emotion is a physiological state of human beings accompanied by cognition and consciousness. People's daily cognitive and behavioral activities are almost driven by emotion, which also affects interpersonal interaction and group activities (Guozhen et al., 2016). Affective computing is a representative field, which aims to give computer systems the ability to automatically recognize, understand and respond to human emotions, so as to realize intelligent human-computer interaction. As the core and important component of affective computing, emotion recognition has a wide range of applications in psychology, emotional computing, artificial intelligence, computer vision, medical, and other fields (Ramirez et al., 2001; Hu et al., 2019; Fürbass et al., 2020).

Physiological signals mainly include electrocardiogram (ECG), electromyography (EMG), and electroencephalogram (EEG). Compared with facial expressions and voice signals, physiological signals are not easy to disguise, and are more objective and reliable in capturing the real emotional state of human beings. With the rapid development of wearable devices, long-term monitoring of physiological signals has become a reality, which makes it feasible and practical to judge emotional state based on EEG signals. In the medical field, EEG classification models play a role in automatic diagnosis of psychiatric disorders. Depression is one of the largest health problems in the world. It is a serious mental illness, and there is a problem of untimely treatment. Severe patients often have thoughts of suicide.

Current diagnostic criteria for depression are still based on subjective clinical rating scales, such as The Hamilton Depression Rating Scale (Hamilton, 1960), and require physician input (Sung et al., 2005). Some research focuses on automatic diagnosis of depression based on EEG (Alhaj et al., 2011; Mohammadi et al., 2015), which can enable patients to quickly diagnose and understand their own condition, so as to carry out scientific treatment in advance.

The main components of the EEG signal are brain rhythm from different brain regions, reflecting the activity of that region (Niedermeyer and da Silva, 2005). The electrical activity of the cerebral cortex is transmitted to the scalp through the anatomical structure. Therefore, the acquired EEG is a mixture of source signals from different brain regions, carrying a large amount of spatial location information (Xing et al., 2019). In the research field of emotion recognition based on EEG, some studies have explored asymmetric features of brain regions, such as DASM (differential asymmetry), RASM (rational asymmetry), DCAU (differential causality) (Gogna et al., 2016; Li et al., 2018b). And other works studied the connectivity of EEG signals (Nolte et al., 2004, 2008; Supp et al., 2007; Haufe et al., 2013). Castelnovo et al. finds that the electrical activity of the brain is mainly concentrated in specific brain regions when people are in different sleep states, scalp EEG analysis of all night NREM (non-rapid eye movement) sleep revealed a localized decrease in slow wave activity (SWA) power (1–4 Hz) over centro-parietal regions relative to the rest of the brain in SADs compared to good sleeping healthy controls (Castelnovo et al., 2016). Nowadays, there are also some works that make better use of the spatial domain information of EEG channels in EEG classification tasks. In order to learn the spatiotemporal characteristics of EEG signals, Salama et al. divided the original EEG signals into multiple frames, and combined the original EEG signals of multiple channels into a two-dimensional matrix in each frame, where the first dimension represents the number of channels, and the second dimension Indicates the time length of a frame. Multiple frames are then superimposed to form a three-dimensional matrix, with the third dimension representing time. Finally, the 3D matrix is used as the input for 3D-CNN (3d convolutional neural networks) training. Since the left and right hemispheres of the human brain respond asymmetrically to emotion, a bi-hemisphere domain adversarial neural network (BiDANN) model is proposed to learn the discriminative emotional features of each hemisphere, BiDANN contains one global and two local domain discriminators, and learns discriminative sentiment features for each hemisphere by adversarial with local domain discriminators and classifiers (Li et al., 2018c). Li et al. (2017) captures the spatial domain information contained in electrode positions by mapping into EEG multidimensional feature image following a 10/20 system. First, the spatial features, frequency domain and time features of the EEG signal are integrated, and mapped into a feature matrix according to the international 10/20 system, and then the EEG multidimensional feature image is generated using the interpolation method, using a combination of convolutional neural network (CNN) and long-term and short-term A hybrid deep network of memory (LSTM) recurrent neural network (RNN) recognizes emotional states. Li et al. (2018a) also used the distribution of electrodes on the scalp to extract the spatial domain information of electrode locations. First, the differential entropy features from 62 EEG signal channels are organized into a two-dimensional map of 8×9, and are mapped to a 20×20 input map through sparse operations to avoid information leakage in convolution and pooling operations. Finally, hierarchical convolutional neural network (HCNN) is used to classify positive, neutral and negative emotional states.

To a certain extent, the above research has applied the extraction of the spatial domain information of the EEG channel, and used the multi-dimensional feature matrix mapped according to the international 10/20 system and CNN to fuse the information of the neighbor nodes. However, there still exist several challenges in multi-channel EEG-based emotion recognition. First of all, the brain activity in emotional state is complex, and multiple brain regions are involved in the interaction. How to effectively characterize the interaction between brain regions is a problem to be considered. Furthermore, due to the local perception characteristics, CNN (Convolutional Neural Networks) tends to pay more attention to adjacent electrode channels and is good at learning local spatial patterns. Therefore, in the process of extracting electrode spatial position information, CNN can mine the significant information of correlation and interaction of different EEG signals in the same brain region. However, it cannot effectively capture the intrinsic connection relationship between EEG channels located in different brain regions and the global spatial position information of electrodes. Finally, the features extracted from the EEG signal and the distance between different electrodes are a kind of non-Euclidean data, only mapping the features extracted from each channel into a multi-dimensional sparse feature matrix according to the international 10/20 system ignores the distance information between electrodes, and ignores that all electrodes are not positioned in an absolute plane on the scalp.

To solve the above problems, this paper proposes a noval emotion recognition method based on residual graph attention neural network (ResGAT). In the proposed method, the residual network is utilized to achieve the spatial position information of the electrode channel and the correlation information of the adjacent EEG channels through the 3D feature matrix. Considering that the graph neural network (GAT) can update the state of vertices by periodically exchanging neighborhood information without being limited by vertex distance, it is employed to learn the neural functional connections between different brain regions, and the multi-head self-attention mechanism is used to adaptively adjust the adjacency matrix in the network. Therefore, the ResGAT model makes full use of the electrode spatial position information and the intrinsic connection relationship between EEG channels located in different brain regions. Moreover, when the EEG channel aggregates the characteristics of neighboring nodes, it pays more attention to the channel that is more relevant to itself. Finally, the high-level abstract features representing electrode space domain information and the high-level abstract features representing intrinsic connection relationship between EEG channels located in different brain regions are fused to judge the emotional state.

2 Datasets and feature extraction

2.1 Data set

The DEAP data set used in the experiment is an open data set collected through experiments by Koelstra et al. from Queen Mary University of London, University of Twente, University of Geneva, Switzerland, and Swiss federal Institute of Technology in Lausanne to analyze human emotional states (Koelstra et al., 2011). The dataset records multimodal physiological signals of 32 volunteers under the stimulation of selected music videos, including EEG and peripheral physiological signals, and 22 of the 32 volunteers also record facial expression videos. Each volunteer needs to watch 40 1-min long videos using 32 active AgCl electrodes (Fp1, AF3, F3, F7, FC5, FC1, C3, T7, CP5, CP1, P3, P7, PO3, O1, Oz, Pz, Fp2, AF4, Fz, F4, F8, FC6, FC2, Cz, C4, T8, CP6, CP2, P4, P8, PO4, and O2) recording EEG signals, these electrodes were placed on the scalp according to the international 10/20 system. At the end of each trial, the valence, arousal, dominance, and liking of the video were evaluated on a scale of 0–9. Physiological signals were sampled at 512 Hz and resampled at 128 Hz. The physiological signal matrix of each subject is 40×40×8064 (40 trials, 40 channels, 8,064 sampling points). Eighty thousand and sixty-four is 63 s data at 128 Hz sampling rate, 3 s silent time.

In addition, the effectiveness of ResGAT was verified using the SEED-IV brain emotion dataset (Zheng et al., 2018). This data set selected 72 movie clips containing 4 emotions (Happy, Sad, Neutral, and Fear) as EEG-induced materials. A total of 15 subjects recorded 62-channel EEG signals and eye movements when watching movie clips.

2.2 Feature extraction

In the DEAP set, each person watched 40 emotion-inducing videos, and the duration of EEG signals recorded in each video was 60 s. In the experiment, a sliding window divides the raw EEG signal of each channel into several segments, and the duration of each sliding window is set to 6 s. The segments do not over lap. Each segment is considered an independent sample, and the six new samples inherited the labels of the original. Thus, 12,800 samples were be obtained. A set of time domain features can be extracted from the 32-channel EEG signals of each sample, specifically including mean, variance, first difference value, second difference value, standard deviation, and fuzzy entropy. Among them, Fuzzy Entropy was proposed by Chen et al. (2007) and applied to the representation of EMG signals. Fuzzy Entropy introduces the concept of fuzzy sets. Based on the exponential function and its shape, the similarity of vectors is vaguely defined in FuzzyEn, compared with ApEn and SampEn, the FuzzyEn is an effective measure algorithm for analyzing chaotic sequence complexity, it has better robustness and measure value continuities.

The soft continuous boundary of the fuzzy function ensures the continuity and effectiveness of the fuzzy entropy under small parameters, so the more details obtained by the fuzzy function also make the fuzzy entropy a more accurate definition of entropy. Assuming that the EEG signal of each channel is represented by s(T), t = 1, 2..., T, T is the signal length, which is 128×60 (frequency×second), the measure of the length of the EEG signal subsequence in fuzzy entropy m is 2. By reconstructing the original sequence, we can get

$$X_i^m = \left\{ s(i), s(i+1), ..., s(i+m-1) \right\} - s_0(i) \tag{1}$$

Among them, i=1,2,...,N-m+1. X_i^m represents m consecutive s values, $s_0(i)$ represents the average value, calculated as follows,

$$s_0(i) = \frac{\sum_{j=0}^{m-1} s(i+j)}{m} \tag{2}$$

Define the maximum difference d_{ij}^m between elements in two m-dimensional vectors X_i^m and X_j^m as the distance between them,

$$d_{ij}^m = \underset{k\in(0,m-1)}{max} \left\{ | \; s(i+j) - s_0(j) - (s(i+k) - s_0(i)) \; | \right\} \tag{3}$$

The similarity between X_i^m and X_j^m can be defined by a fuzzy function,

$$D_{ij}^m = \mu(d_{ij}^m, r) \tag{4}$$

Structure $\varphi_m(r)$ and $\varphi_{m+1}(r)$,

$$\varphi_m(r) = (N-m)_{-1} \sum_{i=1}^{N-m} \phi_i^m(r) \tag{5}$$

$$\varphi_{m+1}(r) = (N-m)_{-1} \sum_{i=1}^{N-m} \phi_i^{m+1}(r) \tag{6}$$

Then can define the parameter $FuzzyEn(m, r)$ of the time series as:

$$FuzzyEn(m, r) = \lim_{x\to -\infty} [\ln \varphi^m(r) - \ln \varphi^{m+1}(r)] \tag{7}$$

Among them, when N is finite, it can be estimated by statistics,

$$FuzzyEn(m, r, N) = \ln \varphi^m(r) - \ln \varphi^{m+1}(r) \tag{8}$$

Two emotion dimensions are used in the experiments. For each sample, if the self-assessment value of the arousal is greater than 5, the category label of the sample is set as the high arousal (HA), otherwise it is set as the low arousal (LA). In the valence-sentiment dimension, the same label division is used for samples, including the high valence (HV) and the low valence (LV).

3 ResGAT emotion recognition framework

The structure of the proposed ResGAT model is described in Figure 1. The framework includes feature extraction and feature mapping module, ResNet modules, GAT modules and Classification modules. The first part is feature extraction and feature mapping modules, which extracts 6 kinds of temporal features from 32 EEG signals. Then, a 2D electrode position mapping matrix and a 3D sparse feature matrixand are constructed according to the temporal features. The 2D electrode position mapping matrix is input into the GAT modules to extract high-level abstract features, which contain the intrinsic connection

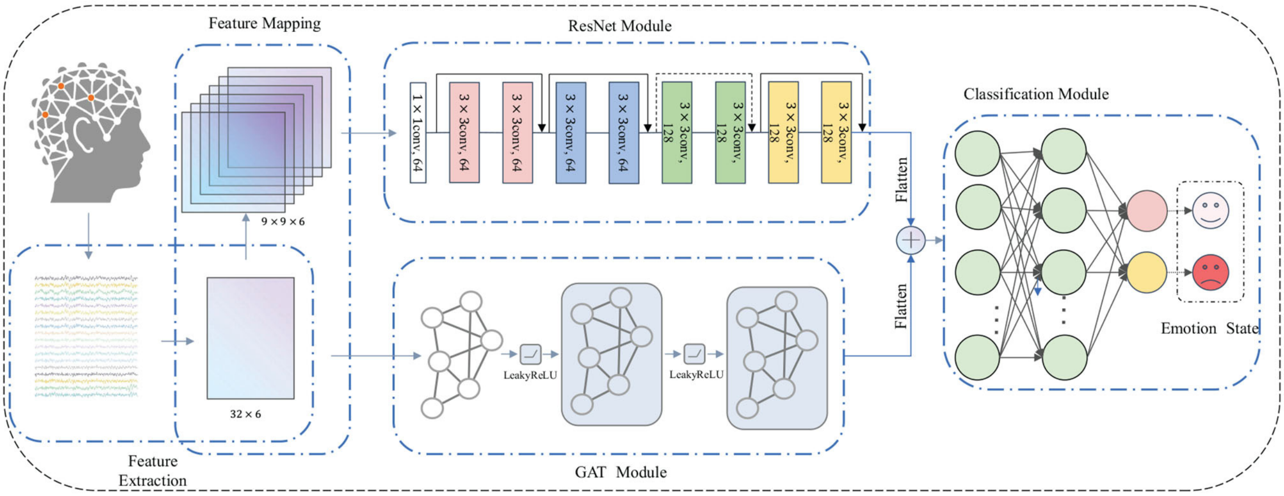
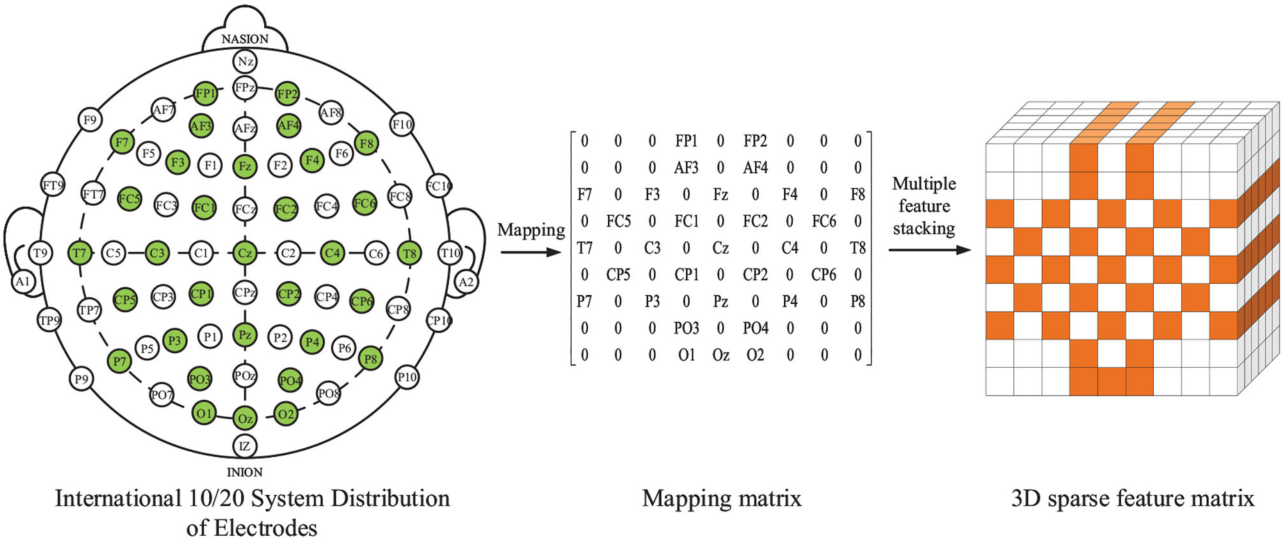


FIGURE 1
ResGAT emotion recognition framework.



$$\begin{bmatrix} 0 & 0 & 0 & \mathrm{FP1} & 0 & \mathrm{FP2} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{AF3} & 0 & \mathrm{AF4} & 0 & 0 & 0 \\ \mathrm{F7} & 0 & \mathrm{F3} & 0 & \mathrm{Fz} & 0 & \mathrm{F4} & 0 & \mathrm{F8} \\ 0 & \mathrm{FC5} & 0 & \mathrm{FC1} & 0 & \mathrm{FC2} & 0 & \mathrm{FC6} & 0 \\ \mathrm{T7} & 0 & \mathrm{C3} & 0 & \mathrm{Cz} & 0 & \mathrm{C4} & 0 & \mathrm{T8} \\ 0 & \mathrm{CP5} & 0 & \mathrm{CP1} & 0 & \mathrm{CP2} & 0 & \mathrm{CP6} & 0 \\ \mathrm{P7} & 0 & \mathrm{P3} & 0 & \mathrm{Pz} & 0 & \mathrm{P4} & 0 & \mathrm{P8} \\ 0 & 0 & 0 & \mathrm{PO3} & 0 & \mathrm{PO4} & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{O1} & \mathrm{Oz} & \mathrm{O2} & 0 & 0 & 0 \end{bmatrix}$$

FIGURE 2
International 10/20 system, 9×9 mapping matrix and 3D sparse feature matrix.

relationships between EEG channels located in different brain regions. The ResNet modules is employed to receive the 3D sparse feature matrix and generate high-level abstract features representing electrode spatial position information. Finally, the classification modules is utilized to fuse the two high-level abstract features and judge the emotional state.

3.1 Extraction of spatial domain information based on ResNet Module

Figure 2 shows the international 10/20 system plan, 2D electrode position mapping matrix and 3D sparse feature matrix $X^R \in \mathbb{R}^{h \times w \times c}$. The values of parameter h and parameter w are both set to 9, and the value of parameter c is 6, indicating that the shape of the 3D feature matrix is. The left side of Figure 2 shows the International 10/20 system, where the EEG electrodes marked by green circles are the test points used in the DEAP dataset. Some researches (Li et al., 2017; Chao and Dong, 2020; Cui et al., 2020) have found that spatial features of EEG channels can improve the performance of emotion recognition. In order to represent the spatial location information of all EEG signal channels, a feature matrix is constructed according to the positions of electrodes on the brain, and the spatial parts of different EEG signal channels are mined. In the feature matrix, the time-domain features extracted from different EEG channels are put into the corresponding positions in the matrix by name, and the positions of unused electrodes in the matrix are set to 0. Finally, a 9×9×6 three-dimensional feature matrix is constructed according to the six extracted features, as shown in the right of Figure 2. For each 9×9×1 matrix, different time domains are arranged according to the mapping rules shown in Figure 2 mapping matrix. Finally, the extracted 3D sparse feature matrix is represented by X^R, $h = w = 9$, $c = 6$, indicating that the shape of the 3D feature matrix is 9×9×6.

After constructing the three-dimensional sparse mapping matrix, it is input into the residual network to extract high-level abstract features. The residual network structure adopted is shown in Figure 1. It is composed of multiple residual blocks. Each residual block is composed of multiple convolution layers, batch normalization layers and activation layers. The size of the convolution kernels used in this part is 3×3. The residual block is calculated as follows:

First, the 3D sparse feature map $U \in \mathbb{R}^{h' \times w' \times c'}$ is obtained from $X^R \in \mathbb{R}^{h \times w \times c}$ by transforming F_{tr}. For transform F_{tr}, it is a Convolution operation. Use $V = [v_1, v_2, ..., v_{c'}]$ to represent the filter set, where v_i refers to the parameter of the i_{th} filter. The output is $U = [u_1, u_2, ..., u_{c'}]$, and

$$u_i = v_i * X^R \tag{9}$$

Here, * means convolution, the filter can learn the spatial position information of electrodes and the interaction information between electrodes in local spatial position through convolution operation.

The normalized network response after batching is $Z = BN(U) = [z_1, z_2, ..., z_{c'}]$.

Batch normalization can effectively prevent the gradient explosion and gradient disappearance in the network, and speed up the convergence speed of the network. Finally, the nonlinear interaction between the feature map channels is learned through the activation layer, and the complete dependence between the channels is obtained. It is expressed by the following formula:

$$S = WZ, W \in \mathbb{R}^{c'} \tag{10}$$

Among them, δ refers to the relu activation function. After multiple convolutions and activation calculations, the final EEG signal characteristics are expressed as $S^R = [s_1, s_2, ..., s_{c'}]$.

3.2 Dynamic learning of the intrinsic connection relationship between EEG channels located in different brain regions

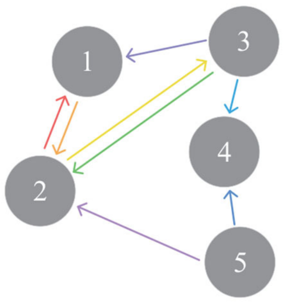
As the basis of ResGAT method, some basic knowledge about graph representation is introduced first. A directed connected graph can be defined as $G = V, E, W$, where V represents the node set with the number of $|V| = N$, and E represents the edge set connecting these nodes. Let $W \in \mathbb{R}^{N \times N}$ represents the adjacency matrix describing the connection between any two nodes in V, in which the entry of W in row i and column j measures the importance of the connection between node i and j. Figure 3 shows five nodes and edges connecting those nodes, as well as the adjacency matrix associated with the graph. The different colored arrows on the left side of the figure represent the edges connecting the source node and the target node, while the corresponding adjacency matrix is on the right side of the figure.

In the past, convolutional neural networks have been applied in many fields due to their powerful modeling capabilities, such as computer vision, speech recognition, and natural language processing. Due to its locality and translation invariance properties, it is very suitable for processing Euclidean data.However, many elements in the real world exist in the form of graph data, such as social networks, transportation networks, and drug discovery.The features extracted from the EEG signal and the distance between different electrodes are non-European data. Although the number of features on each signal channel is consistent, the distance between each adjacent electrode is uneven, and brain functional connectivity tends to capture global relationships among EEG channels. Therefore, the graph neural network is more suitable for learning the potential internal connections between different channels.At present, the graph attention network (GAT) (Veličković et al., 2017) is a widely used graph neural network. GAT achieves information aggregation in the spatial domain by introducing an attention mechanism, making the model pay more attention to the mutual influence between neighbor nodes, and applying it to EEG data to make the channels aggregate the characteristics of neighbor nodes and pay more attention to channels that are more relevant to themselves. Each EEG electrode can be regarded as a node of the graph, and the connection between the electrodes corresponds to the edge of the graph. The weights of all edges, which representing the functional relationship between electrodes, constitute the adjacency matrix of the graph. Therefore, GAT can learn the internal relationship between different EEG electrodes. As shown in the attention neural network in Figure 1, although GAT can describe the connection between different nodes according to their spatial positions, the connection between EEG channels should be determined in advance before applying it to the construction of emotion recognition model. In addition, it should be noted that the spatial location connection between EEG channels is different from the functional connection between them. In other words, closer spatial relationships may not guarantee closer functional relationships.

The flow of processing EEG signal features with GAT is shown in Figure 1. After data acquisition, preprocessing and feature extraction, EEG data are represented by undirected graph $G = V, E, W$. The data on can be represented as feature matrix $X^G \in \mathbb{R}^{n \times d}$, where n represents the number of electrodes and d represents the number of features extracted on each electrode channel. The constructed initial adjacency matrix $W^G \in \mathbb{R}^{n \times n}$, where n represents the number of electrode channels, characterizes the correlation between 2D space electrodes. Assume that each electrode channel has an internal relationship with the other 31 electrode channels, and is initialized as a diagonal matrix with the main diagonal of 0 and other values of 1. The feature combination extracted from each EEG channel is represented as a node in the graph neural network model, can be expressed as:

$$X^G = \vec{x_1}, \vec{x_2}, ..., \vec{x_n}, \vec{x_i} \in \mathbb{R}^d \tag{11}$$

In order to obtain sufficient expression ability to transform input features into higher-level features, at least one learnable linear



	1	2	3	4	5
1		🟧			
2	🟥		🟨		
3	🟪	🟩		🟦	
4					
5		🟪		🟦	

FIGURE 3
A directed graph and the corresponding adjacency matrix.

transformation is needed. A shared $H \in \mathbb{R}^{d' \times d}$ applies to all nodes to increase the expression ability of node features.

Then, a self attention mechanism is used on all nodes. At this time, the dimension of features on the nodes remains unchanged, which is $\mathbb{R}^{d'}$. The self attention mechanism is described as:

$$e_{ij} = Att(H\vec{x_i}, H\vec{x_j}) \tag{12}$$

where Att stands for self attention mechanism, and e_{ij} represents the importance of the characteristics of node j to i. Only the first-order neighbors of each node is calculated. In order to make the coefficients easy to compare between different nodes, the softmax function is used to normalize the attention coefficients of node j to other neighbor nodes.

$$a_{ij} = softmax(e_{ij}) = \frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_{ik})} \tag{13}$$

where a_{ij} is the coefficient of attention mechanism. In fact, the attention mechanism a is composed of a single-layer feedforward neural network, and the leakyrelu activation function is used for non-linear processing. Finally, the coefficient of attention mechanism can be expressed as:

$$\begin{aligned} a_{ij} &= \frac{exp(e_{ij})}{\sum_{k \in N_i} exp(e_{ik})} \\ &= \frac{exp(LeakyReLU(\vec{a}^T[H\vec{x^i} \| H\vec{x^j}]))}{\sum_{k=1}^{k \in N_i} exp(exp(LeakyReLU(\vec{a}^T[H\vec{x^i} \| H\vec{x^j}])))}, \vec{a} \in \mathbb{R}^{2d} \end{aligned} \tag{14}$$

where $\|$ indicates connection operation.

Then apply the normalized attention coefficient to the features corresponding to the node, and get the output after feature recalibration:

$$\vec{x_i'} = \sigma(\sum_{j \in N_i} a_{ij} H \vec{x_j}) \tag{15}$$

Veličković et al. (2017) found that it is beneficial to use multi head attention mechanism in graph neural network. Using K independent attention mechanisms at the same time, Formula 14 will produce K outputs. Then splice the above K outputs together, as shown in the following formula:

$$\vec{x_i'} = \|_{k=1}^{K} \sigma(\sum_{j \in N_i} a_{ij}^k H^k \vec{x_j}) \tag{16}$$

The output of each node changes to Kd'. In the experiment, the K is 2.

The aggregation process of multi head attention mechanism on nodes is shown in Figure 4.

The above is a complete graph convolution process. After multi-layer graph convolution, the EEG features will be further transmitted to the full connection layer, fused and classified with the extracted high-level abstract spatial features, and the S^G is obtained by batch normalization before full connection.

3.3 Feature fusion

The deep features extracted from convolution network and graph network are flattened and spliced, as shown below:

$$Output(S^R, S^G) = Concat(flatten(S^R), flatten(S^G)) \tag{17}$$

Finally, the softmax function is used to output the emotional state. The loss function of this model is the cross-entropy function, and the loss function is minimized using the Adam optimizer with an initial learning rate of 0.0001.

4 Experimental results and analysis

4.1 Performance analysis

The emotion classification network in the experiment consists of residual network and graph attention neural network. The residual network consists of multiple blocks, and each block contains two convolutional layers. In order to increase the fitting ability of the network, an activation layer is added after all convolutional layers. The first two residual blocks employ 64 filters with a size of 3×3 for convolution calculations, and the last two residual blocks use 128 filters of the same size.

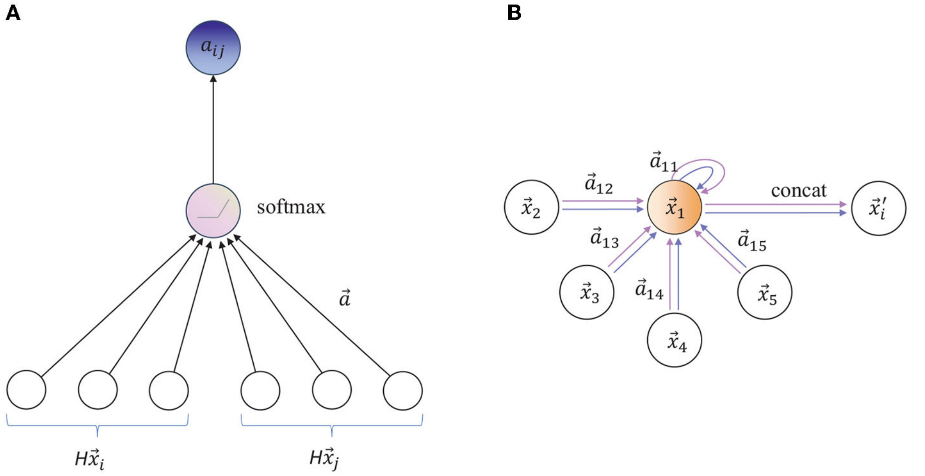


FIGURE 4
(A) The attention mechanism $a(H\vec{x}_i, H\vec{x}_j)$ is parameterized by the weight vector $a \in \mathbb{R}^{2d'}$. **(B)** Illustration of multi headed attention (k = 2) of node 1 in its neighborhood. Arrows of different colors indicate independent attention calculation. Aggregate features from each head are connected to obtain $\vec{x}_i'$.

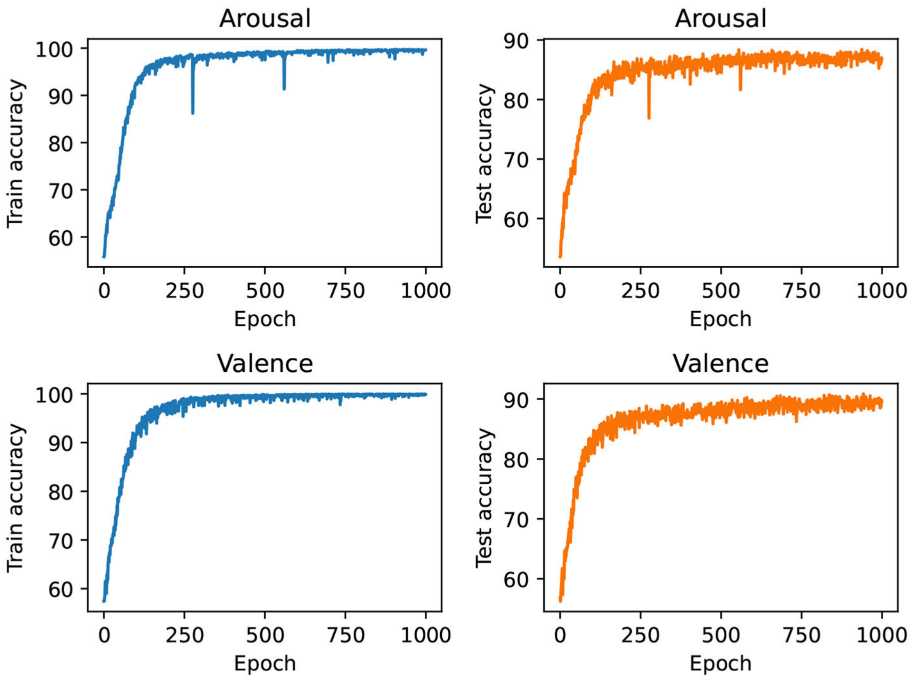


FIGURE 5
The training process of the proposed network in the two dimensions of Arousal and Valence.

The effect of the proposed network is verified on the DEAP dataset. In order to make the experimental results more objective, 10-fold cross-validation technique is used.

Figures 5, 6 show the training process of the proposed network on the dataset, Figure 5 shows the training process of the two emotional dimensions of arousal and valence on the DEAP data set, and Figure 6 shows the training process of the SEED data set process. Among them, when the training period is less than 750 in the DEAP dataset, the training accuracy and validation accuracy increase with the increase of the epoch. When the epoch is greater than 750, the training accuracy and validation longitude tend to be stable.

The classification accuracy (Acc) and F1 score (F1) are used to evaluate the performance of the proposed model. The emotion recognition results are shown in Figure 7, respectively. In the arousal dimension, the accuracy is 0.8706 and the F1 score is 0.8833. In the valence dimension, the recognition accuracy and F1 score are 0.8926 and 0.9042, respectively. In addition, 0.9773 Acc and 1.0 F1

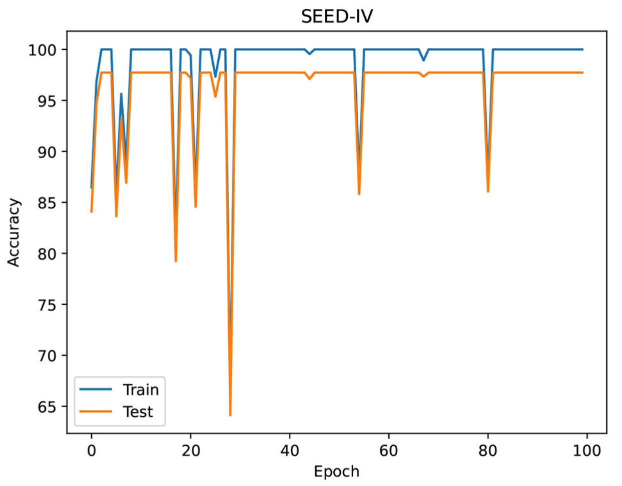


FIGURE 6
The training process of the proposed network in the SEED-IV.

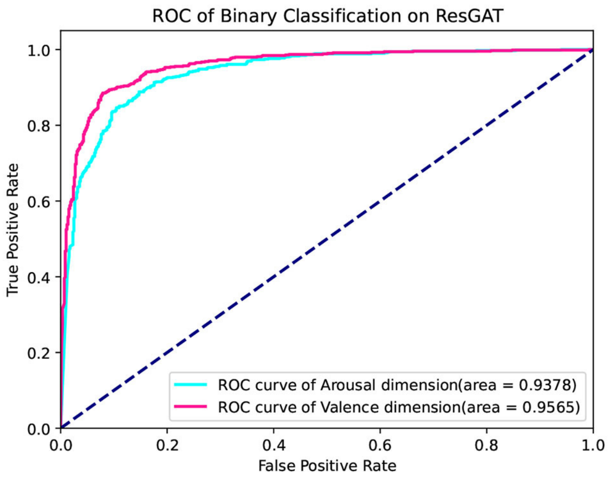


FIGURE 8
ROC curve of the proposed network.

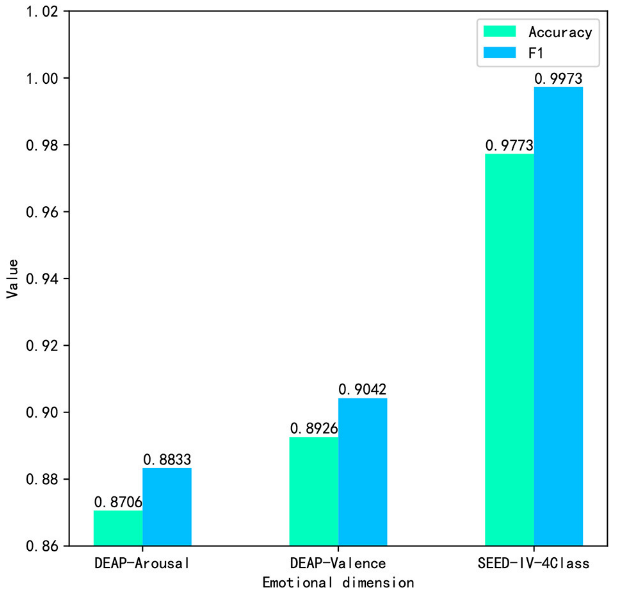


FIGURE 7
Emotion recognition results of the proposed network for binary classification.

were achieved on the four-category task of the SEED-IV dataset. The results of three classified tasks demonstrate the effectiveness of the proposed method.

The receiver operating curve (ROC) is also used to evaluate the performance of the proposed network. The ROC curve is located at the upper left triangle of the square, which reflects a more satisfactory classification rule. The higher the area under ROC Curve (AUC) value, the better the classification effect. Figure 8 shows the receiver operating curves on the two classifications of arousal and valence. The values of AUC in the two dimensions are 0.9378 and 0.9565, respectively. The relatively convex curve and high AUC value prove the excellent classification performance of the proposed classification network.

4.2 Comparison between the ensemble method and the single network

In the experiment, an independent GAT model and an independent ResNet model are constructed, respectively. The network structures of the independent GAT and the independent ResNet used in the experiment are consistent with those in the proposed ensemble ResGAT network. When these two independent models are used for emotion classification, the high-level abstract features are flattened and fed into a fully connected layer for classification. The 10-fold cross-validation technique are also used here, and other hyperparameters remain the same.

Firstly, the comparison is carried out on the emotion recognition accuracy. Compared with the GAT model, the proposed ResGAT improves the emotion recognition accuracies by 21.85% in the arousal dimension and 24.68% in the valence dimension. Compared with the ResNet model, the proposed ResGAT improves the emotion recognition accuracies by 1.64% in the arousal dimension and 2.99% in the valence dimension. Secondly, the comparison is carried out on the F1 scores. Compared with the GAT model, the proposed ResGAT improves the emotion recognition accuracies by 17.5% in the arousal dimension and 21.54% in the valence dimension. Compared with the ResNet model, the proposed ResGAT improves the emotion recognition accuracies by 0.9% in the arousal dimension and 2.26% in the valence dimension. The results show that the performance of the proposed ResGAT is obviously better than that of GAT, and it is also improved compared with ResNet.

In addition, it was also verified on the SEED-IV dataset, and the model recognition results are shown in Table 2. The

TABLE 1 The results of ResGAT and the two single networks.

Emotion dimension	Recognition results					
	ResGAT		GAT		ResNet	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Arousal	0.8706	0.8833	0.6521	0.7083	0.8542	0.8743
Valence	0.8926	0.9042	0.6458	0.6888	0.8627	0.8816

TABLE 2 The results of ResGAT and the two single networks (SEED-IV).

Emotion dimension	Recognition results					
	ResGAT		GAT		ResNet	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Four classification	0.9773	0.9973	0.9522	0.9635	0.9021	0.9251

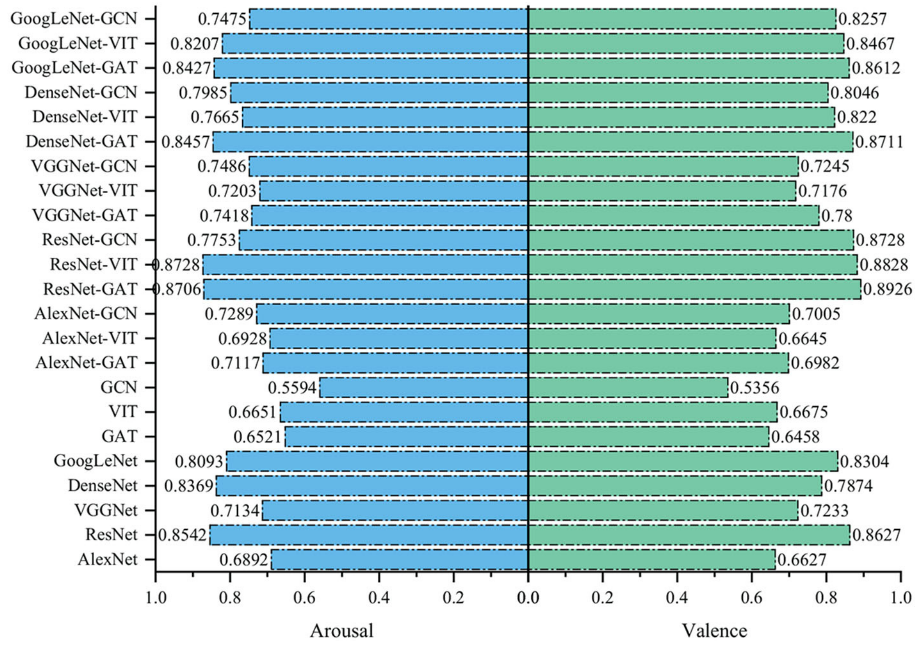


FIGURE 9
Recognition results of the single networks and the ensemble models integrated by the single networks.

experimental results in Tables 1, 2 show that the performance of the integrated network is better than that of a single network, because each network in the integrated network can extract different information, that is, the spatial position information of electrodes and the internal relationship of EEG channels in different brain regions. These two kinds of information complement each other to improve the model recognition performance.

To further verify the complementarity between the electrode spatial position information and the intrinsic connection relationship between EEG channels, a variety of CNN networks, graph convolution network (GCN) (Kipf and Welling, 2016), vision in transformer network (VIT) (Dosovitskiy et al., 2020) and the ensemble models integrated by the above networks are constructed. The CNN networks constructed specifically include Alex (Krizhevsky et al., 2017), VGGNet (Simonyan and Zisserman, 2014), DenseNet (Huang et al., 2017), and GoogLeNet (Szegedy et al., 2015), which focus on extracting the electrode spatial position information. Similar to GAT, GCN and VIT focus on extracting the intrinsic connection relationship between EEG channels. An ensemble model is constructed by a CNN network and GCN, or by a CNN network and VIT, which means these ensemble models can capture both the electrode spatial position information and the intrinsic connection relationship between EEG channels.

In the CNN networks, the AlexNet structure is affected by the size of the input data. Compared with the structure in the reference (Krizhevsky et al., 2017), the maximum pooling is removed, and the size of the convolution kernel is modified. Other structures remain unchanged. Compared with the residual network in this paper, the structure of VGG only removes the spanning connection. DenseNet employs 1×1 convolutions for better data representation, where the depth of the convolutional layers is 9. GoogLeNet contains a multi-branch convolution structure, which uses convolution kernels of 3×3, 5×5, and 7×7, respectively.

GCN is a natural extension of convolutions on graph structure. Because GCN is suitable for extracting structural features of graphs and can customize local receptive fields, it is widely used in network analysis, traffic prediction. and recommender systems. Inspired by the reference (Kipf and Welling, 2016), a spectral domain-based graph convolutional network is constructed, which contains two convolutional layers. Transformer has been successfully applied in natural language processing and computer vision. Therefore, in the experiment, the standard transformer is directly applied to the EEG signal features with minimal modification. Similar to the reference (Dosovitskiy et al., 2020), the 3D feature matrix is divided into small blocks, and the linear embedding sequence of these blocks is provided as the input of the transformer.

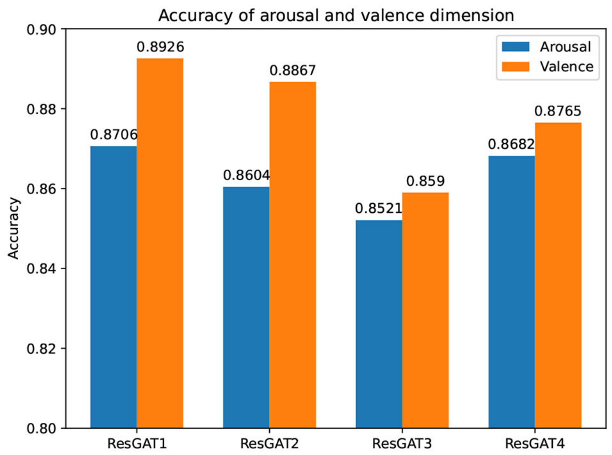


FIGURE 10
Recognition results of ResGAT1, ResGAT2, ResGAT3, and ResGAT4.

TABLE 3 Recognition results using residual networks.

	3D	2D
DEAP-Arousal	0.8542	0.6203
DEAP-Valence	0.8627	0.6136
SEED-IV	0.9021	0.9020

The recognition results of the above networks and the ensemble models are shown in Figure 9. Most of the ensemble models have higher classification accuracy than the corresponding single network, which proves that the electrode spatial position information and the intrinsic connection relationship between EEG channels are complementary to emotion classification. ResNet has the highest classification accuracy in a single network, which achieves 85.42% classification accuracy in arousal dimension and 86.27% classification accuracy invalence dimension, respectively. In the integrated models, ResGAT achieves the highest classification accuracy in the valence dimension, and ResNet-VIT achieves the highest classification accuracy in the arousal dimension. Among the above ensemble models, the ensemble models including GAT performs well in emotion recognition tasks. Combining GAT with any kind of CNN, the classification accuracy can be improved. The experimental results show that GAT has better information capture ability than GCN and VIT, and is more suitable for combining with convolutional networks, which makes the extracted high-level abstract features contain relatively less redundant and irrelevant components.

4.3 ResGAT with different model structures

In addition to the ResGAT (ResGAT1) proposed in this paper, three other ResGAT models (ResGAT2, ResGAT3, and ResGAT4) are also constructed. In the ResGAT1, all 3×3 filters are used in the residual network, and the number of multi-head attention in all graph attention layers in GAT is 2. ResGAT2 sets the multi-head attention number of GAT to 4. ResGAT3 is twice as deep as ResGAT1. ResGAT4 uses 3×3

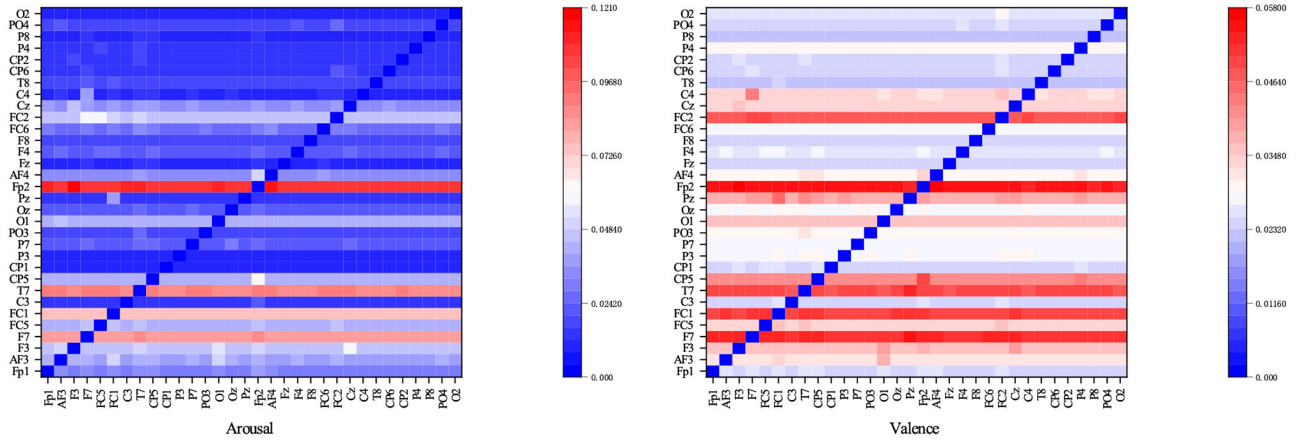


FIGURE 11
Heatmap representation of adjacency matrices in GAT on DEAP-Arousal and DEAP-Valence affective dimensions.

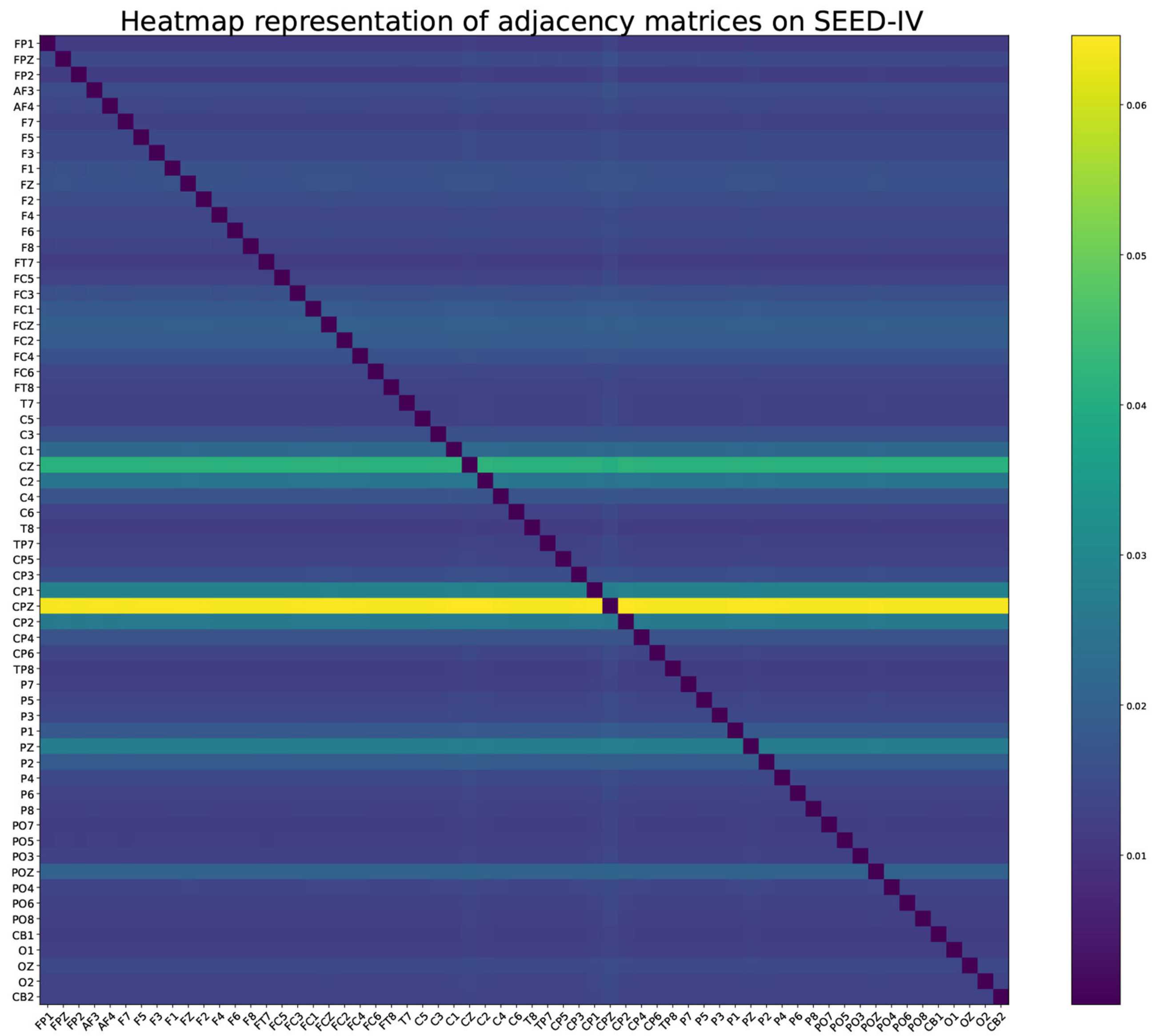


FIGURE 12
Heatmap representation of adjacency matrices in GAT on SEED-IV.

and 5×5 filters to cross the residual structure on the basis of increasing the network depth. The other parameters in ResGAT for the three comparisons remain unchanged. All samples of subjects and 10-fold cross-validation technique are also used here. The recognition results of ResGAT with different structures under the two sentiment annotation schemes are shown in Figure 10.

Compared with ResGAT2, ResGAT3, and ReGAT4, the recognition accuracy of the proposed ResGAT in the arousal dimension is improved by 1.02, 1.85, and 0.24%, respectively. In the dimension of valence, the recognition accuracy of the proposed ResGAT is improved by 0.59, 3.36, and 1.61%, respectively. It can be seen from the comparison results that increasing the complexity and depth of the network will not necessarily improve the accuracy, but will increase the calculation of the model. Therefore, it is very important to choose the appropriate network structure.

4.4 Sensitivity analysis

To further prove that the proposed network can extract the spatial domain information of EEG signal channels and learn the internal relationship of different EEG signal channels, the information extraction ability of the proposed model is analyzed.

In the proposed method, the three-dimensional sparse feature matrix and deep residual network are used to capture the dependence between local EEG signal channels. As a contrast, the deep residual network model is also used to deal with the

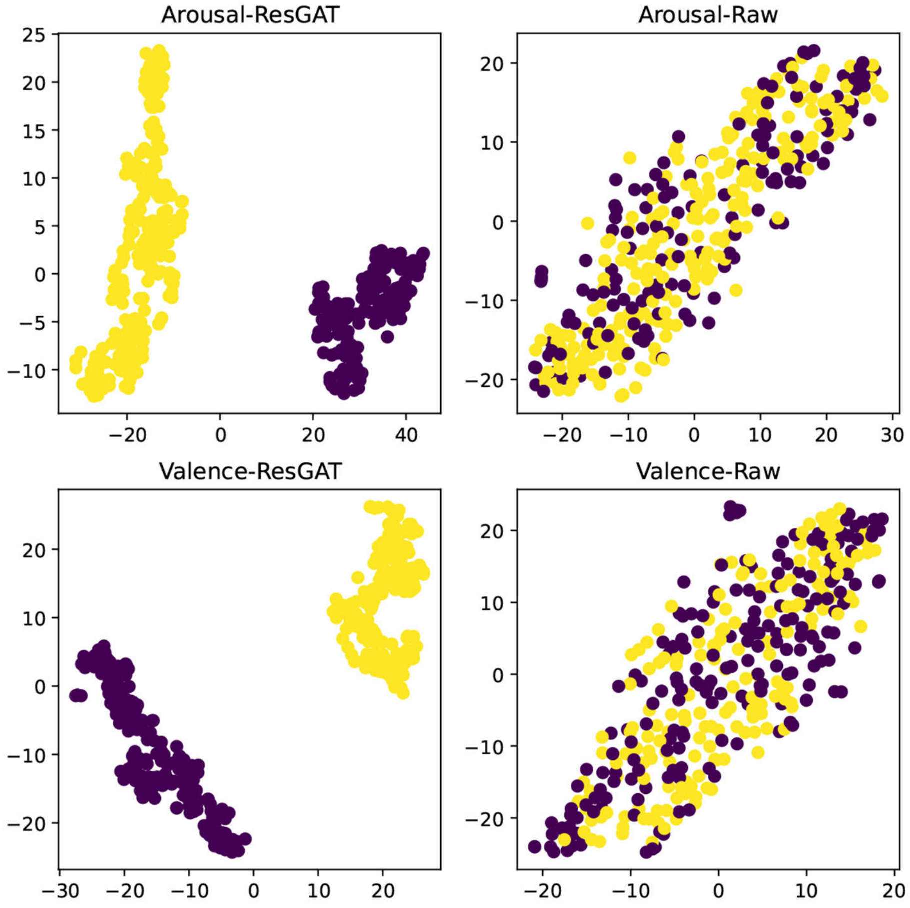


FIGURE 13
t-SNE analysis on the emotional dimensions of DEAP-Arousal and DEAP-Valence.

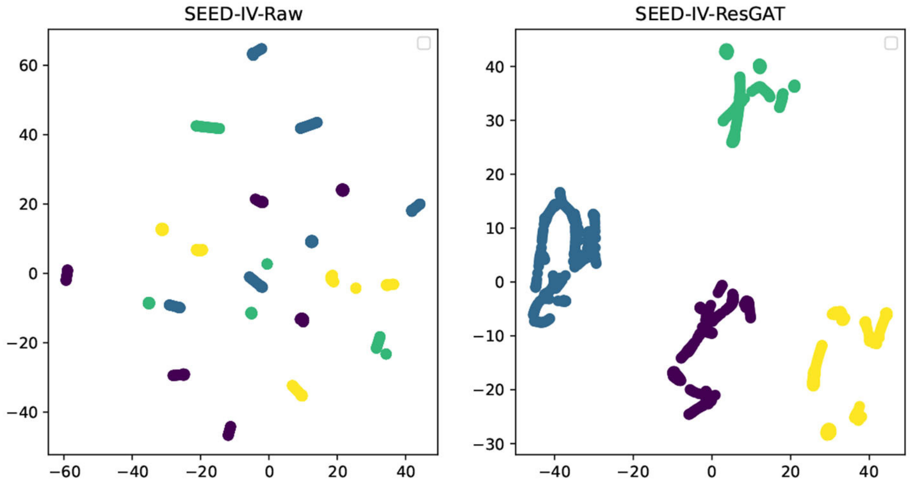


FIGURE 14
t-SNE analysis on the emotional dimensions of SEED-IV.

same time-domain characteristics without mapping and arranging according to the international 10/20 standard. Six features of 32 channels can construct a two-dimensional feature matrix with a size of 32×6. The hyperparameters in the experiment remain unchanged. The recognition results using 3D feature matrix and 2D feature matrix, respectively, are shown in Table 3.

TABLE 4 Details of previous research.

Study	Feature	Classifier	DEAP		SEED-IV
			Arousal	Valence	
Samara et al. (2016)	Band power	SVM	0.7367	0.8599	–
Guo et al. (2017)	DWT	SVM	0.6279	0.6021	–
Alhagry et al. (2017)	Raw EEG signals	LSTM	0.8565	0.8545	–
Yang and Liu (2019)	Differential entropy	TCN	0.7140	0.7440	–
Tripathi et al. (2017)	Statistical parameters	DNN	0.7313	0.7578	
		CNN	0.7336	0.8141	0.8599
Gao et al. (2022)	Differential entropy	GCN	0.8193	0.8177	-
Zhong et al. (2020)	Differential entropy	RGNN	-	-	0.7750
Du et al. (2022)	Differential entropy	MD-GCN	-	-	0.9083
Li et al. (2023)	Differential entropy	FGCN	-	-	0.7714
Vafaei et al. (2023)	Time domain features	SAETM	0.8037	0.8173	-
The proposed method	Time domain features	ResGAT	0.8706	0.8926	0.9773

Compared with the two-dimensional feature matrix, the accuracy of emotion recognition of the three-dimensional feature matrix in the arousal dimension is increased by 23.39%, the accuracy of emotion recognition in the valence dimension is increased by 24.91%, 0.01% improvement on the SEED-IV dataset. The results show that three-dimensional feature matrix and deep residual network can effectively extract local dependency information of signal channels.

In order to illustrate the intrinsic connection relationship between EEG channels mined by GAT, the adjacency matrix learned during the training process is displayed. The adjacency matrix is affected by the input data. Input all training data into the GAT in turn to obtain the adjacency matrix corresponding to each sample. The average value of all adjacency matrices can construct a heat map, as shown in Figures 11, 12.

It can be clearly seen that the graph neural network is not limited by distance when collecting neighbor node information in 32 electrode channels. In terms of arousal and valence emotion, C4 electrode channel pays more attention to FC5 channel when aggregating neighbor node information, and FC2 channel pays more attention to F7 and FC5 channels, and CP5 pays more attention to FP2 channel. In the SEED-IV dataset, all nodes focus more on the four channels CZ, CPZ, PZ, and POZ.

4.5 t-SNE analysis

In order to demonstrate the effectiveness of ResGAT in extracting high-level abstract features, the t-SNE tool is used to visually analyze the features in two-dimensional space, these features extract all data from a single person. As shown in Figures 13, 14, the input data of the model and the high-level abstract features extracted by ResGAT are displayed in the two emotional dimensions of arousal and valence. The results in the figure demonstrate the effectiveness of the proposed ResGAT in extracting affective state discriminative features.

4.6 Comparison with existing methods

The recognition performance of the proposed method is compared with several existing studies. The dataset and labeling scheme are the same for all reported studies. Table 4 details the features and classifiers used in the comparative study. Since recognition accuracy and F1 score are the most commonly used, these two indicators are adopted for comparison. As shown in Table 4, the performance of our approach is better than the comparison methods in both the arousal dimension and the valence dimension. The comparison results show that our approach is excellent in multichannel EEG emotion recognition.

5 Conclusion

A novel ensemble deep learning framework is proposed in this work. In the framework, the residual network is employed to extract the spatial position information of the electrode channel through the 3D characteristic matrix. The graph neural network is utilized to learn the neural functional connections between different brain regions, and the multi-head self-attention mechanism is used to adaptively adjust the adjacency matrix in the network. The results show the proposed ResGAT framework makes full use of the electrode spatial position information and the intrinsic connection relationship between EEG channels located in different brain regions. Moreover, the emotion recognition performance of the proposed method is compared with some existing methods and shows advantages, which proves the feasibility and effectiveness of the proposed emotion recognition method.

The experiments in this manuscript were conducted on public datasets DEAP and SEED, and the proposed emotion recognition method demonstrated good performance. However, the number of subjects on the dataset is limited, and the effectiveness of its use in a large population needs further verification. The monitoring and regulation of emotional state is of great significance for the psychological and physiological health of individuals. For example, in clinical treatment, monitoring and regulating emotional states can help doctors better understand patients' emotional states, thereby providing more personalized and effective treatment plans for patients. In daily life, monitoring and regulating emotional states can help individuals better manage their emotions and improve their quality of life. Moreover, the emotion recognition performance of the proposed method is compared with some existing methods and shows advantages, which proves the feasibility and effectiveness of the proposed emotion recognition method. In addition, the proposed emotion recognition classification model can also be applied in disease diagnosis, such as identification of patients with depression; issuing execution commands to control external devices, helping patients to carry out active rehabilitation training; diagnosis of schizophrenia; quantifying the neurophysiological changes associated with a variety of work-related physical activities (Ismail et al., 2023).

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: http://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

Author contributions

HC designed and organized the study and wrote this article. HC, YL, and YC collected and analyzed the data. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by National Natural Science Research Fund (61872126), Key Scientific Research Project of Colleges and Universities in Henan Province (19A520004), the Fundamental Research Funds for the Central Universities (NSFRF1616), and Programs for Science and Technology Development of Henan province No. 222102210078.

Acknowledgments

We would like to thank all participants of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alhagry, S., Fahmy, A. A., and El-Khoribi, R. A. (2017). Emotion recognition based on EEG using LSTM recurrent neural network. *Int. J. Adv. Comput. Sci. Appl.* 8, 355–358. doi: 10.14569/IJACSA.2017.081046

Alhaj, H., Wisniewski, G., and McAllister-Williams, R. H. (2011). The use of the EEG in measuring therapeutic drug action: focus on depression and antidepressants. *J. Psychopharmacol.* 25, 1175–1191. doi: 10.1177/0269881110388323

Castelnovo, A., Riedner, B. A., Smith, R. F., Tononi, G., Boly, M., and Benca, R. M. (2016). Scalp and source power topography in sleepwalking and sleep terrors: a high-density EEG study. *Sleep* 39, 1815–1825. doi: 10.5665/sleep.6162

Chao, H., and Dong, L. (2020). Emotion recognition using three-dimensional feature and convolutional neural network from multichannel EEG signals. *IEEE Sens. J.* 21, 2024–2034. doi: 10.1109/JSEN.2020.3020828

Chen, W., Wang, Z., Xie, H., and Yu, W. (2007). Characterization of surface EMG signal based on fuzzy entropy. *IEEE Trans. Neural Syst. Rehabil. Eng.* 15, 266–272. doi: 10.1109/TNSRE.2007.897025

Cui, H., Liu, A., Zhang, X., Chen, X., Wang, K., and Chen, X. (2020). EEG-based emotion recognition using an end-to-end regional-asymmetric convolutional neural network. *Knowledge Based Syst.* 205, 106243. doi: 10.1016/j.knosys.2020.106243

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929.*

Du, G., Su, J., Zhang, L., Su, K., Wang, X., Teng, S., et al. (2022). A multi-dimensional graph convolution network for EEG emotion recognition. *IEEE Trans. Instrument. Meas.* 71, 1–11. doi: 10.1109/TIM.2022.3204314

Fürbass, F., Kural, M. A., Gritsch, G., Hartmann, M., Kluge, T., and Beniczky, S. (2020). An artificial intelligence-based EEG algorithm for detection of epileptiform EEG discharges: validation against the diagnostic gold standard. *Clin. Neurophysiol.* 131, 1174–1179. doi: 10.1016/j.clinph.2020.02.032

Gao, Y., Fu, X., Ouyang, T., and Wang, Y. (2022). EEG-GCN: spatio-temporal and self-adaptive graph convolutional networks for single and multi-view EEG-based emotion recognition. *IEEE Signal Process. Lett.* 29, 1574–1578. doi: 10.1109/LSP.2022.3179946

Gogna, A., Majumdar, A., and Ward, R. (2016). Semi-supervised stacked label consistent autoencoder for reconstruction and analysis of biomedical signals. *IEEE Trans. Biomed. Eng.* 64, 2196–2205. doi: 10.1109/TBME.2016.2631620

Guo, K., Candra, H., Yu, H., Li, H., Nguyen, H. T., and Su, S. W. (2017). "EEG-based emotion classification using innovative features and combined SVM and HMM classifier," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (IEEE), 489–492.

Guozhen, Z., Jinjing, S., Yan, G., Yongjin, L., Lin, Y., and Tao, W. (2016). Advances in emotion recognition based on physiological big data. *J. Comput. Res. Dev.* 53, 80–92. doi: 10.7544/issn1000-1239.2016.20150636

Hamilton, M. (1960). A rating scale for depression. *J. Neurol. Neurosurg. Psychiatry* 23, 56.

Haufe, S., Nikulin, V. V., Müller, K.-R., and Nolte, G. (2013). A critical assessment of connectivity measures for EEG data: a simulation study. *Neuroimage* 64, 120–133. doi: 10.1016/j.neuroimage.2012.09.036

Hu, X., Chen, J., Wang, F., and Zhang, D. (2019). Ten challenges for EEG-based affective computing. *Brain Sci. Adv.* 5, 1–20. doi: 10.1177/2096595819896200

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4700–4708.

Ismail, L., Karwowski, W., Hancock, P. A., Taiar, R., and Fernandez-Sumano, R. (2023). Electroencephalography (EEG) physiological indices reflecting human physical performance: a systematic review using updated prisma. *J. Integr. Neurosci.* 22, 62. doi: 10.31083/j.jin2203062

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*. doi: 10.48550/arXiv.1609.02907

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2011). DEAP: a database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90. doi: 10.1145/3065386

Li, J., Zhang, Z., and He, H. (2018a). Hierarchical convolutional neural networks for EEG-based emotion recognition. *Cogn. Comput.* 10, 368–380. doi: 10.1007/s12559-017-9533-x

Li, M., Qiu, M., Kong, W., Zhu, L., and Ding, Y. (2023). Fusion graph representation of EEG for emotion recognition. *Sensors* 23, 1404. doi: 10.3390/s23031404

Li, M., Xu, H., Liu, X., and Lu, S. (2018b). Emotion recognition from multichannel EEG signals using k-nearest neighbor classification. *Technol. Health Care* 26, 509–519. doi: 10.3233/THC-174836

Li, Y., Huang, J., Zhou, H., and Zhong, N. (2017). Human emotion recognition with electroencephalographic multidimensional features by hybrid deep neural networks. *Appl. Sci.* 7, 1060. doi: 10.3390/app7101060

Li, Y., Zheng, W., Zong, Y., Cui, Z., Zhang, T., and Zhou, X. (2018c). A bi-hemisphere domain adversarial neural network model for EEG emotion recognition. *IEEE Trans. Affect. Comput.* 12, 494–504. doi: 10.1109/TAFFC.2018.2885474

Mohammadi, M., Al-Azab, F., Raahemi, B., Richards, G., Jaworska, N., Smith, D., et al. (2015). Data mining EEG signals in depression for their diagnostic value. *BMC Med. Inform. Decis. Mak.* 15, 108. doi: 10.1186/s12911-015-0227-6

Niedermeyer, E., and da Silva, F. L. (2005). *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Lippincott Williams & Wilkins.

Nolte, G., Bai, O., Wheaton, L., Mari, Z., Vorbach, S., and Hallett, M. (2004). Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clin. Neurophysiol.* 115, 2292–2307. doi: 10.1016/j.clinph.2004.04.029

Nolte, G., Ziehe, A., Nikulin, V. V., Schlögl, A., Krämer, N., Brismar, T., et al. (2008). Robustly estimating the flow direction of information in complex physical systems. *Phys. Rev. Lett.* 100, 234101. doi: 10.1103/PhysRevLett.100.234101

Ramirez, P. M., Desantis, D., and Opler, L. A. (2001). EEG biofeedback treatment of add: a viable alternative to traditional medical intervention? *Ann. N. Y. Acad. Sci.* 931, 342–358. doi: 10.1111/j.1749-6632.2001.tb05789.x

Samara, A., Menezes, M. L. R., and Galway, L. (2016). "Feature extraction for emotion recognition and modelling using neurophysiological data," in *2016 15th International Conference on Ubiquitous Computing and Communications and 2016 International Symposium on Cyberspace and Security (IUCC-CSS)* (IEEE), 138–144.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. doi: 10.48550/arXiv.1409.1556

Sung, M., Marci, C., and Pentland, A. (2005). *Objective Physiological and Behavioral Measures for Identifying and Tracking Depression State in Clinically Depressed Patients*. Massachusetts Institute of Technology Media Laboratory, Cambridge, MA.

Supp, G. G., Schlögl, A., Trujillo-Barreto, N., Müller, M. M., and Gruber, T. (2007). Directed cortical information flow during human object recognition: analyzing induced EEG gamma-band responses in brain's source space. *PLoS ONE* 2, e684. doi: 10.1371/journal.pone.0000684

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Tripathi, S., Acharya, S., Sharma, R. D., Mittal, S., and Bhattacharya, S. (2017). "Using deep and convolutional neural networks for accurate emotion classification on deap dataset," in *Twenty-ninth IAAI conference*.

Vafaei, E., Nowshiravan Rahatabad, F., Setarehdan, S. K., Azadfallah, P., et al. (2023). Extracting a novel emotional EEG topographic map based on a stacked autoencoder network. *J. Healthcare Eng.* 2023, 9223599. doi: 10.1155/2023/9223599

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*. doi: 10.48550/arXiv.1710.10903

Xing, X., Li, Z., Xu, T., Shu, L., Hu, B., and Xu, X. (2019). SAE+ LSTM: a new framework for emotion recognition from multi-channel EEG. *Front. Neurorobot.* 13, 37. doi: 10.3389/fnbot.2019.00037

Yang, L., and Liu, J. (2019). "EEG-based emotion recognition using temporal convolutional network," in *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)* (IEEE), 437–442.

Zheng, W.-L., Liu, W., Lu, Y., Lu, B.-L., and Cichocki, A. (2018). Emotionmeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybernet.* 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zhong, P., Wang, D., and Miao, C. (2020). EEG-based emotion recognition using regularized graph neural networks. *IEEE Trans. Affect. Comput.* 13, 1290–1301. doi: 10.1109/TAFFC.2020.2994159

TYPE Brief Research Report
PUBLISHED 28 August 2023
DOI 10.3389/fpsyg.2023.1129830

Vowel production: a potential speech biomarker for early detection of dysarthria in Parkinson's disease

Virginie Roland[1,2]*, Kathy Huet[1,2], Bernard Harmegnies[2], Myriam Piccaluga[1,2], Clémence Verhaegen[1,2] and Véronique Delvaux[1,2,3]

[1]Metrology and Language Sciences Unit, Mons, Belgium, [2]Research Institute for Language Science and Technology, University of Mons, Mons, Belgium, [3]National Fund for Scientific Research, Brussels, Belgium

OPEN ACCESS

EDITED BY
Kevin Tang,
Heinrich Heine University of Düsseldorf,
Germany

REVIEWED BY
Min Ney Wong,
Hong Kong Polytechnic University,
Hong Kong SAR, China
Emily Wang,
Rush University Medical Center, United States
Rita Cardoso,
Campus Neurológico Sénior (CNS), Portugal

*CORRESPONDENCE
Virginie Roland
✉ Virginie.roland@umons.ac.be

RECEIVED 22 December 2022
ACCEPTED 26 July 2023
PUBLISHED 28 August 2023

CITATION
Roland V, Huet K, Harmegnies B, Piccaluga M, Verhaegen C and Delvaux V (2023) Vowel production: a potential speech biomarker for early detection of dysarthria in Parkinson's disease.
Front. Psychol. 14:1129830.
doi: 10.3389/fpsyg.2023.1129830

Objectives: Our aim is to detect early, subclinical speech biomarkers of dysarthria in Parkinson's disease (PD), i.e., systematic atypicalities in speech that remain subtle, are not easily detectible by the clinician, so that the patient is labeled "non-dysarthric." Based on promising exploratory work, we examine here whether vowel articulation, as assessed by three acoustic metrics, can be used as early indicator of speech difficulties associated with Parkinson's disease.

Study design: This is a prospective case–control study.

Methods: Sixty-three individuals with PD and 35 without PD (healthy controls-HC) participated in this study. Out of 63 PD patients, 43 had been diagnosed with dysarthria (DPD) and 20 had not (NDPD). Sustained vowels were recorded for each speaker and formant frequencies were measured. The analyses focus on three acoustic metrics: individual vowel triangle areas (tVSA), vowel articulation index (VAI) and the Phi index.

Results: tVSA were found to be significantly smaller for DPD speakers than for HC. The VAI showed significant differences between these two groups, indicating greater centralization and lower vowel contrasts in the DPD speakers with dysarhtria. In addition, DPD and NDPD speakers had lower Phi values, indicating a lower organization of their vowel system compared to the HC. Results also showed that the VAI index was the most efficient to distinguish between DPD and NDPD whereas the Phi index was the best acoustic metric to discriminate NDPD and HC.

Conclusion: This acoustic study identified potential subclinical vowel-related speech biomarkers of dysarthria in speakers with Parkinson's disease who have not been diagnosed with dysarthria.

KEYWORDS
Parkinson's disease, dysarthria, early detection, acoustic analyses, vowel production

1. Introduction

Parkinson's disease (PD) is a disease that causes degeneration of the nervous system, especially the substructures that control movement. This disease is characterized with the progressive loss of dopaminergic neurons. One of the strongest risk factors associated with disease is aging. Therefore, with the advancing age of the world's population, the early detection of characteristic patterns of the disease is a health challenge. Moreover, as pointed out by Agüera-Ortiz et al. (2021), neurodegenerative diseases are an increased global economic and healthcare system burden.

The motor symptoms associated with PD may involve bradykinesia accompanied by resting tremor and/or rigidity. Patients with PD may experience symptoms that significantly affect their quality of life. Hypokinetic dysarthria, which includes a wide variety of speech disorders associated with PD, is one of them (Sapir, 2014; Postuma et al., 2015). Classical perceptual and acoustic studies have repeatedly shown that dysarthria affects the respiratory, phonatory and/or articulatory aspects of speech on both segmental and suprasegmental levels, i.e., dysprosody (e.g., Duffy, 2019). Dysarthric speech is then characterized by reduced loudness, monopitch and/or monoloudness, harsh voice, imprecise speech articulation or inappropriate silences. On the articulatory level, most of previous studies have focused on imprecision in consonant production (e.g., Ackermann and Ziegler, 1991) and vowel articulation (e.g., Skodda et al., 2011), in particular at moderate and advanced stages of the disease and for patients with moderate dysarthria (e.g., Martel-Sauvageau et al., 2015; Dias et al., 2016; Martel-Sauvageau and Tjaden, 2017; Duez et al., 2020). One of the most commonly reported impairments in individuals with PD who have hypokinetic dysarthria is difficulty in producing consonants accurately as typically evidenced by oral diadochokinetic tasks (e.g., Ackermann and Ziegler, 1991; McRae and Tjaden, 1998; Wong et al., 2011; Karlsson and Hartelius, 2019). The stops, affricates, and fricatives are often distorted, potentially due to the reduced range and strength of the movements used to produce them. Ackermann and Ziegler (1991), Ackermann et al. (1995) have suggested that this may be caused by PD patients trying to maintain a fluid speaking rate, at the risk of causing articulatory undershoot. However, research on muscle activity during speech in PD has yielded inconsistent results (for a review, see Walsh and Smith, 2012). For example, Mcauliffe et al. (2006) demonstrated that listeners perceived consonants as being produced with undershoot but did not find a corresponding reduction in tongue-palate contact on EPG examination. Wong Ackermann and Ziegler, 1991(2011) even found that some individuals with PD had increased distance of tongue movement when producing certain coronal and velar consonants. More research is needed to fully understand the dynamics of supra-laryngeal articulators in PD.

Interestingly, aspects of speech production related to sound resonance – central to the production of vowels, diphthongs and approximants – are thought to be preserved in PD, but have been little studied (Goberman et al., 2002). However, accurate execution of motor plans involving the jaw, tongue, and lips is as essential to the production of vowel-like sounds as it is to the production of consonants. The present study concerns vowel production in francophone speakers with PD. The study of vowel-like sounds, especially if it encompasses both stable and dynamically changing phases, could be a valuable area of research for better understanding speech motor control in PD. Indeed, adequate vowels and diphtongs can only be produced if one can both maintain stable articulatory configurations over time and properly execute dynamic sequences of coordinated articulatory gestures. Note that the available evidence about dysarthria in PD is often based on the English language although the variability across different languages should be considered in a speech assessment framework (Rusz et al., 2021). As it happens, English and French differ in their vowel inventory as well in the phonological structures associated with dynamic vocalic sounds: English counts a dozen vowels and several diphtongs, whereas the French inventory contains only monophtongs but also three approximants /w, ɥ, j/ resulting in sequences such as V.C$_{[glide]}$V (kayak, brouillard, kiwi, etc.) (Fougeron and Smith, 1999).

A previous acoustic study was conducted on the speech productions of PD patients with mild dysarthria compared to healthy speakers (Delvaux et al., 2016). We specifically focused on the production of steady vowels and intervocalic glides, based on the hypothesis that parkinsonian speech production may be characterized by vowel centralization resulting in a reduction of the vowel space (Kent and Kim, 2003; Skodda et al., 2011; Mollaei et al., 2016). The study involved two groups of participants: 9 people (6 men and 3 women) with intermediate-stage Parkinson's disease (according to the Hoehn and Yahr scale), and a healthy group of 10 people (5 men and 5 women) who had no speech or language disorders. Acoustic measurements were taken for sustained oral vowels, including overall duration and frequencies of formants (F1, F2) at the midpoint of the vowel, and individual triangular vowel space areas (tVSA) were calculated. Results showed that the mean areas did not differ significantly between the PD group and the control group. These results suggest that although there is more variation in the production of sustained vowels among persons with PD (here, with mild dysarthria), the size of their vowel spaces is not significantly different from those of HC. Other, complementary acoustic metrics would have to be used to capture subtle alterations in vowel production when dysarthria is mild.

In fact, a variety of acoustic metrics can help identify alterations in the productions of PD compared to HC speakers. In some studies, vowels metrics are calculated to identify a possible marker of the progression of the disease in PD (Sapir et al., 2010; Skodda et al., 2012; Rusz et al., 2013; Rountrey and Molett, 2020). The tVSA is one of the most frequently used acoustic indicator for the evaluation of imprecision in vowel production, as it can reflect major changes in articulatory movements in speech disorders. However, some researchers suggest that the tVSA is not sensitive enough to signal mild and moderate forms of dysarthria (Sapir et al., 2007; Neel, 2008; Skodda et al., 2011). Sapir et al. (2007) suggest that variations across speakers can statistically reduce the differences between those with mild dysarthria and those without dysarthria. Yet, a better understanding of the potential impairment in patients with mild dysarthria and those without dysarthria in PD is essential to identify speech deteriorations in the early stages of the disease. As far as we know, only the study conducted by Audibert and Fougeron (2012) proposes a direct comparison of several metrics derived from F1/F2 measurements to describe and quantify the possible distortions to be observed in the vowel space of French dysarthric speakers.

In the present study, we have selected three acoustic metrics which have been previously tested with PD patients for the complementary information they provide: (1) the triangular vowel space area (tVSA) representing the maximum working space of each individual, (2) the vowel articulation index (VAI), which is the reciprocal of the formant centralization ratio (Roy et al., 2009; Sapir et al., 2010) and (3) the PHI index which expresses the relationship between inter-category distance and intra-category variability within the vowel space considered as an organized system of phonemic categories (Huet and Harmegnies, 2000). Inter-category distance can be considered as a centralization metric and intra-category variability as an index of (in) consistency in the production of acoustic targets. Audibert and Fougeron (2012) suggest that metrics of intra-category dispersion and centralization are complementary. Their results show that intra-category variability is only weakly correlated with other metrics, arguing for its informational potential since it cannot be predicted by other measures.

Note that in clinical practice, the detection of alterations by an objective approach is intended to complete the perceptual analysis made by the clinician. In fact, the methodology of this study is designed to be applicable to a professional practice. The productions requested from the speakers are those that would be expected from a typical speech assessment by a speech therapist (recommendations by the American Speech-Language-Hearing Association, 2004). The number of productions per participant is intentionally limited, less than in typical experimental phonetics studies, which allows to eliminate a possible fatigue effect among participants.

Besides, while the majority of patients interviewed declare themselves dissatisfied with their communication performance (Miller et al., 2011), only a few individuals initiate speech therapy, even though available statistics tend to show an increase in speech therapy over the past few decades (Hartelius and Svensson, 1994; Kalf et al., 2011; Sunwoo et al., 2014; Schalling et al., 2017). Also, when it is present, speech therapy appears rather late in the course of dysarthria, with patients presenting moderate to severe dysarthria, whereas many recommendations suggest early speech therapy (Gentilhomme et al., 2020).

The purpose of this study is to use acoustic metrics to objectively identify speech biomarkers in oral vowel production in PD patients who do not have hypokinetic dysarthria, in order to identify speech alterations that are difficult to detect by even careful listening by the clinician. The long-term goal of this research is to identify early, subtle symptoms of dysarthria as a prodromal marker of PD. Indeed, recent evidence suggests that speech atypicalities might be the first motor signs to emerge (Sapir, 2014; Hlavnička et al., 2017; Rusz et al., 2021).

2. Methods

2.1. Participants

This study included 98 participants, divided into two groups. There were 63 participants diagnosed with idiopathic PD and 35 healthy controls (HC). The group of PD speakers was composed of Belgian French native speakers ranging in age from 38 to 85 years (mean age: 70), with an average disease duration of 7 years (ranging from 1 to 25 years) and representing all stages of Parkinson's disease on the Hoehn and Yahr (1967) disability scale. All patients were diagnosed by the same neurologist following the UK Parkinson's Disease Society brain bank criteria. Of the 63 participants with PD, 43 were dysarthric (DPD) and 20 were not dysarthric (NDPD) as determined by expert perceptual assessment during a complete speech assessment (respiratory aspects, articulatory aspects, oro-linguo-facial and pneumo-phono-articulatory coordination) and with the speech item (item 3.1) of the Movement Disorders Society-Unified Parkinson's Disease Rating Scale part III/MDS-UPDRS (Goetz et al., 2008). All patients were evaluated by the same speech therapist during a speech assessment. The neurologist and the speech therapist are both specialized in the assessment and management of individuals with Parkinson's disease. Both work in a day hospital department dedicated to individuals with PD.

Table 1 presents the characteristics of participants with PD in terms of sex, stage of disease (referring to Hoehn & Yahr stages), time since first diagnosis and dysarthria. It also provides scores on the original versions of UPDRS-III (motor score), Beck Depression Inventory/BDI-II (Beck et al., 1996), Montreal Cognitive Assessment/MoCA (Nasreddine et al., 2005), as well as the Parkinson's Disease Questionnaire/PDQ-39 (Auquier et al., 2002) specific to quality of life.

HC participants were aged 41 to 84 years (mean: 66) and presented nor reported any previous speech-language pathology.

2.2. Tasks

All PD patients were met in the ON (dopaminergic treatment) phase. Study participants were subjected to a variety of speech tasks, one of which was to repeat the cardinal French vowels/a, i, u/ five times. Steady oral vowels are the most easy-to-collect speech material in clinical settings. Furthermore, this production number allows for a compromise between clinical care and evaluation constraints while ensuring a sufficient number of repetitions to allow for robust statistical analysis of the collected data. Each participant thus performed fifteen isolated vowel productions. Only the results of this controlled task will be presented in this study. PD participants were assessed individually in a quiet room in the hospital and HC subjects were recorded under similar conditions, at home. The two groups were recorded with the same Zoom H5 portable recorder.

2.3. Acoustic measurements and acoustic metrics

Acoustic measurements were performed using Praat formant tracking and customized Praat scripts. The F1 and F2 values were obtained through a semi-automatic procedure from the steady state portion of each vowel. Specifically, the stable part of each vowel was manually identified based on information from the speech waveform and spectrogram, excluding unstable phases characterized by creaky voice, voicing interruption, breathing resumption, etc. The formant frequencies were automatically detected and manually verified, and their average value over the whole stable part was calculated.

Three different acoustic metrics were computed from the vowels produced by each speaker:

- The triangular Vowel Space Area (tVSA, in Hz^2), which gives the size of the working vowel space for each participant (e.g., Kent and Vorperian, 2018). The tVSA is calculated using the formula:

$$\text{tVSA} = \left|0.5 \times \left[(\text{F2u} + \text{F2i}) \times (\text{F1u} - \text{F1i}) - (\text{F2a} + \text{F2u}) \times (\text{F1u} - \text{F1a}) - (\text{F2a} + \text{F2i}) \times (\text{F1a} - \text{F1i})\right]\right|$$

The higher the tVSA, the larger the participant's vowel space.

- The Vowel Articulation Index (VAI), which concerns the tendency for vowel centralization, was developed by Sapir et al. (2010, 2011) to account for inter-speaker variability. According to these authors, since the measure of maximum vowel space is sensitive to inter-individual variability, the

TABLE 1 Characteristics of participants with PD in terms of sex, stage of disease, time since first diagnosis, dysarthria and scores of UPDRS-III, BDI-II, MoCA and PDQ-39.

Sexe	Stage (Hoehn and Yahr)	Duration_PD	UPDRS_III	Dysarthria	Severity_dysarthria	BECK (cut-off mild depression: 10–18)	MoCA (cut-off detecting MCI ≤ 25/30)	PDQ_39 (QoL deteriorated >50)
M	3	6	10	No	N/A	6	24	17
F	3	7	16	No	N/A	13	27	33
F	1,5	9	5	No	N/A	12	30	12
M	0	6	3	No	N/A	3	30	11
F	2,5	2	7	No	N/A	11	24	21
M	2	8	6	No	N/A	0	28	5
F	2,5	4	12	No	N/A	2	30	17
M	2,5	7	15	No	N/A	5	28	19
M	1	5	1	No	N/A	3	29	5
M	2	11	2	No	N/A	3	28	7
M	3	6	15	No	N/A	7	29	5
M	3	5	10	No	N/A	7	29	11
F	1,5	6	2	No	N/A	1	30	15
F	1,5	2	4	No	N/A	7	28	18
F	2	2	8	No	N/A	5	28	5
F	3	19	13	No	N/A	16	27	36
M	2	3	16	No	N/A	6	28	7
M	2	7	9	No	N/A	2	30	6
M	3	6	33	No	N/A	4	26	8
F	3	2	16	No	N/A	9	25	25
F	1,5	10	6	Yes	mild	6	29	26
M	2,5	13	15	Yes	moderate	6	27	13
M	3	24	4	Yes	mild	10	30	43
M	3	7	15	Yes	mild	11	29	19
M	4	2	45	Yes	moderate	8	30	23
F	4	7	20	Yes	mild	10	29	25
M	2	7	12	Yes	mild	11	27	14
F	2,5	11	11	Yes	mild	6	23	14
M	2,5	11	10	Yes	moderate	1	30	7
M	1,5	15	9	Yes	moderate	4	27	15
F	4	9	16	Yes	moderate	9	28	42
F	4	10	26	Yes	mild	12	27	37
F	1,5	9	7	Yes	moderate	8	30	21
M	2	3	16	Yes	moderate	25	26	30
F	2,5	3	13	Yes	mild	15	25	32
M	1,5	3	8	Yes	mild	5	29	12
M	2	3	1	Yes	moderate	0	29	3
M	3	6	35	Yes	mild	6	28	20
M	2,5	6	14	Yes	moderate	9	29	32
F	2,5	8	16	Yes	mild	15	30	30
F	3	9	20	Yes	moderate	9	28	23

(Continued)

TABLE 1 (Continued)

Sexe	Stage (Hoehn and Yahr)	Duration_PD	UPDRS_III	Dysarthria	Severity_dysarthria	BECK (cut-off mild depression: 10–18)	MoCA (cut-off detecting MCI ≤ 25/30)	PDQ_39 (QoL deteriorated >50)
M	1,5	2	11	Yes	mild	1	28	0
M	1,5	12	7	Yes	mild	1	30	9
M	1,5	6	7	Yes	mild	1	30	3
M	3	4	19	Yes	moderate	6	27	37
F	3	11	16	Yes	mild	11	26	30
F	2	8	5	Yes	mild	7	28	18
M	4	4	27	Yes	mild	18	24	49
M	3	4	18	Yes	mild	7	23	19
M	2	7	12	Yes	mild	3	26	7
F	4	15	35	Yes	mild	16	17	46
F	2	7	5	Yes	mild	2	28	3
F	5	18	48	Yes	moderate	10	21	46
M	4	6	14	Yes	mild	6	29	16
M	4	6	41	Yes	moderate	12	28	60
F	3	4	18	Yes	mild	6	30	21
M	1,5	1	5	Yes	mild	2	30	3
F	2,5	2	14	Yes	mild	11	27	25
F	2,5	2	13	Yes	mild	9	25	28
M	2	5	6	Yes	mild	11	26	24
F	1,5	5	8	Yes	moderate	8	27	13
M	1	4	2	Yes	mild	7	30	23
M	4	25	43	Yes	moderate	10	24	26

VAI allows to better represent any centralization of vowel formants. The goal of this index is to minimize sensitivity to interindividual variability and maximize sensitivity to vowel centralization with respect to tVSA (Sapir et al., 2010). Caverlé and Vogel (2020), in a study in which they compared several metrics to quantify vowel production (including tVSA and VAI), suggest that VAI is the most stable and sensitive measure under fatigue and noise conditions in healthy participants. According to Skodda et al. (2012), the VAI is considered to be a more effective measure than the Triangular Vowel Space Area (tVSA) for identifying speech difficulties in individuals with PD.

The VAI is calculated using the formula:

$$\text{VAI} = \left(\text{F2i} + \text{F1a}\right) / \left(\text{F1i} + \text{F1u} + \text{F2u} + \text{F2a}\right)$$

The lower the calculated value, the higher the vowel centralization, and vice versa.

- The PHI index, which characterizes the level of organization of the vowel space, was calculated by determining the ratio between inter-category and intra-category dispersion within the vocalic system (Huet and Harmegnies, 2000). In addition to inter-category variability (e.g., variability due to vowel centralization), it can account for intra-category variability (e.g., variability due to vowel distortions). The phi index is the ratio between inter- and intra-categorical variability computed by analogy with the Fisher-Snedecor F-statistic in an analysis-of-variance model:

$$\Phi = \frac{\text{inter\_MS}}{\text{intra\_MS}}$$

$$\text{Where:}\quad \text{inter\_MS} = \frac{\text{inter} - \text{category sum of squares}}{\text{inter} - \text{category degrees of freedom}}$$

$$\text{And:}\quad \text{intra\_MS} = \frac{\text{intra} - \text{category sum of squares}}{\text{intra} - \text{category degrees of freedom}}$$

The inter-category mean square (inter_MS) is defined as the sum of the squares of the differences between the centroid of each vowel category and the general centroid of the entire vowel space, weighted by the number of vowels in each category and standardized by the total number of categories minus 1.

The intra-category mean square (intra_MS), on the other hand, is defined as the sum of the squares of the differences between each repetition of the same vowel and the centroid of the corresponding category, normalized by the number of vowels considered minus the number of categories.

Therefore, a lower PHI value suggests a lower degree of vocalic organization.

2.4. Statistical analysis

In order to assess the differences in acoustic parameters between PD patients and HC, statistical analyses were performed on all collected measurements using SPSS software (IBM SPSS Statistics 25). Because of the non-normality of the distributions non-parametric tests were chosen. Specifically, a series of Mann Whitney U tests were performed in order to make all possible pairwise comparisons between the three groups of participants.

3. Results

The demographic data (Table 1) allow us to observe a link between disease stage and motor symptoms (proportion of variance accounted η^2: 0.677) and between disease stage and quality of life (η^2: 0.468). Only a marginal fraction of the total variance was explained by the relationship between disease stages and time since first diagnosis (η^2: 0.138). Moreover, no link is found between disease progression stages and presence/absence of dysarthria (η^2: 0.047), nor between the presence/absence of dysarthria and disease duration (η^2: 0.023), depressive symptoms (η^2: 0.040), cognitive impairment (η^2: 0.012), or quality of life (η^2: 0.098).

The proportion of participants did not differ significantly in the groups either in terms of sex (Pearson chi-square test, $\chi^2 = 0.075$; $p = 0.785$) or age ($\chi^2 = 43.151$; $p = 0.298$).

3.1. Triangular vowel space area

The calculation of the triangular vowel space area (tVSA) showed that on average, the mean area was significantly smaller for DPD patients than for HC participants (U = 1,400, $p = 0.027$). The area values were significantly greater for HC participants (mean: 363679 Hz^2) compared to those in the DPD group (mean: 306501 Hz^2), except for the first repetition. Indeed, when the five iterations per vowel produced by the participants were considered separately, we found that, the first production of the phonemes /a, i, u/ had similar characteristics in both groups. The four other productions were significantly different between the two groups.

However, these differences were only found for DPD speakers compared to HC speakers. No differences were observed between the productions of NDPD and HC participants. We also observe no significant differences between the productions of DPD and NDPD participants, which is in contradiction with the distinction made by clinicians regarding the presence or absence of dysarthric symptoms in these patients. Overall, we also observe a high interindividual variability in PD speakers.

3.2. Vowel articulation index

The VAI values were found to be significantly different between DPD patients and HC speakers (U = 1,519, $p = 0.001$), indicating that the dysarthric speakers with PD had more centralized vowel productions and less contrast between vowels when compared to HC participants.

As observed from the tVSA metric, we were unable to identify differences between NDPD and HC participants from the VAI centralization index. However, unlike the results obtained from the calculation of the tVSA metric, the VAI centralization index allows us to uncover significant differences between the productions of the DPD and NDPD speakers.

3.3. Index of the level of organization of the vowel space (PHI)

Regarding the PHI index, there was no difference between the productions of DPD and NDPD speakers. However, the PHI values were found to be significantly higher for HC speakers (mean: 1477) compared to DPD patients (mean: 150) (U = 1960, $p < 0.001$). Indeed, a high level of formant centralization was observed in DPD speakers, resulting in lower inter-category differentiation than in HC speakers (U = 1,511, $p = 0.001$). Furthermore, intra-category dispersion was significantly lower in HC speakers than in the DPD group (intra_MS: mean: 8751 vs. mean: 31125; U = 278, $p < 0.001$).

The PHI metric also showed a significant difference between NDPD and HC speakers (U = 639, $p < 0.001$). This difference was primarily due to a higher intra-categorical dispersion in NDPD patients, likely resulting from larger variability in vowel production (U = 86, $p < 0.001$) (see Figure 1).

4. Discussion

The purpose of this study was to identify objective vocal biomarkers in the production of oral vowels among parkinsonian speakers. The aim was to support the clinicians in identifying subtle acoustic alterations that may be difficult to detect perceptually, in order to allow an early diagnosis of dysarthria, even when clinical symptoms are subclinical. Furthermore, the relationships between PD and dysarthria are not bidirectional: not all Parkinson's patients necessarily develop dysarthria, and the presence and severity of dysarthria can vary from one patient to another and evolve at a different pace than the progression of the disease (Dias et al., 2016; Karan et al., 2022). Moreover, the analysis of demographic data highlights a lack of correlation between the progression of the disease (disease stages and duration since diagnosis) and the presence or absence of dysarthric symptoms. Therefore, the sole progression of the disease does not appear to be a reliable indicator of the progression of dysarthria.

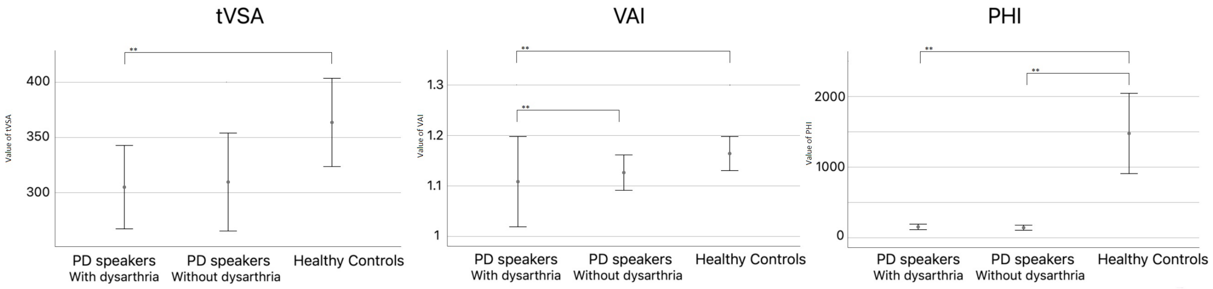


FIGURE 1
Mean tVSA [KHz^2], VAI and PHI across the three groups of participants. Error bars represent 95% confidence intervals. Significant pairwise comparisons are represented by asterisks (** 0.01 significant threshold).

Through an acoustic analysis of the productions of the vowels /a, i, u/, we computed three acoustic metrics considered as complementary because of the information they provide: information on the maximum vowel working space (tVSA), information on the accuracy during the productions (PHI, and more particularly intra_MS, the intra-category variability), information on a possible phenomenon of centralization of the vowel targets (VAI as well as inter_MS, the component of the PHI metric that reflects inter-category variability).

Using these combined metrics, the overall goal was to better identify global variations in the exploitation of the vowel system in the three groups of participants. The results demonstrate the benefits of combining several acoustic metrics to characterize the vowel system of PD speakers. First, the tVSA metric, which is the most frequently used in research on the vowel system in pathological speech, enables to uncover alterations in DPD speakers compared to HC speakers. In fact, both groups of speakers had similar tVSA values for the first repetition of the phonemes /a, i, u/, but differed significantly for the other four productions, DPD speakers exhibiting smaller vowel space areas than healthy controls. This result pattern can be interpreted as DPD speakers transiently resorting to hyperarticulation (relative to their own routines) on their first attempt to repeat the vowel. The significant differences observed on subsequent repetitions suggest that they could not maintain this strategy for the remainder of the vowel sequence.

Importantly, tVSA does not allow to distinguish NDPD speakers from parkinsonian participants with dysarthria or from healthy speakers. Thus, this metric is not sensitive enough to identify subclinical manifestations of dysarthria, supporting Skodda et al. (2011) suggestion of a low informative potential of tVSA in detecting slight changes during vowel production by PD speakers. The lack of significant differences in our study between DPD and NDPD in terms of tVSA would result from the fact that dysarthria-related alterations in steady-state vowel production are too subtle to be highlighted by tVSA calculation.

Second, the VAI centralization metric is valuable in that it reflects the categorization made by the speech therapists during speech assessment between PD patients with and without dysarthria. These findings which corroborates the perceptual distinction between the groups as formulated by the speech therapist, in accordance with Skodda et al. (2011), suggest that speech therapists may use vowel centralization as a cue of dysarthria, i.e., a form of hypoarticulation characterized by a general shift of vowel targets toward the center of the vowel space. However, this metric does not appear to be useful in searching for potential early, subtle speech alterations that might distinguish NDPD speakers from HC, which suggests that NDPD speakers produce vowels as dispersed in the vowel space as those of typical, healthy participants.

Third, the PHI metric yields very different results depending on whether participants are healthy controls or Parkinsonian participants (both with and without dysarthria). PHI values were found to be significantly lower for parkinsonian speakers which indicates that their vocalic system is substantially less organized than that of control speakers.

For DPD speakers, inter-category dispersion was reduced and intra-category variability was increased. Significantly lower inter-category dispersion is in line with higher centralization, in accordance with the results of the VAI metric for these speakers. Greater intra-category variability suggests difficulty in repeatedly producing the same vowel in the same way, which may reflect articulatory instability and/or more variable speech targets.

As to NDPD speakers, PHI was the only metric that showed a significant difference between their vocalic productions and those of healthy speakers. However, this difference was primarily due to a higher intra-categorical dispersion in NDPD patients, likely resulting from larger variability in vowel production. Therefore, what was significantly reduced among NDPD speakers was not so much the overall articulatory range/workspace (indexed by tVSA), but the internal organization of the vowel system itself due to the lack of accuracy around vowel targets.

Unlike the other two metrics, PHI accounts for intra-category variability (intra_MS) in vowel production, which appears to be substantially increased for all PD participants, even for those who have not been diagnosed with dysarthria.

In summary, following our acoustic analyses based on a diversity of metrics, we confirm in the present study the presence of potential speech biomarkers of dysarthria in NDPD. The PHI metric could be considered a potential biomarker for the early stages of dysarthria in people with PD as it is the only measure capable of detecting subtle differences in vowel production between NDPD and HC speakers,

even though it does not allow for the differentiation of DPD and NDPD speakers. Those differences reside in larger intra-categorical variability presumably due to a difficulty in reaching vowel targets with accuracy and consistency. Such alterations seem to occur in the initial stages of PD, or at least when the dysarthria is still subclinical, which is in line with recommendations for early evaluation of dysarthria in PD, so that early speech therapy can be considered.

It should be noted that the limited number of data points collected per speaker should be considered, in our opinion, not as a limitation, but as an asset of the present study. Indeed, our goal was to propose an analysis based on a procedure that could be easily integrated into the clinical practice of speech therapists. Faced with a clinical problem, the intention is to propose an early detection method for a systematic screening of hypokinetic dysarthria with a semi- automatic acoustic analyses routine. Such semi-automatic screening procedures involving manually supervised acoustic measures to be integrated into clinical practice of speech therapists are currently tested in the framework of the MonPaGe protocol so that they require no more than a few minutes of analysis per patient for the clinician, the intervention of the speech therapists required to check the automatic segmentation as well as adjusting some key parameters (Laganaro et al., 2021).

Among the limitations of the present study, the most significant one concerns the evolution of NDPD patients. A longitudinal study confirming or refuting the subsequent appearance of dysarthric symptoms would allow us to reinforce or qualify our results.

Moreover, we ensured that the relative proportions of men and women in each group were identical to ensure the relevance of comparisons between groups even though the data was not standardized. Examining the effects of normalizing formant values, as recently proposed by Kuo and Berry (2023), may be relevant to a future study.

Furthermore, the characteristics of our PD patients allow us to identify participants with mild cognitive impairment ($N = 11$). However, the results on the MoCA do not appear to be correlated with the presence/absence of dysarthria ($\eta^2 = 0.012$). A future study focusing on the effects of cognitive impairments and the progression of dysarthria in Parkinson's disease could be conducted, as speech motor control requires significant cognitive resources.

The perspectives of the present work relate to the potential value of the PHI index for the differential diagnosis of Parkinson's disease. Currently, we are conducting an acoustic analysis of spontaneous vowels produced by the same participants in a picture description task. The objective is to consolidate the findings of the present study concerning the interest of the PHI index for the detection of subtle, subclinical speech alterations in PD, i.e., even in patients without dysarthria. Next, we will recruit patients in the diagnostic phase as well as previously diagnosed patients in order to identify biomarkers that can be used to guide the diagnosis of PD vs. other related pathologies (e.g., Parkinson +, progressive supranuclear palsy, multiple system atrophy). Indeed, there are still few studies comparing the productions of these patients for differential diagnosis purposes as highlighted by Daoudi et al. (2022).

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by CHU-Charleroi and Erasme-ULB (P2015/527/B406201526528). The patients/participants provided their written informed consent to participate in this study.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Funding

This study has partially benefited from the funding attributed to the Research Project « Parolpathos », Action de Recherche Concertée AUWB- 2012-12/17-UMONS-N°1.

Acknowledgments

We thank the CHU-Charleroi for their help in recruitment.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ackermann, H., Hertrich, I., and Hehr, T. (1995). Oral diadochokinesis in neurological dysarthrias. *Folia Phoniatrica et Logoaedica.* 47, 15–23. doi: 10.1159/000266338

Ackermann, H., and Ziegler, W. (1991). Articulatory deficits in parkinsonian dysarthria: an acoustic analysis. *J. Neurol. Neurosurg. Psychiatry* 54, 1093–1098. doi: 10.1136/jnnp.54.12.1093

Agüera-Ortiz, L., García-Ramos, R., Grandas Pérez, F. J., López-Álvarez, J., Montes Rodríguez, J. M., Olazarán Rodríguez, F. J., et al. (2021). Depression in Alzheimer's disease: a Delphi consensus on etiology, risk factors, and clinical management. *Front. Psych.* 12:638651. doi: 10.3389/fpsyt.2021.638651

American Speech-Language-Hearing Association. (2004). *Preferred practice patterns for the profession of speech-language pathology [preferred practice patterns]*. Available at: https://www.asha.org/policy/pp2004-00191/ (Accessed December 22, 2022).

Audibert, N., and Fougeron, C. (2012). Distorsions de l'espace vocalique: quelles mesures? Application à la dysarthrie. Actes de la conférence conjointe JEP-TALN-RECITAL 2012, 1, 217–224. Grenoble.

Auquier, P., Sapin, C., Ziegler, M., Tison, F., Destée, A., Dubois, B., et al. (2002). Validation en langue française d'un questionnaire de qualité de vie dans la maladie de Parkinson: le Parkinson's Disease Questionnaire-PDQ-39. *Rev. Neurol.* 158, 41–50.

Beck, AT, Steer, RA, and Brown, GK. (1996). *BDI-II Manual.* London: The Psychological Corporation.

Caverlé, M. W. J., and Vogel, A. P. (2020). Stability, reliability, and sensitivity of acoustic measures of vowel space: a comparison of vowel space area, formant centralization ratio, and vowel articulation index. *J. Acoust. Soc. Am.* 148, 1436–1444. doi: 10.1121/10.0001931

Daoudi, K., Das, B., de Saint, M., Victor, S., Foubert-Samier, A., Fabbri, M., et al. (2022). A comparative study on vowel articulation in Parkinson's disease and multiple system atrophy. *Interspeech*, 2248–2252. doi: 10.21437/Interspeech.2022-845

Delvaux, V., Roland, V., Huet, K., Piccaluga, M., Haelewyck, M. C., and Harmegnies, B. (2016). The production of intervocalic glides in non Dysarthric parkinsonian speech. *Proc. Interspeech*, 253–256. doi: 10.21437/Interspeech.2016-349

Dias, A. E., Barbosa, M. T., Limongi, J. C. P., and Barbosa, E. R. (2016). Speech disorders did not correlate with age at onset of Parkinson's disease. *Arq. Neuropsiquiatr.* 74, 117–121. doi: 10.1590/0004-282X20160008

Duez, D., Ghio, A., and Viallet, F. (2020). Perception des consonnes dans la dysarthrie parkinsonienne: Effets du contexte phonémique, prosodique et lexical. 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition), 10

Duffy, J. R. (2019). *Motor speech disorders: Substrates, differential diagnosis and management* (*4th*). St Louis, MO: Mosby.

Fougeron, C., and Smith, C. L. (1999). *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet (French)*. Cambridge University Press, England.

Gentilhomme, A., Tir, M., and Renard, A. (2020). Dysarthrie dans la maladie de Parkinson. Quelle est la sévérité de la dysarthrie au moment de la première évaluation en orthophonie pour prise en soin? *Neurologies* 23:213.

Goberman, A., Coelho, C., and Robb, M. (2002). Phonatory characteristics of parkinsonian speech before and after morning medication: the ON and OFF states. *J. Commun. Disord.* 35, 217–239. doi: 10.1016/S0021-9924(01)00072-7

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., et al. (2008). Movement Disorder Society-sponsored revision of the Uni-fied Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. *Mov. Disord.* 23, 2129–2170. doi: 10.1002/mds.22340

Hartelius, L., and Svensson, P. (1994). Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey. *Folia Phoniatr. Logop.* 46, 9–17. doi: 10.1159/000266286

Hlavnička, J., Čmejla, R., Tykalová, T., Šonka, K., Růžička, E., and Rusz, J. (2017). Automated analysis of connected speech reveals early biomarkers of Parkinson's disease in patients with rapid eye movement sleep behaviour disorder. *Sci. Rep.* 7:12. doi: 10.1038/s41598-017-00047-5

Hoehn, M. M., and Yahr, M. D. (1967). Parkinsonism: onset, progression, and mortality. *Neurology* 17, 427–442. doi: 10.1212/WNL.17.5.427

Huet, K., and Harmegnies, B. (2000). Contribution à la quantification du degré d'organisation des systèmes vocaliques. *XXIIIèmes Journées d'Etude sur la Parole* 1, 225–228. Aussois

Kalf, H., de Swart, B., Bonnier-Baars, M., Kanters, J., Hofman, M., and Kocken, J., … Munneke, M. (2011). *Guidelines for speech-language therapy in parkinson's disease*. Nijmegen (The Netherlands)/Miami (USA): National Parkinson Foundation.

Karan, B., Sahu, S. S., and Orozco-Arroyave, J. R. (2022). An investigation about the relationship between dysarthria level of speech and the neurological state of Parkinson's patients. *Biocybernetics and Biomed. Eng.* 42, 710–726. doi: 10.1016/j.bbe.2022.04.003

Karlsson, F., and Hartelius, L. (2019). How well does Diadochokinetic task performance predict articulatory imprecision? Differentiating individuals with Parkinson's disease from control subjects. *Folia Phoniatr Logop* 71, 251–260. doi: 10.1159/000498851

Kent, R. D., and Kim, Y.-J. (2003). Toward an acoustic typology of motor speech disorders. *Clin. Linguist. Phon.* 17, 427–445. doi: 10.1080/0269920031000086248

Kent, R. D., and Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: a review. *J. Commun. Disord.* 74, 74–97. doi: 10.1016/j.jcomdis.2018.05.004

Kuo, C., and Berry, J. (2023). The relationship between acoustic and kinematic vowel space areas with and without normalization for speakers with and without dysarthria. *Am. J. Speech Lang. Pathol.* 1-15, 1–15. doi: 10.1044/2023_AJSLP-22-00158

Laganaro, M., Fougeron, C., Pernon, M., Levêque, N., Borel, S., Fornet, M., et al. (2021). Sensitivity and specificity of an acoustic- and perceptual-based tool for assessing motor speech disorders in French: the MonPaGe-screening protocol. *Clin. Linguist. Phon.* 35, 1060–1075. doi: 10.1080/02699206.2020.1865460

Martel-Sauvageau, V., Roy, J.-P., Cantin, L., Prud'Homme, M., Langlois, M., and Macoir, J. (2015). Articulatory changes in vowel production following STN DBS and levodopa intake in Parkinson's disease. *Parkinsons Dis* 2015, 1–7. doi: 10.1155/2015/382320

Martel-Sauvageau, V., and Tjaden, K. (2017). Vocalic transitions as markers of speech acoustic changes with STN-DBS in Parkinson's disease. *J. Commun. Disord.* 70, 1–11. doi: 10.1016/j.jcomdis.2017.10.001

McAuliffe, M. J., Ward, E. C., and Murdoch, B. E. (2006). Speech production in Parkinson's disease: I. an electropalatographic investigation of tongue-palate contact patterns. *Clin. Linguist. Phon.* 20, 1–18. doi: 10.1080/02699200400001044

McRae, P., and Tjaden, K. (1998). Spectral properties of fricatives in Parkinson's disease. *Am. J. Speech Lang. Pathol.* 104:1854. doi: 10.1121/1.424480

Miller, N., Deane, K. H. O., Jones, D., Noble, E., and Gibb, C. (2011). National survey of speech and language therapy provision for people with Parkinson's disease in the United Kingdom: therapists' practices. *Int. J. Lang. Commun. Disord.* 46, 189–201. doi: 10.3109/13682822.2010.484849

Mollaei, F., Shiller, D. M., Baum, S. R., and Gracco, V. L. (2016). Sensorimotor control of vocal pitch and formant frequencies in Parkinson's disease. *Brain Res.* 1646, 269–277. doi: 10.1016/j.brainres.2016.06.013

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal cognitive assessment, MoCA: a Brief screening tool for mild cognitive impairment: MOCA: a BRIEF SCREENING TOOL FOR MCI. *J. Am. Geriatr. Soc.* 53, 695–699. doi: 10.1111/j.1532-5415.2005.53221.x

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *J. Speech Lang. Hear. Res.* 51, 574–585. doi: 10.1044/1092-4388(2008/041)

Postuma, R. B., Berg, D., Stern, M., Poewe, W., Olanow, C. W., Oertel, W., et al. (2015). MDS clinical diagnostic criteria for Parkinson's disease: MDS-PD clinical diagnostic criteria. *Mov. Disord.* 30, 1591–1601. doi: 10.1002/mds.26424

Rountrey, C., and Molett, M. (2020). Vowel articulation index and conversational spontaneous speech intelligibility in parkinson's disease. *Parkinsonism Related Dis.* 79:e123. doi: 10.1016/j.parkreldis.2020.06.444

Roy, N., Nissen, S. L., Dromey, C., and Sapir, S. (2009). Articulatory changes in muscle tension dysphonia: evidence of vowel space expansion following manual circumlaryngeal therapy. *J. Commun. Disord.* 42, 124–135. doi: 10.1016/j.jcomdis.2008.10.001

Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., et al. (2013). Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. *J. Acoust. Soc. Am.* 134, 2171–2181. doi: 10.1121/1.4816541

Rusz, J., Hlavnička, J., Novotný, M., Tykalová, T., Pelletier, A., Montplaisir, J., et al. (2021). Speech biomarkers in rapid eye movement sleep behavior disorder and Parkinson disease. *Ann. Neurol.* 90, 62–75. doi: 10.1002/ana.26085

Sapir, S. (2014). Multiple factors are involved in the dysarthria associated with Parkinson's disease: a review with implications for clinical practice and research. *J. Speech Lang. Hear. Res.* 57, 1330–1343. doi: 10.1044/2014_JSLHR-S-13-0039

Sapir, S., Ramig, L. O., Spielman, J. L., and Fox, C. (2010). Formant centralization ratio: a proposal for a new acoustic measure of dysarthric speech. *J. Speech. Lang. Hear. Res.* 53:114. doi: 10.1044/1092-4388(2009/08-0184)

Sapir, S., Ramig, L. O., Spielman, J., and Fox, C. (2011). "Acoustic metrics of vowel articulation in parkinson's disease: Vowel space area (VSA) vs. vowel articulation index (VAI)," in *Models of Analysis of Vocal Emissions for Biomedical Applications: 7th International Workshop*. Florence, Italy: Firenze University Press, 173–5.

Sapir, S., Spielman, J. L., Ramig, L. O., Story, B. H., and Fox, C. (2007). Effects of intensive voice treatment (the Lee Silverman voice treatment [LSVT]) on vowel articulation in Dysarthric individuals with idiopathic Parkinson disease: acoustic and perceptual findings. *J. Speech Lang. Hear. Res.* 50, 899–912. doi: 10.1044/1092-4388(2007/064)

Schalling, E., Johansson, K., and Hartelius, L. (2017). Speech and communication changes reported by people with Parkinson's disease. *Folia Phoniatr. Logop.* 69, 131–141. doi: 10.1159/000479927

Skodda, S., Grönheit, W., and Schlegel, U. (2012). Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease. *PLoS One* 7:e32132. doi: 10.1371/journal.pone.0032132

Skodda, S., Visser, W., and Schlegel, U. (2011). Vowel articulation in Parkinson's disease. *J. Voice* 25, 467–472. doi: 10.1016/j.jvoice.2010.01.009

Sunwoo, M. K., Hong, J. Y., Lee, J. E., Lee, H. S., Lee, P. H., and Sohn, Y. H. (2014). Depression and voice handicap in Parkinson disease. *J. Neurol. Sci.* 346, 112–115. doi: 10.1016/j.jns.2014.08.003

Walsh, B., and Smith, A. (2012). Basic parameters of articulatory movements and acoustics in individuals with Parkinson's disease. *Mov. Disord.* 27, 843–850. doi: 10.1002/mds.24888

Wong, M. N., Murdoch, B. E., and Whelan, B.-M. (2011). Lingual kinematics in Dysarthric and Nondysarthric speakers with Parkinson's disease. *Parkinson's Dis* 2011, 1–8. doi: 10.4061/2011/352838

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 30 August 2023
DOI 10.3389/fpsyg.2023.1167902

Exploring the acoustic and prosodic features of a lung-function-sensitive repeated-word speech articulation test

Biao Zeng[1]*, Edgar Mark Williams[2†], Chelsea Owen[1], Cong Zhang[3], Shakiela Khanam Davies[1], Keira Evans[1] and Savannah-Rose Preudhomme[2]

[1]Department of Psychology, University of South Wales, Pontypridd, United Kingdom, [2]School of Care Sciences, University of South Wales, Pontypridd, United Kingdom, [3]School of Education, Communication and Language Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom

Introduction: Speech breathing is a term usually used to refer to the manner in which expired air and lung mechanics are utilized for the production of the airflow necessary for phonation. Neurologically, speech breathing overrides the normal rhythms of alveolar ventilation. Speech breathing is generated using the diaphragm, glottis, and tongue. The glottis is the opening between the vocal folds in the larynx; it is the primary valve between the lungs and the mouth, and by varying its degree of opening, the sound can be varied. The use of voice as an indicator of health has been widely reported. Chronic obstructive pulmonary disease (COPD) is the most common long-term respiratory disease. The main symptoms of COPD are increasing breathlessness, a persistent chesty cough with phlegm, frequent chest infections, and persistent wheezing. There is no cure for COPD, and it is one of the leading causes of death worldwide. The principal cause of COPD is tobacco smoking, and estimates indicate that COPD will become the third leading cause of death worldwide by 2030. The long-term aim of this research program is to understand how speech generation, breathing, and lung function are linked in people with chronic respiratory diseases such as COPD.

Methods: This pilot study was designed to test an articulatory speech task that uses a single word ("helicopter"), repeated multiple times, to challenge speech-generated breathing and breathlessness. Specifically, a single-word articulation task was used to challenge respiratory system endurance in people with healthy lungs by asking participants to rapidly repeat the word "helicopter" for three 20-s runs interspersed with two 20-s rest periods of silent relaxed breathing. Acoustic and prosodic features were then extracted from the audio recordings of each adult participant.

Results and discussion: The pause ratio increased from the first run to the third, representing an increasing demand for breath. These data show that the repeated articulation task challenges speech articulation in a quantifiable manner, which may prove useful in defining respiratory ill-health.

KEYWORDS

speech breathing, COPD, respiration, pause, helicopter task

Check for updates

OPEN ACCESS

EDITED BY
Kevin Tang,
Heinrich Heine University of Düsseldorf, Germany

REVIEWED BY
Fangfang Li,
University of Lethbridge, Canada
Anthony P. Salvatore,
University of Louisiana at Lafayette, United States
Vikram N. Dayalu,
Seton Hall University, United States

*CORRESPONDENCE
Biao Zeng
✉ biao.zeng@southwales.ac.uk

†Deceased

RECEIVED 16 February 2023
ACCEPTED 03 August 2023
PUBLISHED 30 August 2023

CITATION
Zeng B, Williams EM, Owen C, Zhang C, Davies SK, Evans K and Preudhomme S-R (2023) Exploring the acoustic and prosodic features of a lung-function-sensitive repeated-word speech articulation test. *Front. Psychol.* 14:1167902.
doi: 10.3389/fpsyg.2023.1167902

1. Introduction

Chronic obstructive pulmonary disease (COPD) is the most common long-term respiratory disease. The main symptoms of COPD are increasing breathlessness, a persistent chesty cough with phlegm, frequent chest infections, and persistent wheezing.[1] There is no cure for COPD, and it is one of the leading causes of death in the world. The principal cause of COPD is tobacco smoking, and estimates indicate that it will become the third leading cause of death worldwide by 2030. This pilot study tested an articulatory speech, which uses a single word "helicopter," repeated multiple times, to challenge speech-generated breathing and breathlessness.

1.1. COPD and lung function

COPD develops slowly, appearing in middle age; initially, it has little effect on lung function, and its impact on lifestyle is minor. As the disease develops, the associated lung dysfunction becomes disabling; the person with COPD becomes increasingly immobile and eventually requires oxygen support. This slow decline can be marked by exacerbations that require acute healthcare intervention. The ability to identify these exacerbations before they occur or early on would improve the quality of life of those with COPD and help reduce healthcare costs.

Severe COPD leads to pronounced breathlessness and alters pulmonary ventilation. These COPD-induced changes subtly affect other breathing-related functions, such as speech articulation and the pause time between words.

A clinical test used to diagnose and stage the severity of COPD is spirometry, which is performed under the supervision of a trained healthcare practitioner. Although speech production is altered by COPD and other lung diseases, signals from speech production have not been used as a diagnostic tool, partially because the changes in voice are subtle.

1.2. Speech breathing

The term "speech breathing" is usually used when referring to the manner in which expired air and respiratory mechanics are utilized to produce the airflow necessary for phonation. During speech breathing, a quick inspiration is followed by a prolonged expiration. Quick inspiration can reduce pause time and allows a speaker to retain the floor in a speaking exchange. A volume of air is taken into the lungs and then pushed out through the glottis to enable utterance of speech sounds. The variable amounts of air inhaled are based on the content to be produced.

The cycle of inspiration and expiration in speech breathing is generated using the abdominal muscles, diaphragm, glottis, mouth, and nose. The abdominal muscles (rectus abdominus, external oblique, internal oblique, and transverse abdominus) are located between the ribs and the pelvis on the front of the body. These muscles support the rib cage to expand during inspiration (Hixon et al., 1973). The diaphragm is a large, dome-shaped muscle located at the base of the lungs. When the diaphragm contracts and flattens and the chest cavity enlarges, this contraction creates a vacuum and pulls air into the lungs. Upon exhalation, the diaphragm relaxes and returns to its domelike shape, and air is forced out of the lungs. The abdominal muscles can move the diaphragm and provide more power to empty the lungs. The glottis is the opening between the vocal folds in the larynx; it is the primary valve between the lungs and the mouth, and the sound can be varied by varying its degree of opening.

Usually, speech begins once the lungs have been filled upon the end of inspiration. It therefore begins with a large lung volume (LV), which is associated with longer voice onset times, increased subglottal pressure, increased sound pressure levels, a higher fundamental sound frequency, and increased glottal leakage. In contrast, speech produced at low LVs has been found to be associated with a more adducted vocal state compared with speech produced at high LVs. For instance, Iwarsson et al. (1998) studied the effects of lung volume on the glottal voice source and found that the closed quotient increases with decreasing lung volume, while subglottal pressure, peak-to-peak flow amplitude, and glottal leakage tend to decrease. In addition, Murray et al. (2018) asked speakers to read passages with two speaking voices: typical (baseline and return phases) and breathy vocals (experimental phase). They found that the participants spoke with larger LV excursions during the experimental phase, characterized by increased LV initiation and decreased LV termination compared with the baseline phase.

Regarding the airstream mechanism of speech breathing, many studies have explored the possibility of using speech breathing to predict and diagnose lung function. For specific groups, e.g., patients with asthma or Parkinson's disease, speech measures offer promising monitoring and diagnosis methods. Tayler et al. (2015) reported that healthcare professionals can estimate the predicted forced expiratory volume in one second (FEV_1 %) based on speech samples from asthma patients. This finding provides evidence that speech is altered in acute asthma.

1.3. Effects of age and sex on speech breathing

Across an individual's lifespan, the anatomy of the respiratory system changes, and the functioning of breathing can become limited in association with these changes. For instance, larger bodies typically result in larger lungs and respiratory systems (McDowell et al., 2008). Until age 14, lung length and width both expand linearly (Polgar and Weng, 1979; Zeman and Bennett, 2006). Boys' lungs continue to grow until between the ages of 18 and 20, while girls' lung growth patterns settle at ∼14 years of age (Polgar and Weng, 1979). However, men and women typically start to lose weight beyond the age of 60, and this loss continues into the seventh and eighth decades of life. Lung volume, static recoil pressure, and respiratory muscle strength all undergo physiological changes because of the respiratory system. Along with anatomical and physiological changes in the respiratory system, speech breathing patterns and features change across the lifespan.

Vocal intensity is an acoustic measure. During everyday speech production, speakers have to raise their vocal intensity to ensure

[1] https://www.nhs.uk/conditions/chronic-obstructive-pulmonary-disease-copd/

that they are heard in noisy environments. To increase vocal intensity, the respiratory system will generate higher subglottal air pressures (Finnegan et al., 2000). There are differences between age groups in terms of the way in which intensity is increased during speech production: children, teenagers, and young adult speakers can use larger lung and rib cage volume excursion to increase intensity, but older adults do not show the same pattern. Utterance length is a measure of the linguistic (prosodic) feature of speech, and there is a significant correlation between utterance length and respiratory function. Utterance length is defined as the number of syllables or words produced in one speech breath.

Speech breathing patterns change with age, but little consideration has been given to sex-based differences within COPD clinical research (Somayaji and Chalmers, 2022). There is, however, emerging and considerable evidence to suggest that sex contributes to disease pathogenesis, risk, diagnosis, prevalence, severity, and clinical outcomes (Fletcher and Peto, 1977; Doyal, 2001; Carey et al., 2007; Townsend et al., 2012). In addition, there have been calls for researchers to better understand the mechanisms underpinning these observed differences (Silveyra et al., 2021). Despite these calls, the literature is scant, but it points to anatomical differences between sexes and the influence of sex hormones, the menstrual cycle, and other diseases (e.g., asthma), which are said to modulate these sex differences in COPD (LoMauro and Aliverti, 2018). For example, standard morphometric measures have shown that males have larger lungs than females (Thurlbeck, 1982). In addition, females have smaller airway diameters and lung volume, resulting in lower peak expiratory flow than males. Furthermore, respiratory symptoms in females (e.g., wheezing, dyspnea, and cough) vary significantly with menstrual cycle-induced hormonal changes: specifically, these COPD symptoms tend to get worse in the mid-phase of the cycle (Macsali et al., 2012). Understanding the contribution of sex and gender to COPD will help with the development of precision medicine and the effective daily management of COPD.

1.4. Three-tier feature measures

Speech breathing is a special breathing function. It is a multi-faceted phenomenon integrating breathing, speech production, and articulation. Distinct types of information can be extracted from the utterance of speech sounds by examining the characteristics of speech, e.g., acoustic features, prosodic features, and certainly breathing-related features. Therefore, we developed a systematic analysis method with three tiers of feature measures. The three tiers consist of acoustics, prosody, and breathing. Acoustics refers to the physical properties of sounds, and this measure captures speech-related information, e.g., vowel formants, intensity (perceived as loudness), or fundamental frequency (perceived as pitch). Prosodic features, in this study, refers to how speech sounds are organized, including length of run and pause ratio. Measures of breathing features, specifically in relation to speech breathing, are under development. People usually take around 10–15 breaths per minute when resting. This is described as the respiratory rate. In the current study, we adopted respiratory rate as the key measure of breathing.

Vocal intensity is the most widely investigated acoustic feature in studies of speech breathing. Speakers use larger lung and rib cage volume excursions when increasing their vocal intensity (Stathopoulos and Sapienza, 1997). Further studies on prosodic information have revealed the correlation of these measures with respiratory functions. For instance, studies have revealed a correlation between utterance length and respiratory function (e.g., Sperry and Klich, 1992; Whalen and Kinsella-Shaw, 1997). Age-related effects also occur, with older adults producing shorter utterances than young adults do. Huber (2008) examined age-related changes in speech breathing by measuring utterance length and loudness, and found that age-related effects increased as utterances became longer. These results suggest that older adults have a more challenging time when the speech system is being taxed by both utterance length and loudness. The data were also consistent with the hypothesis that both young and older adults use utterance length in premotor speech planning processes.

A wide range of speech features have been judged relevant for and investigated in relation to health status. Farrús et al. (2021) proposed two types of speech features, acoustic and prosodic information, and applied them for the detection and classification of bipolar disorder. They argued that prosodic information, which is conveyed through intonation, stress, and rhyme, could reflect the emotional aspects of the individual. In this study, we investigated acoustic and prosodic features and focused on the prosodic information.

In recent years, computerized deep learning methods have offered new ways of modeling speech and analyzing it for healthcare applications (Cummins et al., 2018). For instance, Nallanthighal et al. (2021) proposed using deep learning to investigate breathing. Breathing (i.e., inhalation and expiration) is essential and these are the primary mechanisms driving speech production. These authors explored techniques for sensing the breathing cycle and extracting breathing metrics from speech using deep learning architectures, and addressed the challenges involved in establishing the usefulness of applying this technology. Estimating breathing patterns from speech provides information about the corresponding respiratory parameters, which would enable assessment of the speaker's respiratory health using speech alone.

Specifically, in the present study, pause ratio was measured as one key feature of rhythm. Fuchs and Rochet-Capellan (2021) reviewed the respiratory foundations of spoken language and highlighted the fact that breathing interacts with respiration, syntax, and planning. We should distinguish respiratory and linguistic pauses in a breathing cycle. A typical respiratory pause occurs during a breathing cycle. In a normal breathing cycle at rest, there is an in-breath (inhalation) followed by an out-breath (exhalation). The out-breath is followed by an automatic pause (or period of no breathing) lasting ∼1 to 2 s. In contrast, a linguistic pause is a silent pause or filled pause containing *um* or *uh*. In terms of speech breathing, an in-breath may play the role of such a linguistic pause, which can inspire speech and empower the following articulation. In the following section, we introduce the potential use of prosodic information as a measure of lung function and analyse the potential correlation between prosodic features and lung function.

In the present study, a speech breathing task, namely the "helicopter task," was designed to measure the acoustic, prosodic, and breathing characteristics of speech. The helicopter task requires participants to repeat the word "helicopter" as quickly as possible for 20 s, followed by a 20-s break of silence. This was repeated twice, creating a task consisting of three runs lasting ∼100 s in total. Based on previous studies, acoustic features include a wide range of parameters, e.g., frequency and vocal intensity. Vocal intensity is of specific interest. Utterance length is the key umbrella concept of prosodic features; in this study, it was calculated in terms of speech rate and word duration. A pause, as a critical parameter for measurement of speech breathing, was defined as a silence filled with no utterance of the word "helicopter" and measured in terms of pause ratio. The pause ratio was calculated as the duration of the pause divided by the entire 20-s duration of word repetition.

Three main research hypotheses were addressed:

1. Run effect: It was assumed that, with airflow consumed over the course of the task, prosodic measures would be affected: in particular, pauses would become longer, and, correspondingly, pause ratio would increase and more breaths would be taken in the later runs.
2. Sex differences: It was predicted that female speakers would produce higher-frequency speech, lower-intensity speech, shorter syllable durations, and longer pause ratios compared to male speakers.
3. As the three tiers of features included a wide range of variables in the study, we predicted that acoustic and prosodic features would be significant predictors of measures of lung function. A multiple regression method was employed to explore these predictors.

2. Materials and methods

2.1. Participants

A total of 27 healthy, native English-speaking participants (12 men, 15 women; mean age: 26, range 19–55 years; height: 1.68 ± 0.12 m; weight: 68.8 ± 14.0 kg, $n = 24$) were recruited from the University of South Wales community through random sampling. Two participants did not follow the instructions, and their data were not included in the analysis. All participants filled out the Clinical Report Form One (Appendix 1), which consisted of seven questions: age, sex, height, weight, respiratory condition, smoking history, and breathing status. No other general health status, medication, or physical activity parameters were investigated.

Predicted lung function was calculated using the Global Lung Function Initiative index (European Respiratory Society). Two measures of forced expired volume (FEV1) and forced vital capacity (FVC) were predicted from weight and height data. A t-test showed that the means of FEV1, FVC, and FEV1/FVC ratio were statistically different between sexes (p-values < 0.01).

The study was approved by the Faculty of Life Science and Education Ethics Panel, University of South Wales (No 210901HR), in accordance with the Declaration of Helsinki. Written and verbal informed consent were obtained from each participant.

TABLE 1 Phonemes in ten common words used in speech therapy tests.

Word	Number of syllables	IPA (International Phonetic Alphabet)
Ambulance	3	/ˈambjƱl(ə)ns/
Hippopotamus	5	/ḩIpə'pɐtəməs/
Computer	3	/kəmˈpjuːtə/
Spaghetti	3	/spəˈgɛti/
Vegetables	3	/ˈvɛdʒtəb(ə)l/z//
Helicopter	4	/hɛlIkɐptə/
Animal	3	/ˈanIm(ə)l/
Caravan	3	/ˈkarəvan/
Caterpillar	4	/katəpIlə/
Butterfly	3	/ˈbʌtəfl/ai//

2.2. Protocol

The study used a speech articulation task designed to test lung health. The design was derived from that of the diadochokinesis (DDK) task, which is one of the oldest and most frequently used tasks for evaluating various types of speech communication problems. It often involves fast repetition of single words or of non-speech oral movements such as opening and closing of the lips. It has numerous variations and is also referred to as verbal, oral, or phonoarticulatory DDK. It has cross-disciplinary applications in areas such as aging, biomedical engineering, biological sciences, communication sciences and disorders, computational methods in biomedicine, craniofacial surgery, dentistry, neurology and neurosciences, and oral surgery (Kent et al., 2022).

In the current study, we used a task involving the repetition of a single polysyllabic word to explore its potential in measuring lung function. This word "helicopter" was chosen from a list of words commonly used in speech therapy testing. This list is shown in Table 1. In this list, the words "ambulance," "vegetable," and "animals" can be pronounced either with a mid-central vowel, the schwa /ə/, or without. The uncertainty of the presence of the schwa could result in changes in syllable structure, duration, loudness, and other aspects of articulation. In order to keep all measurements consistent and accurate, these words were not chosen for the current task. Among the rest of the words, "hippopotamus" and "helicopter" contain the voiceless glottal fricative /h/, which requires the maximum amount of airflow to maintain articulation compared with the consonants in the other words, which are mostly stops, nasals, or approximants. Between the two words containing /h/, "hippopotamus" is a low-frequency word and may prevent speakers from articulating fluently. Therefore, "helicopter" was chosen as the word for our task.

2.3. Procedures

Each participant was asked to sit still in a quiet room at a distance of ∼40 cm from their computer. Their task was to produce the word “helicopter” as quickly as possible and as many times as possible within a 20-s production period. Each participant was given three 20-s production periods with a 20-s forced break between the first and second production periods and between the second and third production periods. The experimenter gave a signal for the start of each production period and the start of each break. The instructions were as follows:

> *“For this exercise, you will be asked to repeat the word ‘helicopter,’ as before, but this time for only 20 s at a time. For this exercise, I want you to do this three times but with a short break in between. When I say ‘go,’ please start repeating ‘helicopter’ until I say ‘stop.’ After a short interval I will say ‘go,’ so like before keep going until I say ‘stop.’ After another short rest, I will ask you to do this one more time. I will time and record the whole exercise, so please sit quietly during the two resting intervals.”*
>
> *“Do you have any questions you would like to ask?”*
>
> *“Are you ready to begin?”*

The recordings were made online via Microsoft Teams, with both the experimenter and the participant in the session. During testing, the camera was switched off and only speech was recorded. Only participants’ speech was analyzed.

The audio recordings were converted into.wav files using the audio editing software package Audacity. The three 20-s runs were extracted as separate files for acoustic analysis. Only instances of the word produced in full were kept for analysis; instances with disfluencies, prominent background noise, or overlap with the experimenter’s instructions were excluded from acoustic analysis. Audible breaths (inspirations) in each run were identified and matched to the digital recording, allowing the number of breaths to be counted and the duration of each breath to be calculated. Word boundaries and pause boundaries were manually segmented in all audio files in Praat (Boersma and Weenink, 2022) by a first annotator; a second annotator, who was a trained phonetician, then checked the boundaries and made corrections to the annotation.

For the acoustic analysis, F0 and formants were both extracted as mean values across each word as a whole, using a Praat script. F0 data were extracted with a range of 60–500 Hz, and the formants were extracted with a ceiling of 5,000 Hz. Mean F1 and mean F2 values were measured with a ceiling of 5,000 Hz. Intensity data were extracted using a minimum F0 of 60 Hz, with a 0.01 step. The length of pauses was measured, including only periods when the speaker was taking a breath between words; the initial and final silence periods at the start and end of each production period were excluded. The pause ratio was the total duration of the pauses in the production period divided by the total duration of the production period from the production of the first instance.

2.4. Statistical analysis

Means ± SD are reported. Acoustic features (e.g., intensity, F0, F1, F2, and F0 range) and prosodic features, e.g., speech rate (number of words per second for the entire run), word duration (mean duration for the word “helicopter” across the entire run), and pause ratio (mean pause duration for the entire run), are reported in Table 2. As the lung growth pattern stabilizes at 13–14 years of age for females and 18–20 years for males (Polgar and Weng, 1979), age is not a factor that we aimed to investigate in the study. Instead, the two independent variables in this study were sex and run (the repetition order in the task). A two-way repeated measure ANOVA (analysis of variance) was conducted to test for the effects of these variables on acoustic, prosodic, and breathing measures. If there was an effect of run on any feature, a regression was conducted to predict the effect on the lung function measures.

TABLE 2 Comparison of voice characteristics between male and female speakers.

Mean		Female		Male	
		Mean	SD	Mean	SD
Height (cm)		161.00	9.18	176.60	8.90
Weight (kg)		58.67	8.72	79.80	11.41
Predicted FEV1 (liters)		3.31	0.46	4.53	0.36
Predicted FVC (liters)		3.83	0.58	5.45	0.40
FEV1/FVC ratio		0.88	0.02	0.84	0.03
Intensity (dB)	Run 1	66.52	6.65	68.70	4.50
	Run 2	67.87	6.48	69.23	3.71
	Run 3	67.64	6.20	68.47	3.73
F0 (Hz)	Run 1	194.39	28.22	133.27	10.85
	Run 2	195.94	37.01	137.84	15.39
	Run 3	198.39	30.76	141.76	8.36
F1 (Hz)	Run 1	762.62	81.90	722.10	73.01
	Run 2	752.83	89.48	726.87	81.66
	Run 3	769.09	88.08	754.37	89.28
F2 (Hz)	Run 1	1868.28	148.26	1826.54	174.60
	Run 2	1868.10	123.51	1843.20	151.93
	Run 3	1887.05	141.15	1865.62	133.49
F0 range (Hz)	Run 1	89.08	47.66	154.49	106.14
	Run 2	80.42	41.09	187.64	98.44
	Run 3	101.17	44.23	191.69	82.15
Speech rate (number of words per second)	Run 1	1.47	0.20	1.72	0.22
	Run 2	1.53	0.19	1.69	0.26
	Run 3	1.47	0.17	1.67	0.26
Word duration (milliseconds)	Run 1	164	21	140	19
	Run 2	157	18	139	22
	Run 3	159	18	142	20
Pause ratio (%)	Run 1	6.5	3.1	5.4	3.7
	Run 2	5.9	2.8	8.9	3.7
	Run 3	8.0	2.3	9.1	4.0

3. Results

3.1. Voice characteristics: acoustic and prosodic features

Among the acoustic features, mains effect of run were found on intensity and F1. Specifically, a two-way repeated measures ANOVA showed a significant main effect of run on intensity, F (2.38) = 5.35, $p = 0.009$, $\eta_p^2 = 0.220$. Pairwise comparison showed that the intensity of the first run (Mean = 67.61, SE = 1.34) was significantly lower than that of the second run (Mean = 68.55, SE = 1.26), $p < 0.001$. It is intriguing that a two-way repeated measures ANOVA also showed a significant main effect of run on F1, F (2.38) = 5.53, $p = 0.008$, $\eta_p^2 = 0.225$. Pairwise comparison showed that F1 was significantly higher for the third run (Mean = 761.73, SE = 19.90) than for the first run (Mean = 742.36, SE = 17.69, $p = 0.018$) and the second run (Mean = 739.85, SE = 19.48, $p = 0.006$), but there was no difference between the first and second runs in terms of F1.

Main effects of sex on F0 and F0 range were observed. Specifically, a two-way repeated measures ANOVA showed a significant main effect of sex on F0, F (1.19) = 25.21, $p < 0.001$, $\eta_p^2 = 0.570$. Pairwise comparison showed that F0 was significantly higher in female speakers (Mean = 196.24, SE = 7.21) than in male speakers (Mean = 137.62, SE = 9.19), which is consistent with previous studies. A two-way repeated measures ANOVA also showed a significant main effect of sex on F0 range, F (1.18) = 9.49, $p = 0.006$, $\eta_p^2 = 0.345$. Pairwise comparison showed that F0 range was significantly wider in male speakers (Mean = 177.94, SE = 22.96) than in female speakers (Mean = 90.22, SE = 16.85).

With regard to prosodic features, there was a significant main effect of run on pause ratio, F (2.36) = 5.26, $p = 0.010$, $\eta_p^2 = 0.226$. Pairwise comparison showed that the third run was associated with the highest pause ratio (Mean = 0.085, SE = 0.007), and that this was significantly higher than that of the first run (Mean = 0.059, SE = 0.008, $p = 0.014$), but not the second run. There was no significant difference between the second and third runs.

A main effect of sex on speech rate was also observed, F (1.19) = 4.85, $p < 0.001$, $\eta_p^2 = 0.203$. The interaction between sex and run was significant, F (2.38) = 3.83, $p = 0.031$, $\eta_p^2 = 0.168$. Pairwise comparison showed that, in the first run only, male speakers articulated the word "helicopter" more quickly than female speakers (men: Mean = 1.72, SE = 0.07 vs. women: Mean = 1.47, SE = 0.06, $p = 0.012$).

A multiple regression analysis was conducted to explore which features contribute significantly to lung function, specifically FEV1 and FVC. Based on the ANOVA results, we selected intensity, F1, and pause ratio for inclusion as predictor variables. The two criterion variables were FEV1 and FVC. A multiple regression with backward elimination was conducted in SPSS for each run. Using backward elimination, we attempted to only include significant predictor variables in the regression model.

The regression results showed that the regression model was significant for the criterion variable of FEV1 only in the case of the second run, F (1.16) = 6.23, p = 0.025; pause ratio was the only significant predictor included, adjusted R-squared = 0.246, Beta = 0.542, $p = 0.025$. For the first and third runs, none of the predictors tested (intensity, F1, and pause ratio) was a significant predictor. A similar pattern occurred for FVC. A significant regression model only emerged in the case of the second run, F (1.16) = 6.61, $p = 0.021$, where pause ratio was the only significant predictor, adjusted R-squared = 0.26, Beta = 0.553, $p = 0.021$.

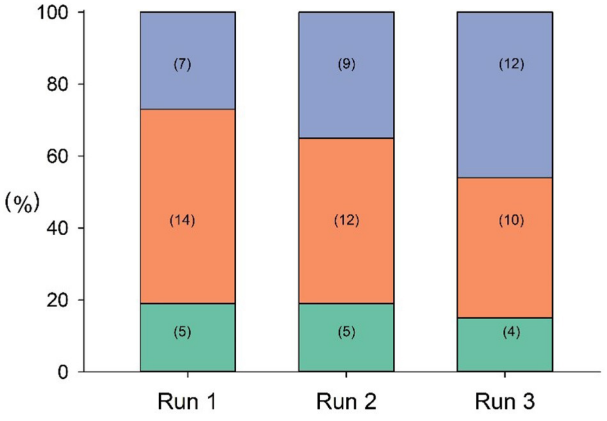


FIGURE 1
Frequency of inspiration in each run (green: no inspirations detected; orange: 1, 2 inspirations; purple: 3+ inspirations). Number of cases is shown in brackets.

3.2. Breathing characteristics

All three runs for each participant ($n = 26$) were analyzed (a total of 168 breathes, split between men and women at a ratio of 85:83). For four participants, no inspirations were recorded.

The mean duration of the inspirations taken while completing the three "helicopter" runs was 0.283 ± 0.161 s (range: 0.04–0.88 s), $n = 168$, with no differences observed between the sexes. There were also no statistically significant differences ($p = 0.149$) between runs in terms of duration or number of inspirations. The occurrence of the first inspiration was significantly earlier in the second run compared to the first (Run 1: 59 ± 19%; Run 2: 43 ± 15 %; Run 3: 49 ± 22 %).

In the first run, approximately half (14/26) of the participants took 1–2 inspirations, but by the third run, this had risen to more than 3 inspirations (12/26) (Figure 1).

4. Discussion

In summary, the effects of both sex and run number on different acoustic features, prosodic features, and breathing measures were investigated in this study. Effects of sex occurred for F0, F0 range, and speech rate. Female speakers showed higher F0 values than male speakers, which is consistent with previous studies. Male speakers articulated the word "helicopter" more quickly than female speakers, but only during the first run, which might suggest that their speech rate decreased as articulation load increased. However, the current study also found that male speakers showed a broader F0 range than female speakers. This result differs from those of some previous studies, and further investigation should be considered.

Effects of run number on intensity, F1, and pause ratio were also observed in the present study. As more runs were taken, intensity and F1 increased among male and female speakers. Correspondingly, as intensity and F1 increased with each run, the speakers took more inspirations, which was indicated by the increase in pause ratio over the course of the three runs. The speakers needed to take more pauses to inhale air as they articulated the words in the later (second and third) runs. The mechanism underlying the effect of run on intensity needs further investigation. Huber et al. (2005) proposed three means of increasing speech loudness: increased recoil pressures, increased expiratory tension, and a combination of both. Unlike their design, in which the speaker was requested to increase loudness, the current "helicopter" task is deliberately designed to induce respiratory load by requiring participants to repeat words quickly over the course of three runs. We need to clarify which approach was adopted by speakers to increase intensity and F1. In addition, we need to clarify whether the increase in air intake was used to increase intensity or to compensate for respiratory load or physiological fatigue.

The multiple regression results indicated that pause ratio was the sole significant variable to predict two lung function measures, FEV1 and FVC, and did so only on the second run. The analysis of breathing characteristics showed that the first inspiration occurred earliest in the second run, among all three runs. The correlation of pause ratio and inspiration indicates that the relationship between run number and inspiration in speech breathing might not be linear. Essentially, to understand this finding, the mechanism underlying the use of pauses in speech breathing should be interpreted in relation to whether it represents an inhaling process or is mixed with another breathing event, for instance, breathiness.

In the present study, three tiers of features have been proposed to extract and categorize rich information from speech breathing and to provide insight into the relevance of factors on each of these tiers to lung function. In the current data analysis, for instance, the weights and roles of acoustic, prosodic, and breathing features are one issue that calls for more work. Even within the category of acoustic features, the changes in various features and measures that could be attributed to speech breathing patterns or loads are not clear. Previous studies have investigated the roles of F0, F1, and F2. Lively et al. (1993) found that, in a workload condition, talkers produced utterances with increased amplitude and amplitude variability, decreased spectral tilt and F0 variability, and increased speaking rate. However, no changes in F1, F2, or F3 were observed across conditions. In contrast, Huttunen et al. (2011) studied the utterances of 13 male military pilots that were recorded during simulated combat flights, and found that the strongest associations were observed between three types of cognitive load and F1 and F2 changes in back vowels.

The present study provides empirical evidence for the use of acoustic and prosodic features of speech as health sensors and indicators. Specifically, the repeated articulation "helicopter" task and the pause ratio measure are sensitive to changes in speech breathing and reflect lung function. Within the range of other available acoustic and prosodic features, we need to further screen for sensitive and specific indicators and investigate their mechanistic link with lung function. Our single-word-based articulation task may potentially represent a rapid tool for prediction of lung health in people with COPD. Therefore, the use of speech breathing and relevant linguistic–prosodic information could be further integrated into future home-based healthcare systems.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

BZ and EW: original idea, original draft, data analysis, and writing. CZ: original speech data analysis and writing. SD and CO: writing and review. BZ, CO, KE, and S-RP: original data collection. All authors approved the final version of the article.

Funding

This study was funded by Wales Innovation Network: Small Grant Funding (G2-230427/414 1.3.1b).

In memoriam

It is sad that Prof. Edgar Mark Williams passed away prior to the publishing of the research paper. This paper is in a deep and sincerely memory for Prof. Williams' research innovation, energy, curiosity and ambition in breathing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1167902/full#supplementary-material

References

Boersma, P., and Weenink, D. (2022). *Praat: Doing Phonetics by Computer [Computer Program]*. Version 6, 3.01. Available online at: http://www.praat.org/

Carey, M. A., Card, J. W., Voltz, J. W., Arbes, S. J. Jr., Germolec, D. R., Korach, K. S., et al. (2007). (2007). It's all about sex: gender, lung development and lung disease. *Trends Endocrinol. Metabol.* 18, 308–313. doi: 10.1016/j.tem.08003

Cummins, N., Baird, A., and Schuller, B. W. (2018). Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. *Methods* 151, 41–54. doi: 10.1016/j.ymeth.07007

Doyal, L. (2001). Sex, gender, and health: the need for a new approach. *BMJ* 323, 1061–1063. doi: 10.1136/bmj.323.7320.1061

Farrús, M., Codina-Filbà, J., and Escudero, J. (2021). Acoustic and prosodic information for home monitoring of bipolar disorder. *Health Informatics J.* 27, 1460458220972755. doi: 10.1177/1460458220972755

Finnegan, E. M., Luschei, E. S., and Hoffman, H. T. (2000). Modulations in respiratory and laryngeal activity associated with changes in vocal intensity during speech. *J. Speech Lang. Hear. Res.* 43, 934–950. doi: 10.1044/jslhr.4304.934

Fletcher, C., and Peto, R. (1977). The natural history of chronic airflow obstruction. *Br. Med. J.* 1, 1645–1648. doi: 10.1136/bmj.1.6077.1645

Fuchs, S., and Rochet-Capellan, A. (2021). The respiratory foundations of spoken language. *Ann. Rev. Ling.* 7, 13–30. doi: 10.1146/annurev-linguistics-031720-103907

Hixon, T. J., Goldman, M. D., and Mead, J. (1973). Kinematics of the chest wall during speech production: volume displacements of the rib cage, abdomen, and lung. *J. Speech Hear. Res.* 16, 78–115. doi: 10.1044/jshr.1601.78

Huber, J. E. (2008). (2008). Effects of utterance length and vocal loudness on speech breathing in older adults. *Respiratory Physiol. Neurobiol.* 164, 323–330. doi: 10.1016/j.resp.08007

Huber, J. E., Chandrasekaran, B., and Wolstencroft, J. J. (2005). Changes to respiratory mechanisms during speech as a result of different cues to increase loudness. *J. Appl. Physiol.* 98, 2177–2184. doi: 10.1152/japplphysiol.01239.2004

Huttunen, K. H., Keränen, H. I., Pääkkönen, R. J., Päivikki Eskelinen-Rönkä, R., and Leino, T. K. (2011). Effect of cognitive load on articulation rate and formant frequencies during simulator flights. *J. Acoust. Soc. Am.* 129, 1580–1593. doi: 10.1121/1.3543948

Iwarsson, J. M., Thomasson, M., and Sundberg, J. (1998). Effects of lung volume on the glottal voice source. *J. Voice* 12, 424–433. doi: 10.1016/S0892-1997(98)80051-9

Kent, R. D., Kim, Y., and Chen, L. M. (2022). Oral and laryngeal diadochokinesis across the life span: a scoping review of methods, reference data, and clinical applications. *J. Speech Lan. Hearing Res.* 65, 574–623. doi: 10.1044/2021_JSLHR-21-00396

Lively, S. E., Pisoni, D. B., Van Summers, W., and Bernacki, R. H. (1993). Effects of cognitive workload on speech production: acoustic analyses and perceptual consequences. *J. Acoust. Soc. Am.* 93, 2962–2973. doi: 10.1121/1.405815

LoMauro, A., and Aliverti, A. (2018). Sex differences in respiratory function. *Breathe* 14, 131–140. doi: 10.1183/20734735.000318

Macsali, F., Svanes, C., Bjørge, L., Omenaas, E. R., and Real, F. G. (2012). Respiratory health in women: from menarche to menopause. *Expert Rev. Resp. Med.* 6, 187–202. doi: 10.1586/ers.12.15

McDowell, M. A., Fryar, C. D., Ogden, C. L., and Flegal K.M. (2008). Anthropometric reference data for children and adults: United States, 2003–2006. *Natl. Health Stat. Report* 22, 1–48. doi: 10.1037/e623932009-001

Murray, E. S. H., Michener, C. M., Enflo, L., Cler, G. J., and Stepp, C. E. (2018). The impact of glottal configuration on speech breathing. *J. Voice* 32, 420–427. doi: 10.1016/j.jvoice.07001

Nallanthighal, V. S., Mostaani, Z., Härmä, A., Strik, H., and Magimai-Doss, M. (2021). Deep learning architectures for estimating breathing signal and respiratory parameters from speech recordings. *Neural Networks* 141, 211–224. doi: 10.1016/j.neunet.03029

Polgar, G., and Weng, T. R. (1979). The functional development of the respiratory system: from the period of gestation to adulthood. *Am. Rev. Resp. Dis.* 120, 625–695.

Silveyra, P., Fuentes, N., and Rodriguez Bauza, D. E. (2021). Sex and gender differences in lung disease. *Lung Inflammation in Health and Disease, Volume II*. (Cham: Springer), (pp. 227-258).

Somayaji, R., and Chalmers, J. D. (2022). Just breathe: a review of sex and gender in chronic lung disease. *Eur. Resp. Rev.* 31, 21. doi: 10.1183./16000617.0111-2021

Sperry, E. E., and Klich, R. J. (1992). Speech breathing in senescent and younger women during oral reading. *J. Speech Lang. Hear. Res.* 35, 1246–1255. doi: 10.1044/jshr.3506.1246

Stathopoulos, E. T., and Sapienza, C. M. (1997). Developmental changes in laryngeal and respiratory function with variations in sound pressure level. *J. Speech Lang. Hear. Res.* 40, 595–614.

Tayler, N., Grainge, C., Gove, K., Howarth, P., and Holloway, J. (2015). Clinical assessment of speech correlates well with lung function durnig induced bronchoconstriction. *NPJ Primary Care Resp. Med.* 25, 1–3. doi: 10.1038/npjpcrm.2015.6

Thurlbeck, W. M. (1982). Postnatal human lung growth. *Thorax* 37, 564–571. doi: 10.1136/thx.37.8.564

Townsend, E. A., Miller, V. M., and Prakash, Y. S. (2012). Sex differences and sex steroids in lung health and disease. *Endo. Rev.* 33, 1–47. doi: 10.1210/er.2010-0031

Whalen, D. H., and Kinsella-Shaw, J. M. (1997). Exploring the relationship of inspiration duration to utterance duration. *Phonetica* 54, 138–152. doi: 10.1159/000262218

Zeman, K. L., and Bennett, W. D. (2006). Growth of the small airways and alveoli from childhood to the adult lung measured by aerosol-derived airway morphometry. *J. Appl. Physiol.* 100, 965–971. doi: 10.1152/japplphysiol.00409.2005

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 11 September 2023
DOI 10.3389/fpsyg.2023.1135469

Check for updates

OPEN ACCESS

EDITED BY
Kevin Tang,
Heinrich Heine University of Düsseldorf, Germany

REVIEWED BY
Melania Cabezas-García,
University of Granada, Spain
Isabel Neitzel,
Technical University Dortmund, Germany

*CORRESPONDENCE
Joshua Cohen
✉ jcohen@clarigenthealth.com
Vikram Ramanarayanan
✉ vikram.ramanarayanan@modality.ai

[†]These authors share first authorship

RECEIVED 31 December 2022
ACCEPTED 14 August 2023
PUBLISHED 11 September 2023

CITATION
Cohen J, Richter V, Neumann M, Black D, Haq A, Wright-Berryman J and Ramanarayanan V (2023) A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations. *Front. Psychol.* 14:1135469. doi: 10.3389/fpsyg.2023.1135469

A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations

Joshua Cohen[1*†], Vanessa Richter[2†], Michael Neumann[2], David Black[1], Allie Haq[1], Jennifer Wright-Berryman[3] and Vikram Ramanarayanan[2,4*]

[1]Clarigent Health, Mason, OH, United States, [2]Modality.AI, Inc., San Francisco, CA, United States, [3]Department of Social Work, College of Allied Health Sciences, University of Cincinnati, Cincinnati, OH, United States, [4]Otolaryngology - Head and Neck Surgery (OHNS), University of California, San Francisco, San Francisco, CA, United States

Background: The rise of depression, anxiety, and suicide rates has led to increased demand for telemedicine-based mental health screening and remote patient monitoring (RPM) solutions to alleviate the burden on, and enhance the efficiency of, mental health practitioners. Multimodal dialog systems (MDS) that conduct on-demand, structured interviews offer a scalable and cost-effective solution to address this need.

Objective: This study evaluates the feasibility of a cloud based MDS agent, Tina, for mental state characterization in participants with depression, anxiety, and suicide risk.

Method: Sixty-eight participants were recruited through an online health registry and completed 73 sessions, with 15 (20.6%), 21 (28.8%), and 26 (35.6%) sessions screening positive for depression, anxiety, and suicide risk, respectively using conventional screening instruments. Participants then interacted with Tina as they completed a structured interview designed to elicit calibrated, open-ended responses regarding the participants' feelings and emotional state. Simultaneously, the platform streamed their speech and video recordings in real-time to a HIPAA-compliant cloud server, to compute speech, language, and facial movement-based biomarkers. After their sessions, participants completed user experience surveys. Machine learning models were developed using extracted features and evaluated with the area under the receiver operating characteristic curve (AUC).

Results: For both depression and suicide risk, affected individuals tended to have a higher percent pause time, while those positive for anxiety showed reduced lip movement relative to healthy controls. In terms of single-modality classification models, speech features performed best for depression (AUC = 0.64; 95% CI = 0.51–0.78), facial features for anxiety (AUC = 0.57; 95% CI = 0.43–0.71), and text features for suicide risk (AUC = 0.65; 95% CI = 0.52–0.78). Best overall performance was achieved by decision fusion of all models in identifying suicide risk (AUC = 0.76; 95% CI = 0.65–0.87). Participants reported the experience comfortable and shared their feelings.

Conclusion: MDS is a feasible, useful, effective, and interpretable solution for RPM in real-world clinical depression, anxiety, and suicidal populations. Facial information is more informative for anxiety classification, while speech and language are more discriminative of depression and suicidality markers. In general, combining speech, language, and facial information improved model performance on all classification tasks.

KEYWORDS

machine learning, multimodal dialog systems, speech features, natural language processing, facial features, suicide, depression, anxiety

1. Introduction

Globally, ∼301 million people and 280 million people were affected by anxiety and depression in 2019, respectively.[1] According to the World Health Organization, over 700,000 people die by suicide every year, with more than 20 suicide attempts per suicide death (World Health Organization, 2021). In 2022 in the United States (US), 5.0% of adults report regular feelings of depression, and 12.5% report regular feelings of worry, nervousness, or anxiety.[2] Regarding the frequency of suicidal thoughts in the US, 3.7% of adults had serious thoughts of suicide in 2021. Initial estimates of the impact of the COVID-19 pandemic show more than a 25% increase in mental disorders, worldwide (World Health Organization, 2022). As the prevalence of these conditions increases, technological solutions are needed to more efficiently identify, monitor, and manage these conditions.

The identification and monitoring of mental health conditions related to depression, anxiety, and suicide risk often rely on self report from individuals or evaluation from a trained professional. Self report scales, such as the Patient Health Questionnaire-9 Item (PHQ-9) for depression (Kroenke et al., 2001) or the Generalized Anxiety Disorder-7 Item (GAD-7) for anxiety (Spitzer et al., 2006), have reported excellent sensitivity and specificity but rely on the honesty of a patient and typically only screen for a single condition, requiring additional time to screen for more than one condition. For suicide risk, a 2017 meta analysis suggests current methods of predicting death by suicide are no better than random chance, and recommends other techniques such as machine learning (ML) to improve predictive capabilities (Franklin et al., 2017).

In clinical settings, information from a patient's visual appearance and body language, verbal communications, and speech may aid clinicians' diagnoses. More recently, these signals have been combined with supervised ML as biomarkers to identify the presence of mental health conditions (Ramanarayanan et al., 2022). In this context, signals from different modalities (e.g., speech or visual inputs) are transformed into features, leading to 100's to 1,000's of data points that describe aspects of the signal. For example, speech (S) features describe characteristics of an acoustic signal, such as pitch or intensity. Facial features (F) describe aspects of face movements, such as the number of eye blinks per second or the speed of the lower lip and jaw center. Text (T) features are derived from a patient's language and may capture relevant semantic information. During supervised ML, features are paired with a clinical label, such as having a condition (case) or not (control), and then used to train a model to allow the discovery of patterns from the data for classification.

Reviews of articles using ML with SFT features for the identification of depression, anxiety, and suicide risk indicate good to excellent model performance, with many investigators reporting areas under the receiver operating characteristic curve (AUC) in the range of 0.7–0.9 (Cummins et al., 2015, 2018; Arif et al., 2020; Bernert et al., 2020; Neumann et al., 2020; Kusuma et al., 2022). Comparatively, under realistic clinical conditions, many traditional mental health diagnostic checklists perform with AUCs in the range of 0.7–0.8 (Rice and Harris, 2005; Youngstrom, 2013). While ML models appear to perform with a similar discriminative ability as traditional methods, they face unique challenges. A key challenge is model overfitting, which occurs when a model learns from idiosyncrasies of a dataset as opposed to clinically meaningful variables. This leads to overly optimistic estimates of model performance and may result in a model that is overly sensitive to specific expressions of mental health conditions, reducing its effectiveness when symptoms are expressed differently (Berisha et al., 2022). Mitigation strategies include cross-validation, regularization, or using models with fewer parameters. Additionally, there are significant challenges to generalizability. Work by Botelho et al. (2022) shows a high degree of separability among six popular speech datasets, demonstrating the limitation that a model trained on one population or dataset might not accurately predict outcomes in another. This necessitates rigorous external validation and diverse, representative data collection. Biases in the dataset represent another source of error. If the training data predominantly represents a specific demographic or cultural group, the model may not perform as well on other groups, leading to misdiagnosis or underdiagnosis. Furthermore, Berisha et al. (2022) recently reported a negative association between model performance and sample size among 77 publications on speech-based ML for the identification of dementia, attributing this to not only model overfitting but also publication bias. This finding underscores the importance of transparency and balanced

Abbreviations: AUC, area under receiver operating characteristic curve; C-SSRS, Columbia-Suicide Severity Rating Scale; CI, Confidence; CRC, Clinical Research Coordinator; F0, fundamental frequency; GAD-7, Generalized Anxiety Disorder-7; HNR, harmonics-to-noise ratio; KW, Kruskal-Wallis; LR, Logistic Regression; MHSAFE, Mental Health Hopes Secrets Anger Fear and Emotional Pain; MDS, Multimodal Dialog System; ML, Machine Learning; NLP, Natural Language Processing; PHQ-9, Patient Health Questionnaire-9; PPT, percent pause time; SFT, speech, facial, and text (modalities); SIT, speech intelligibility test; SVM, support vector machine; TF-IDF, Term Frequency-Inverse Document Frequency.

[1] Global Health Data Exchange (GHDx), https://ghdx.healthdata.org/ [accessed: 2022-12-20].

[2] Centers for Disease Control, https://www.cdc.gov/nchs/fastats/mental-health.htm [accessed: 2023-07-11].

reporting in research publications. Taken together, these issues emphasize the importance of not solely focusing on classification performance but also on selecting clinically meaningful and generalizable features, ensuring a representative dataset, and employing robust validation methodologies. While ML shows promising potential in mental health diagnostics, these challenges must be recognized and addressed to maximize its clinical utility.

Previous research related to this work has found a semi-structured, in-person interview promising for the collection of SFT features to be used with ML models for the identification of suicide risk (Pestian et al., 2010, 2016, 2017; Laksana et al., 2017; Cohen et al., 2020, 2022; Wright-Berryman et al., 2023). In these studies, trained staff (therapists, clinical research coordinators, or licensed behavioral health clinicians) recorded a semi-structured interview with hundreds of suicidal or non-suicidal participants in emergency departments, psychiatric units, and in-school therapy settings with adolescents and adults. Support vector machine (SVM) models were trained to identify suicidal vs non-suicidal participants, with AUCs ranging from 0.69 to 0.93 depending on the features and cross-validation approach used (Pestian et al., 2016, 2017). Notably, two of these investigations included an external validation of the models developed with separately collected corpora (Cohen et al., 2020, 2022). It is also important to note that while these studies involved hundreds of participants, there is no universally accepted minimum sample size for ML analyses. The required sample size can vary greatly depending on the complexity of the model, the number of features, the variability in the data, and the specific research question being addressed. Some studies have successfully applied ML techniques with as few as 60 sessions (Pestian et al., 2016).

While these initial results are encouraging for the use of SFT features for the identification of suicide risk in clinical settings, the procedures relied on trained staff to conduct the interview. There is a shortage of mental health professionals (Satiani et al., 2018), which may limit the uptake of technology requiring more of their time. Therefore, techniques to accurately and autonomously screen for mental health concerns are needed. One option may be to use multimodal dialog systems (MDS), which have recently been developed for remote health screening and monitoring. For example, DeVault et al. (2014) presented the SimSensei Kiosk, a virtual human interviewer specifically built to render clinical decision support. It captures verbal and non-verbal features to extract distress indicators correlated with mental conditions such as depression. Lisetti et al. (2015) presented results of a large-scale effort building a virtual health assistant for “brief motivational interventions,” for example, interviews about a subject’s drinking behavior. The described system uses text input from the subject’s keyboard (or, alternatively, a speech recognition hypothesis) along with facial expression features to determine next steps in the interaction. In addition to cost reduction and scalability, MDSs may reduce participants’ fear associated with the perception of being judged (Cummins et al., 2015). Gratch et al. (2014) found that participants felt more comfortable disclosing personal information with an agent that was framed as autonomous as opposed to one that was framed as human-controlled.

For the present study, the Modality service, a cloud-based MDS (Suendermann-Oeft et al., 2019; Ramanarayanan et al., 2020) was used to conduct automated, structured interviews with participants. Neumann et al. (2020) recently demonstrated the utility of the Modality MDS in differentiating people with mild, moderate and severe depression, and similar studies have also been conducted in ALS (Neumann et al., 2021), Parkinson’s disease (Kothare et al., 2022), schizophrenia (Richter et al., 2022), and autism (Kothare et al., 2021). The Modality MDS can be used with widely available endpoints such as smartphones and laptops as opposed to the dedicated, locally administered hardware used in other studies. Speech, facial, and language data was collected by the MDS for feature analysis and ML classification of depression, anxiety, and suicide risk. Overall, we found participants accepting of the technology and procedures, and ML models using a combination of features led to the greatest discriminative ability.

This case-control study sought to (1) examine the feasibility of collecting a mental health interview with an MDS with participants with and without depression, anxiety, and suicide risk, (2) evaluate candidate features for the identification of these conditions, and (3) internally validate models trained to identify each condition with different modalities (speech, facial, and text).

2. Methods

2.1. Data

Sixty-eight participants enrolled in the study between October 2021 and April 2022, providing a total of 73 sessions. Notably, participants were allowed (but not required) to participate again after 2 weeks, out of whom five participants chose to take part in another session each. The PHQ-9 to measure depression (Kroenke et al., 2001), the GAD-7 to measure anxiety (Spitzer et al., 2006), and the Columbia-Suicide Severity Rating Scale (C-SSRS) Screener (Posner et al., 2011) to measure suicide risk were collected in all sessions. Participant demographics and distributions of case sessions are shown in Table 1. For a more complete picture of the study participants, statistics of control participants and additional demographic information are available in Supplementary Table 1.

Criteria for participant recruitment were: (1) age $\geq$ 18, (2) able to provide informed consent, (3) English as a primary language, and (4) located in the United States. Recruitment for the study was done via ResearchMatch, a national health volunteer registry that was created by several academic institutions and supported by the U.S. National Institutes of Health as part of the Clinical Translational Science Award program. ResearchMatch has a large population of volunteers who have consented to be contacted by researchers about health studies for which they may be eligible. For this study, we specifically targeted individuals who had self-selected to be contacted by studies related to depression, anxiety, and suicide risk. This targeted recruitment strategy was designed to ensure a sufficient number of participants with the conditions of interest. Review and approval for this study and all procedures was obtained from our commercial Institutional Review Board. All participants gave informed consent in accordance with the Declaration of Helsinki before they participated in the study. Participants received a $15 gift card for each session they completed.

TABLE 1 Participant descriptive statistics and case session summaries.

			Case sessions		
Variable	Participants	Sessions	PHQ-9 $\geq$ 10	GAD-7 $\geq$ 10	C-SSRS $\geq$ Mod.
Count (%)	68 (100.0%)	73 (100.0%)	15 (20.6%)	21 (28.8%)	26 (35.6%)
Average age (SD)	38.8 (14.7)	38.7 (14.7)	39.3 (13.3)	34.5 (13.1)	38.8 (15.7)
Average interview length (min) (SD)	9.6 (2.2)	9.3 (2.3)	9.7 (2.5)	9.0 (2.4)	9.7 (2.2)
Average word count (SD)	917.0 (302.06)	925.0 (309.9)	912.1 (374.2)	899.1 (308.5)	964.1 (316.2)
Sex					
Male (%)	15 (22.1%)	16 (21.9%)	3 (4.11%)	6 (8.2%)	9 (12.3%)
Female (%)	52 (76.5%)	56 (76.7%)	12 (16.4%)	15 (20.6%)	16 (21.9%)
Prefer not to answer	1 (1.5%)	1 (1.4%)	- (-%)	- (-%)	1 (1.4%)
Race					
White or Caucasian (%)	50 (73.5%)	54 (74%)	12 (16.4%)	17 (23.3%)	21 (28.8%)
Black or African American (%)	10 (14.7%)	11 (15.1%)	2 (2.7%)	3 (4.1%)	2 (2.7%)
Asian (%)	5 (7.4%)	5 (6.9%)	- (-%)	1 (1.4%)	- (-%)
Other (%)	3 (4.4%)	3 (4.1%)	1 (1.4%)	- (-%)	3 (4.1%)

2.1.1. Study staff

The study staff was composed of three clinical research coordinators (CRC) who are all mental health practitioners or graduate-level students in the mental health field. They are extensively trained in all study procedures, human subjects protection, good clinical practice, and crisis management procedures. The CRCs oversaw all study procedures.

2.2. Study design

Participants invited through ResearchMatch completed informed consent and demographic information electronically and scheduled a remote study session time with a CRC to meet via the video conferencing platform Microsoft Teams. During the study session, the CRC confirmed participant consent and that they understood the study procedures and their rights as participants, and then administered the PHQ-9, GAD-7, and the C-SSRS Screener. These instruments have been administered via video conferencing platforms in a variety of studies. Figure 1 outlines the study procedures.

The PHQ-9 is a rigorously tested, reliable and valid instrument for depression in adults, with a sensitivity and specificity of 88%, corresponding with a threshold score $\geq$ 10 out of 27, which includes "Moderate," "Moderately Severe," and "Severe" levels of depression (Kroenke et al., 2001). Similarly, the GAD-7 has been widely tested with adults to measure anxiety, with a sensitivity and specificity of 89 and 82%, respectively, corresponding with a threshold score $\geq$ 10 out of 21, which includes "Moderate" and "Severe" levels of anxiety (Spitzer et al., 2006). The C-SSRS Screener is a structured interview which has demonstrated high sensitivity and specificity for classifying suicidal ideation and behaviors in a multi-site emergency department study (Posner et al., 2011). The screener asks six questions about the past month to measure suicidal ideation and suicidal behaviors on an ordinal scale, with a final question about lifetime suicidal behavior (more than 3 months ago). The C-SSRS Screener designates suicide risk as "None" if all questions are answered negatively, "Low" if passive suicidal ideation is present, "Moderate" if suicidal ideation with a method OR lifetime suicidal behavior is present, and "High" if suicidal ideation with intent (with or without a method) OR suicidal behavior in the past 3 months is present. In this study a severity threshold $\geq$ "Moderate" was used for the binary identification of all conditions to maximize sensitivity and specificity of the instruments. Table 2 is a summary of the assessments and scores used for case definitions.

For participant safety, all participants received wellness resources such as the 988 Suicide and Crisis Lifeline and the Crisis Text Line. The 988 National Suicide Prevention Lifeline and the Crisis Text Line are U.S.-based single line immediate access to trained crisis counselors. For participants that score "High" risk on the C-SSRS Screener, a more comprehensive contingency plan was followed, including asking additional questions about their mental state, access to lethal means, engagement in mental health services, and protective factors. In the event of imminent risk, the contingency safety plan included a warm hand-off to the 988 Suicide and Crisis Lifeline and/or a call to 911. No participants in this study were at imminent risk and required following of the contingency safety plan.

Following the PHQ-9, GAD-7, and C-SSRS, the CRCs provided a link to the MDS, a web-based program that accesses the participant's computer's microphone and webcam to record their voice and facial video. To supervise this section of the study, CRCs instructed participants to share their computer's screen and audio. The CRC then muted their microphone and turned off their webcam.

Before participants start their conversation with the virtual agent, Tina—implemented via a scalable, cloud-based MDS to conduct automated structured interactions (Suendermann-Oeft et al., 2019; Ramanarayanan et al., 2020)—tests of the speaker, microphone, and camera need to be passed to ensure that the participants' devices are correctly configured so that the collected

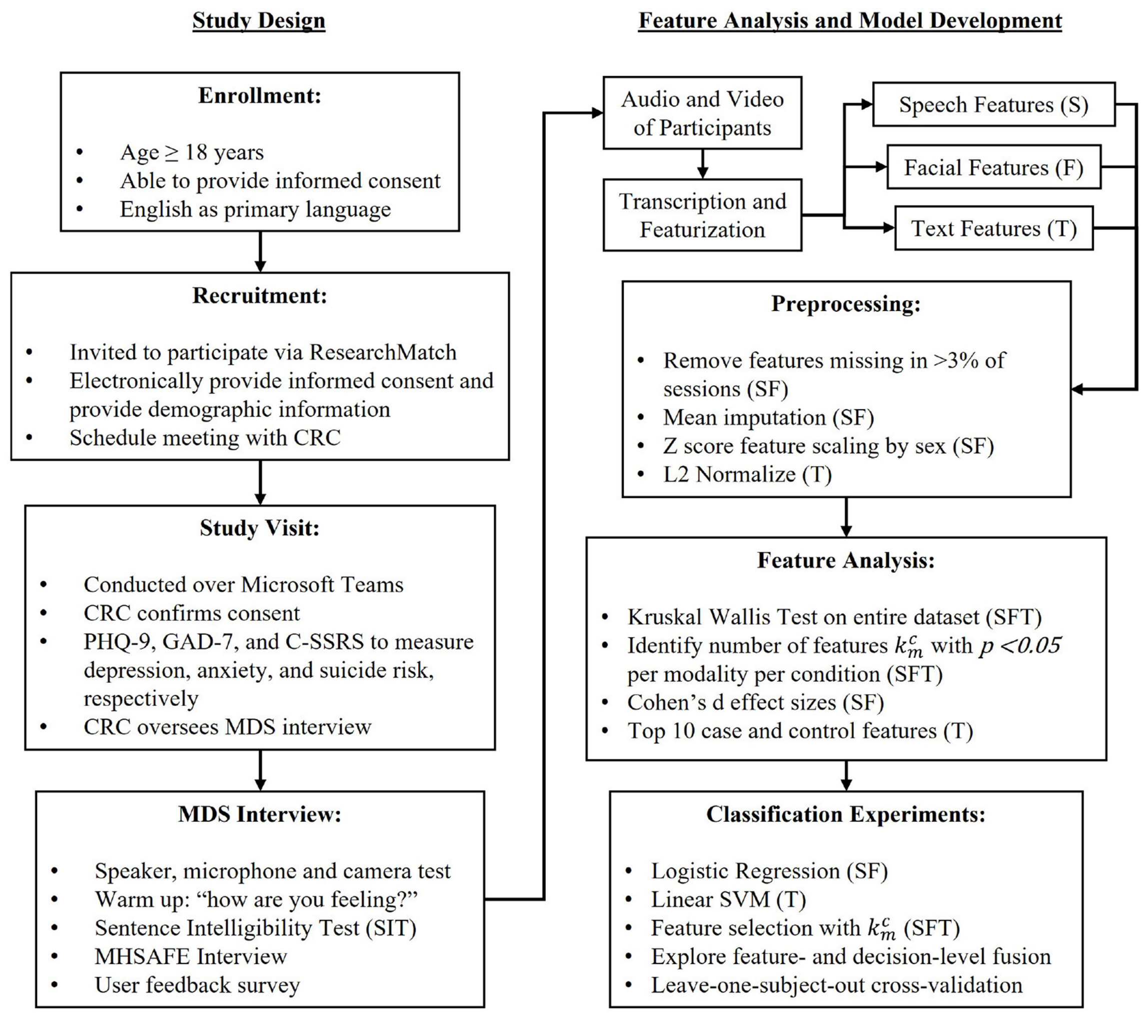

Study Design
Feature Analysis and Model Development
Enrollment:
• Age ≥ 18 years
• Able to provide informed consent
• English as primary language
Recruitment:
• Invited to participate via ResearchMatch
• Electronically provide informed consent and provide demographic information
• Schedule meeting with CRC
Study Visit:
• Conducted over Microsoft Teams
• CRC confirms consent
• PHQ-9, GAD-7, and C-SSRS to measure depression, anxiety, and suicide risk, respectively
• CRC oversees MDS interview
MDS Interview:
• Speaker, microphone and camera test
• Warm up: "how are you feeling?"
• Sentence Intelligibility Test (SIT)
• MHSAFE Interview
• User feedback survey
Audio and Video of Participants
Transcription and Featurization
Speech Features (S)
Facial Features (F)
Text Features (T)
Preprocessing:
• Remove features missing in >3% of sessions (SF)
• Mean imputation (SF)
• Z score feature scaling by sex (SF)
• L2 Normalize (T)
Feature Analysis:
• Kruskal Wallis Test on entire dataset (SFT)
• Identify number of features k_m^c with $p < 0.05$ per modality per condition (SFT)
• Cohen's d effect sizes (SF)
• Top 10 case and control features (T)
Classification Experiments:
• Logistic Regression (SF)
• Linear SVM (T)
• Feature selection with k_m^c (SFT)
• Explore feature- and decision-level fusion
• Leave-one-subject-out cross-validation

FIGURE 1
Schematic of study and modeling procedures.

data has sufficient quality. Once all device tests pass, Tina guides participants through an interactive interview.

In each participant's first session, Tina introduced the graphical interface and asked the participant to read a sentence, taken from a speech intelligibility test (SIT) corpus. At the start of every interview, Tina first asked participants "how they are feeling today" as a warm up question. Participants then began a semi-structured inteview (renamed MHSAFE—hope, secrets, anger, fear, and emotional pain–from the "Ubiquitous Questionnaire"). The MHSAFE interview has been used in previous studies with human interviewers to collect language for ML models to identify suicide risk (Pestian et al., 2010, 2016, 2017; Laksana et al., 2017; Cohen et al., 2020, 2022). The interview asks participants open-ended questions about five topics: hope, secrets, anger, fear, and emotional pain (Pestian, 2010; Cohen et al., 2020, 2022). In the present study, for each topic, Tina asks if they have that topic and how that makes them feel, for example, "do you have hope and how does that feel?" The question about secrets is not intended for participants to reveal what their secrets are, but to gather information about whether they are keeping secrets at all, and how they feel about this. The SIT task and warm up question from the beginning were included in the analysis, because these speech samples may contain useful features in addition to the MHSAFE interview.

Tina is equipped with a voice activity detection system to measure the length of participant responses. To collect enough language for analysis, Tina required a minimum of 1 min of speech for each topic of the MHSAFE interview. Participants that did not speak for the minimum amount of time were nudged up to two times to tell Tina more about that topic. Tina moved onto to the next question if after two nudges the participant's speaking time for that questions was still <1 min. The recorded audio files were manually transcribed using a HIPAA-compliant service.

2.3. User feedback

For user feedback, we used two forms of data collection, a qualitative questionnaire (likes/advantages, dislikes/disadvantages, and improvements) and a five-question survey with Likert scale responses, shown in Table 3. Qualitative data were analyzed

TABLE 2 Summary of completed assessments, associated mental state measured, and case definition for model development.

Assessment	Mental state	Case definition
PHQ-9	Depression	Total ≥ 10
GAD-7	Anxiety	Total ≥ 10
C-SSRS screener	Suicidal risk	Risk ≥ Moderate

TABLE 3 Post interview survey.

Item	Survey questions
Likert scale questions: 1 = most negative to 5 = most positive	
1.	How did it feel to express your emotions of your hope, secrets, anger, fear, and emotional pain to a virtual assistant?
2.	How honest were you in your responses to the virtual assistant?
3.	How comfortable were in your responses to the virtual assistant?
4.	What was your impression of the virtual assistant in terms of visual appearance and voice?
5.	What was your impression of the virtual assistant in terms of pace of interview including interruptions and pauses from the virtual assistant, and your time to respond?
Open-ended questions:	
6.	What did you like about Tina?
7.	What did you not like about Tina?
8.	What could be improved with this experience?

using thematic analysis. Two investigators coded the responses and annotated the emerging themes. Likert scale responses were analyzed using frequency distribution, mean, and standard deviation. Student's *t*-tests were performed with SciPy's *ttest_ind* function to identify any statistically significant differences between case and control groups for Likert scale responses.

2.4. Data preprocessing and featurization

All analysis was performed using the Python programming language (version 3.9.12; Van Rossum and Drake, 1995). The following open-source Python libraries were also used: Pandas (version 1.4.2; McKinney, 2010; The Pandas Development Team, 2020), Numpy (version 1.22.3; Oliphant, 2007; Van Der Walt et al., 2011), scikit-learn (version 1.0.2; Pedregosa et al., 2011), Matplotlib (version 3.5.1; Hunter, 2007), and SciPy (version 1.8.0; Virtanen et al., 2020). For calculating effect sizes, we also used the R package effsize (version: 0.7.6; Torchiano, 2020) and the rpy2 interface (version 2.9.4).[3]

In our methodology, three modalities—acoustic (speech), facial, and textual—were examined, each contributing a distinct set of features to our models and are described in more detail below.

TABLE 4 Overview of speech, facial, and text features.

	Domain	Features
Speech	Energy	Shimmer (%), signal-to-noise ratio (dB)
	Timing	Speaking and articulation duration (sec.), percent pause time (PPT, %)
	Voice quality	Harmonics-to-noise ratio (HNR, dB)
	Frequency	Mean, max., min. fundamental frequency F0 (Hz), jitter (%)
Facial	Mouth (distances)	Lip aperture/opening, lip width, mouth surface area, Mean symmetry ratio between left and right half of the mouth
	Movement	Velocity, acceleration, jerk, and speed of lower lip and jaw center
	Eyes	Number of eye blinks per sec., eye opening, vertical displacement of eyebrows
Text	TF-IDF	$\frac{\text{Count of n-gram in interview}}{\text{Count of interviews containing n-gram}}$

For facial features, functionals (minimum, maximum, and average) are applied to produce one value across all video frames of an utterance.

2.4.1. Speech features

For the acoustic speech analysis, a variety of commonly established measures for clinical voice analysis were extracted (France et al., 2000; Mundt et al., 2007, 2012). These include *timing measures*, such as percentage of pause time (PPT), *frequency domain measures*, such as fundamental frequency (F0) and jitter, *energy-related measures*, such as intensity and shimmer as well as the harmonics-to-noise ratio (HNR) as a measure for *voice quality*. All measures were extracted with Praat (Boersma and Van Heuven, 2001). Table 4 lists all features. More detailed descriptions of speech features are available in Supplementary Table 2.

2.4.2. Facial features

The set of facial features is based on facial landmarks generated in real time by the MediaPipe Face Mesh algorithm (Kartynnik et al., 2019). For each user turn, the following algorithm is applied to compute features. First, MediaPipe Face Detection, which is based on BlazeFace (Bazarevsky et al., 2019), is used to determine the (x, y)-coordinates of the face for every frame. Then, facial landmarks are extracted using MediaPipe Face Mesh. We use 14 key landmarks to compute features like the speed and acceleration of articulators (jaw and lower lip), surface area of the mouth, and eyebrow raises (see Table 4). The key facial landmarks are illustrated in Figure 2. Lastly, the features are normalized by dividing them by the inter-caruncular or inter-canthal distance, which is the distance between the inner canthi of the eyes (see Figure 2 for a visual illustration), to handle variability across participant sessions due to position and movement relative to the camera (Roesler et al., 2022). More detailed descriptions of facial features are available in Supplementary Table 3.

2.4.3. Text features

The natural language processing (NLP)/ML pipeline used in this study focused on the term frequency-inverse document

[3] https://github.com/rpy2/rpy2

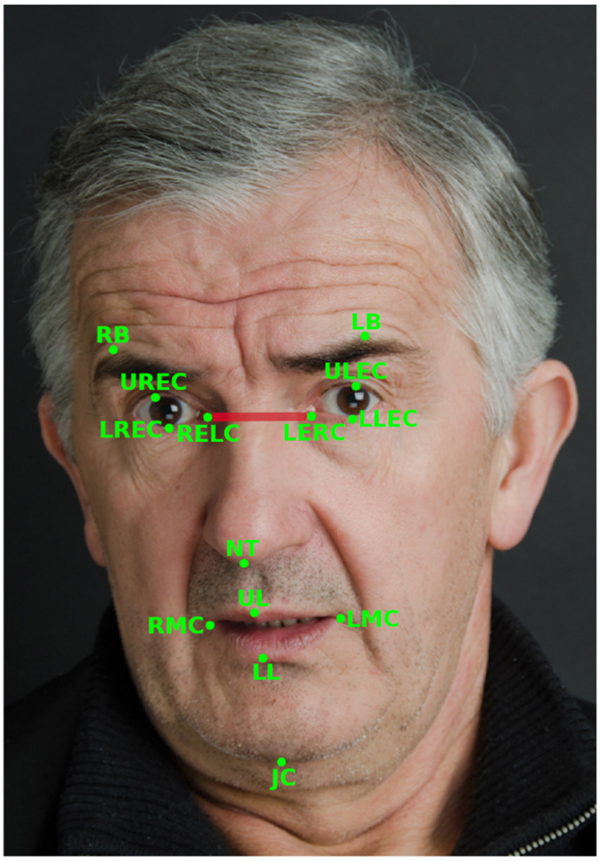


FIGURE 2
Illustration of the 14 facial landmarks used to calculate the facial features used in this study and the inter-caruncular distance (shown in red) between the inner canthi of the eyes (RELC and LERC). Adapted with permission from iStock Photo/meshaphoto.

frequency (TF-IDF) of unigrams (single words), calculated using `scikit-learn`'s *TfidfVectorizer*. TF-IDF is a numerical statistic that reflects how often a term appears in a document (i.e., interview), while also taking into account how common the term is in the entire corpus of documents. This weighting scheme assigns higher importance to terms that are more distinctive to a particular document, and lower importance to terms that are common across many documents (Rajaraman and Ullman, 2011).

The text was preprocessed so all characters were lowercase and to remove any punctuation and non-letter characters. Language was tokenized by splitting on white spaces. Following the preprocessing steps, each session was subject to L2 normalization, a process designed to control for varying response lengths. L2 normalization, also known as Euclidean normalization, works by adjusting the values in the data vector so that the sum of the squares of these values equals one. Specifically, each value in the vector is divided by the Euclidean length (L2 norm) of the vector itself—the square root of the sum of the squared vector values.

2.4.4. Missing data

Features may be missing if a participant skipped a segment or a technical issue arose. To handle missing speech and facial features, `scikit-learn`'s *SimpleImputer* was used to replace the missing feature with its mean value for each cross-validation fold. Any feature missing from $> 3\%$ of sessions was removed prior to model evaluation to ensure the robustness of our analyses and to avoid potential biases or inaccuracies that could arise from imputing a large amount of missing data.[4]

2.5. Feature analysis and classification experiments

Due to the limited size of our dataset, we performed feature analysis on the entire dataset to identify the *number* of significant features (but importantly, not which features). In other words, during classification experiments, we only specified the number of features, and not the specific features, per cross-validation fold to avoid information leakage across training and validation folds. To test statistical meaningfulness of the features, non-parametric Kruskal-Wallis (KW) tests (McKight and Najab, 2010) were conducted on the entire dataset for each feature, which test the hypothesis that feature medians are significantly different between cohorts (cases and controls) at the $\alpha = 0.05$ level. In order to give equal weight to the features of individual participants, we selected only one session per user for this test. This results in k_m^c number of features per modality (m), per condition (c). For speech and facial features, effect sizes were then calculated with Cohen's d (Cohen, 1988), which analyzes the direction and magnitude of effects between cohorts. Cohen's d was introduced to measure effect sizes in units of variability by dividing the difference of cohorts' means by the pooled standard deviation. Because TF-IDF featurization of participant language results in sparse matrices, we did not measure effect sizes, but instead extracted the top 10 case and control features by feature weight per condition from a linear SVM fit to the entire dataset after feature selection for the k_m^c features determined by the KW test. Figure 1 includes a schematic of feature analysis and model development procedures.

Discrimination power was assessed by evaluating the classification performance using a logistic regression (LR) classifier for speech and facial features and a linear SVM for text features. These classifiers were selected for their relative simplicity and promising performance in previous studies (Pestian et al., 2016, 2017; Laksana et al., 2017; Cohen et al., 2020, 2022). To prune our high-dimensional feature set, the number of speech and facial input features for the classifier was determined by k_m^c. We selected the top k_m^c features that resulted from a KW test on $n - 1$ participants' session(s) in each classification fold. To ensure robustness and reliability of our results, we only reported ML experiments if they were based on at least five significant features. This threshold was set to avoid over-reliance on a small number of features or outliers, and to provide a more robust basis for classification.

[4] We consider each combination of a speech/facial measure and a task (interview question) as one feature. Twenty-eight percent of speech features (26) and 33% of facial features (146) were removed because of missing data. The majority of these were from the initial SIT sentence and the warm up question.

The acoustic characteristics of male and female voices have been studied in detail and found to differ in a variety of variables such as pitch, voice quality, and timing measures (see, for example, Titze, 1989; Mendoza et al., 1996; Simpson, 2009). Furthermore, facial behavior as well as classification accuracy based on facial features was found to differ by gender (Dimberg and Lundquist, 1990; Drimalla et al., 2020). To ensure that analyses between case and control cohorts are unbiased with respect to widely reported differences between males and females, we standardized scores for speech and facial features by z-scoring for both groups separately.[5]

Both feature- and decision-level fusion were examined to identify any potential predictive benefits of including information from multiple modalities. During feature fusion, features are independently preprocessed and selected, and then merged into a single matrix prior to model development and evaluation. An LR classifier was used for feature fusion classification. Decision fusion involves independently training models on each modality or a combination thereof (e.g., speech and facial features combined together), and then combining outputs from each model through different rules. For decision fusion, LR classifiers were used for speech and facial features, while a linear SVM was used with text features. Model output combination rules considered include the minimum, maximum, and mean of all model output scores.

Models were trained using different feature combinations paired with each session's label as a case or control. Models were evaluated using a leave-one-subject-out cross-validation approach, where a model is iteratively trained on all but one participants' session(s). The features from the held out subject's session(s) were fed into the model and a probability for belonging to the case group was returned. When done iteratively, this results in a list of probabilities for each session to be compared to the true label to compute overall model performance metrics. Model performance was primarily evaluated with the AUC and Brier score. AUC values range from 0.5 (random chance) to 1.0 (perfect model). The Brier score is a measure of model calibration and ranges from 0 to 1 where low scores indicate less discrepancy between labels and predicted probabilities.

The selection of features in the cross-validation folds in the classification experiments based on a KW test on $n-1$ participants may differ from the result of the KW test on the entire cohort. To identify the most important features for each mental state in terms of robustness and generalizability across experiments and thus independence from participant partitions, we assessed these by determining the intersection of features that (a) were consistently selected across *all* cross-validation folds and (b) were found to be statistically significant in the KW test for the entire cohort. We then examined these features in more detail by reviewing previous research and by testing their association with the respective mental states. For the latter, Pearson correlations were calculated between the assessment total scores and the speech and facial features. A threshold of $|r| \geq 0.2$ and $p < 0.05$ was used to identify weak, but statistically significant correlations. We acknowledge that an $|r|$ value of 0.2 is often considered a "small" effect size. However, in the context of our exploratory analysis with a relatively smaller dataset, we chose this threshold to highlight any potential weak, but statistically significant relationships that may warrant further investigation in larger studies. This approach allows us to focus on potentially clinically meaningful features and gives a more nuanced understanding of the data, rather than focusing exclusively on model performance.

3. Results

3.1. Feature analysis and classification experiments

Figure 3 shows the effect sizes of the speech and facial features that are statistically significantly different between the respective cohorts. For PHQ-9 assessments, we find one facial and 15 speech features, as can be seen in Figure 3A. These features include a higher percent pause time as well as lower shimmer, jitter and F0 standard deviation for cases than controls. Conversely, for comparisons based on GAD-7 scores, more facial features (24) are evident than speech features (seven), as shown in Figure 3B. Similar to the GAD-7 assessments, seven speech and 24 facial features were found to be significant in the statistical analysis based on the C-SSRS scores, which is shown in Figure 3C. For each of the conditions examined in our study - depression, anxiety, and suicide—the top 10 text features (words) for both cases and controls were extracted using linear SVM models fit to the entire corpus, after the feature selection process. Importantly, these textual features do not overlap with or include acoustic or facial features - they are entirely separate. Table 5 provides a list of these top textual features for each condition.

Receiver operating characteristic (ROC) curves can be seen in Figure 4 for the identification of depression, anxiety, and suicide risk for single modalities, feature fusion, and decision fusion models. The results in terms of AUC and Brier score are shown in Table 6. Of the single-modality models, the best performance for depression occurred with speech features (AUC = 0.64; 95% CI = 0.51–0.78); for anxiety, facial features performed best (AUC = 0.57; 95% CI = 0.43–0.71); and for suicide risk, text features performed best (AUC = 0.65; 95% CI = 0.52–0.78). While these AUC values indicate that the models have some predictive power, it's important to highlight that an AUC of 0.5 would be equivalent to random chance and values in the range of 0.7–0.8 are often considered indicative of a good performing model. Thus, it can be seen that some of our single-modality models are performing at near-chance or sub-optimal levels (see Figure 4).

In general, we found a combination of features or models improved discriminative ability, with a decision fusion of all models leading to the best overall performance in the identification of suicide risk (AUC = 0.76; 95% CI = 0.65–0.87). The best discriminative ability for depression (AUC = 0.70; 95% CI = 0.56–0.84) resulted from a decision level fusion of speech and text features. For anxiety, a feature-level fusion of all features performed best (AUC = 0.71; 95% CI = 0.59–0.83). The best performance for all decision-level fusion models resulted by selecting the minimum score of considered mode.

[5] One out of 68 participants did not specify their sex at birth. This participant's session was excluded from the analysis of speech and facial features.

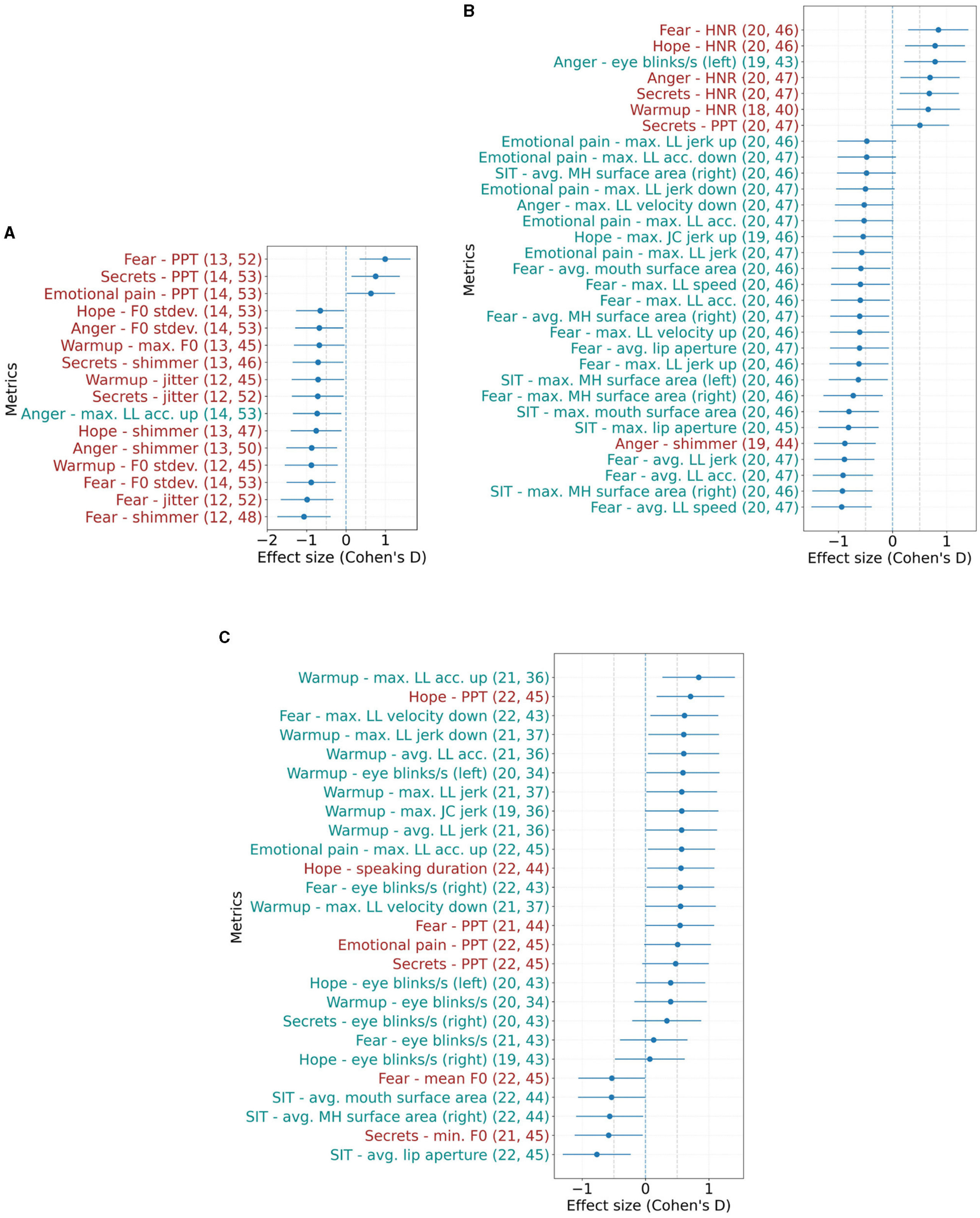


FIGURE 3
Effect sizes (Cohen's *d*) of speech and facial metrics that show statistically significant differences between controls and cases based on **(A)** PHQ-9 ≥ 10, **(B)** GAD-7 ≥ 10, and **(C)** C-SSRS (suicide risk) ≥ *Moderate* at $\alpha = 0.05$. Error bars show the 95% confidence interval. Positive values indicate features where cases had higher mean values than controls. Numbers in parentheses indicate the number of included samples for cases and controls. The respective task/ interview question is specified in the prefix. LL, lower lip; JC, jaw center; MH, mouth half; acc, acceleration. **(A)** Effect sizes based on PHQ-9. **(B)** Effect sizes based on GAD-7. **(C)** Effect sizes based on C-SSRS.

Further, we found the models' performance sensitive to the number of features fed into the classifier. Therefore, we performed a follow-up analysis to determine the optimal number of features for classification performance based on the AUC. We found that a small set of speech features with only three features is most beneficial for classifying depression, and similarly, a small set of nine features from combined speech and facial modalities is useful for classifying anxiety. The performance increased to a maximum AUC of 0.8 and 0.79, respectively.

Speech and facial features selected across experiments (KW tests on the entire sample and selected features in each leave-one-speaker-out cross-validation fold) are shown in Table 7. As can be seen for depression, speech frequency and timing metrics were found to be discriminative across experiments. Percent pause time in speaking about fear as well as about secrets is significantly higher in cases than in controls, while the standard deviation of F0 is lower. These results are in agreement with the conducted Pearson correlation analysis that revealed a statistically significant positive correlation ($p = 0.025$, $r = 0.268$) between percent pause time (fear) and PHQ-9 total scores as well as a negative correlation (p: 0.003, r: −0.346) between standard deviation of F0 (fear) and PHQ-9 total scores.

Individuals with a GAD-7 score $\geq$ 10 reveal a different voice quality compared to controls expressed by a higher harmonics-to-noise ratio when speaking about hope, fear and secrets. Moreover, cases showed reduced movement and facial expression indicated by smaller lip aperture in the SIT task, slower lip movements in terms of velocity, acceleration and jerk measures, in particular while speaking about emotional pain and fear. In line with these findings, we found a statistically significant negative correlation between average absolute acceleration of the lower lip for the fear task and GAD-7 total scores (p: 0.007, r: −0.320). For the cohort with a suicidal risk $\geq$ *Moderate*, we observed a higher percent pause time while speaking about fear and hope than for controls. In addition, we detected reduced movement and facial expression for cases than controls captured by less eyebrow displacement when asked about anger, lip opening in the SIT task (p: 0.004, r: −0.332) and lower maximum downwards velocity of the lower lip when speaking about fear.

3.2. User feedback

3.2.1. Survey

Forty participants (59%) completed the five-question Likert scale survey about their experience, shown in Table 3. Frequency distributions, means, and standard deviations of the responses are shown in Figure 5. Student t-tests yielded no significant difference between case and control Likert scale ratings for all questions, for all conditions, except between suicide risk cases and controls for question four (p = 0.03), which asks about the virtual assistant's appearance and voice. Participants who scored "Moderate" risk or above on the C-SSRS Screener were more likely to rate Tina's visual appearance and voice higher compared to controls.

TABLE 5 Top 10 text model features with the associated mental state.

Mental state	Control features	Case features
Depression	Friends, little, certain, think, right, this, some, unable, suffer, across	Homeless, nice, being, paper, very, NAME, times, following, poetry, should
Anxiety	Right, year, well, family, theres, friends, best, depression, far, thought	Lot, anxious, heart, his, mad, parents, everyone, another, sensitive, worst
Suicide	School, family, money, having, cry, these, changes, loss, worry, point	Yeah, very, at, fear, her, one, whether, still, when, seem

3.2.2. Likes/advantages

The most frequent theme among the things that participants liked about the dialog agent was a "comfortable experience" (N = 43, 60%). One user reported: "I am very impressed by the realism of the experience. It felt almost as if I were talking to a real human being... the voice was pleasant and felt calming." The second-most recurrent theme was "accessibility" (N = 24, 32%). Another respondent stated: "There's value in screening for immediate risk when people aren't available." The third commonly occurring theme was "confidentiality" (N = 17, 22.7%), yet another participant commented: "I felt that I was able to be more open because it wasn't a real person; I didn't feel as though anyone was judging me."

3.2.3. Dislikes/disadvantages

The most common theme was "lack of human likeness" (N = 62, 82.7%). Users felt "awkward with the conversation flow" and were concerned about the virtual agent's "ability to understand nuances in someone's tone." The second theme was "perception of lack of risk intervention" (N = 3, 4%). One participant stated: "If somebody is in crisis, they wouldn't be caught in time to keep them safe."

3.2.4. Improvements

The most frequent theme was "interview flow" (N = 13, 17.3%). Users felt the pressure to speak for a certain time. A respondent conveyed: "I felt like I was grasping at straws trying to make up more things to say." The second-most occurring theme was "diversity in prompts" (N = 12, 16%). A user suggested to have "more specific questions based on responses." The third common theme was "different visual" (N = 9, 12%). One respondent recommended: "have an option on what kind of voice/face to interact with." Another user suggested: "it would be helpful to have an avatar that moved and blinked. It would feel less hollow."

4. Discussion

In this study, we explored the potential of using an MDS to collect speech, facial, and semantic text information to aid in the detection of depression, anxiety, and suicidal risk. Most participants indicated they honestly shared their feelings with the

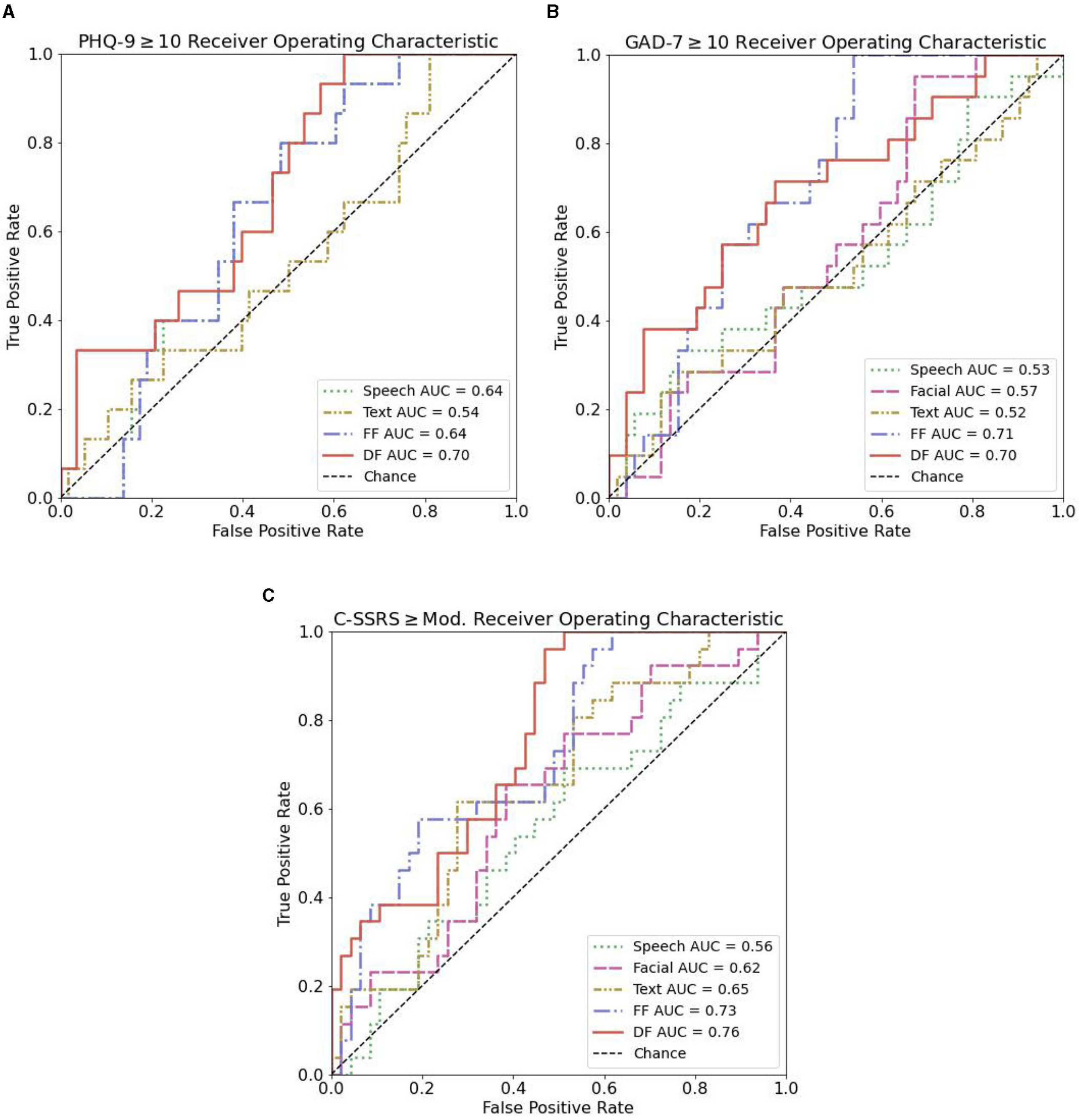


FIGURE 4
ROC curves for text, speech, facial, and the combination of speech and facial features in distinguishing controls from case participants. **(A)** ROC curves for PHQ-9 ≥ 10. **(B)** ROC curves for GAD-7 ≥ 10. **(C)** ROC curves for C-SSRS ≥ Mod.

virtual agent and found the experience comfortable, highlighting the potential acceptability of this approach. However, participants also identified areas for improvement in the conversational agent, such as the need for more contextually appropriate responses, indicating that further refinement of this methodology is needed.

While previous work has examined using MDSs with participants with depression and anxiety (DeVault et al., 2014; Cummins et al., 2015; Neumann et al., 2020), few studies have included individuals with an elevated suicide risk, which pose unique safety concerns. Indeed, some studies avoid any suicide-related questions and use the PHQ-8 (Kroenke et al., 2009) rather than the PHQ-9 (Kroenke et al., 2001), which skips the last question about suicide risk. In this study, participants received resources such as the 988 Suicide and Crisis Lifeline, and CRCs observed the interaction live and were able to follow the safety contingency plan if an acute risk arose. As several participants noted, there are limits to the degree a system such as this could immediately intervene. This is a valid concern and is true of any remote screening or patient monitoring system for suicide risk. Whether used clinically or in future studies where direct

TABLE 6 Evaluation metric scores for all conditions for best performing models.

Condition	Features	No. of features	AUC (95% CI)	Brier score
Depression	Speech	15	0.64 (0.51–0.78)	0.22
	Text	450	0.54 (0.37–0.70)	0.26
	Speech	7	0.53 (0.37–0.69)	0.23
Anxiety	Facial	24	0.57 (0.43–0.71)	0.23
	Text	80	0.52 (0.36–0.67)	0.30
	Speech	7	0.56 (0.42–0.70)	0.26
Su. Risk	Facial	24	0.62 (0.49–0.76)	0.25
	Text	54	0.65 (0.52–0.78)	0.27
Feature fusion (best performing combination)				
Depression	Speech+Text	15+450	0.64 (0.51–0.78)	0.22
Anxiety	All	7+54+80	0.71 (0.59–0.83)	0.22
Su. Risk	All	7+24+54	0.73 (0.61–0.85)	0.22
Decision fusion (best performing combination, min. scores)				
Depression	Speech+Text	15+450	0.70 (0.56–0.84)	0.16
Anxiety	All	7+24+80	0.70 (0.56–0.83)	0.20
Su. Risk	All	7+24+54	0.76 (0.65–0.87)	0.21

Note that we do not report AUC for the depression classification task with facial features because only one feature remained after feature selection, with the resulting AUC less than chance, suggesting that facial features are not as useful as other modalities for depression discrimination in this study cohort.

TABLE 7 Intersection of speech and facial features identified as statistically significant between respective cohorts for the entire sample and selected in every leave-one-speaker-out cross-validation fold.

Mental state	Features	Effect sizes	Categories
Depression	PPT (fear, secrets)	0.99, 0.75	Speech, timing
	F0 stdev. (fear)	−0.89	Speech, frequency
Anxiety	HNR (fear, hope, and secrets)	0.84, 0.78, 0.68	Speech, voice quality
	Max. jerk lower lip down (emotional pain)	0.5	Facial, movement
	Average speed lower lip (fear)	−0.94	Facial, movement
	Max. half mouth surface area right (SIT)	−0.93	Facial, mouth
	Average acc. lower lip (fear)	−0.92	Facial, movement
	Avg. jerk lower lip (fear)	−0.89	Facial, movement
	Max. lip aperture (SIT)	−0.81	Facial, mouth
	Max. mouth surface area (SIT)	−0.81	Facial, mouth
	Max. jerk lower lip up (fear, emotional pain)	−0.62, −0.48	Facial, movement
	Max. velocity lower lip up (fear)	−0.61	Facial, movement
	Abs. max. jerk lower lip (emotional pain)	−0.57	Facial, movement
	Max. acc. lower lip down (emotional pain)	−0.53	Facial, movement
	Abs. max. acc. lower lip (emotional pain)	−0.48	Facial, movement
Suicide	PPT (hope and fear)	0.71, 0.55	Speech, timing
	Max. velocity lower lip down (fear)	0.61	Facial, movement
	Avg. lip aperture (SIT)	−0.77	Facial, mouth
	Avg. half mouth surface area right (SIT)	−0.57	Facial, mouth
	Avg. eyebrow displacement (anger)	−0.38	Facial, eyes

The respective interview question or task is shown in parentheses.

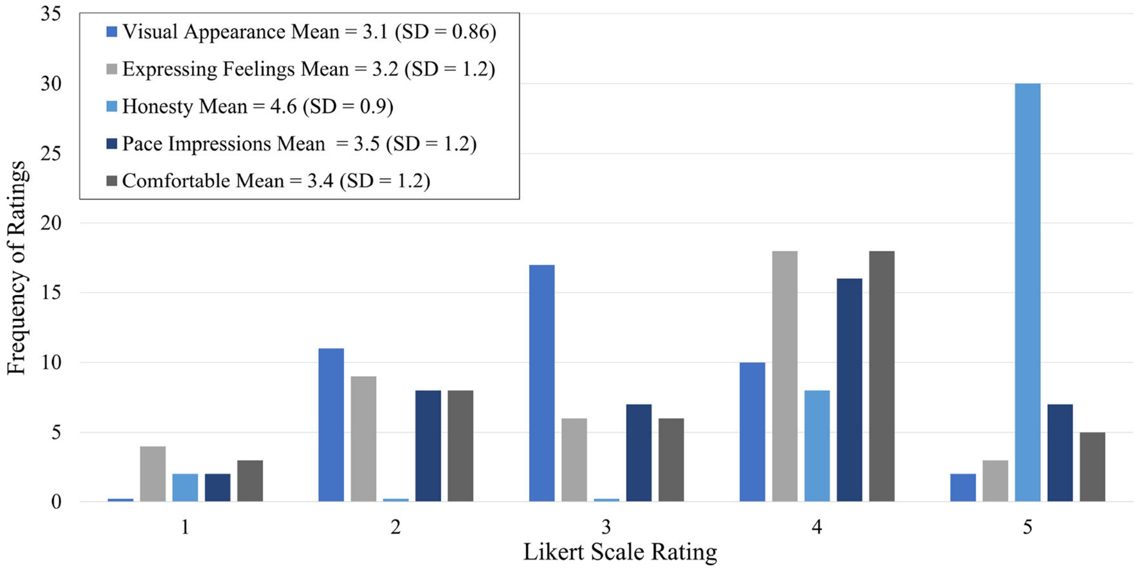


FIGURE 5
Survey distribution.

observation is not possible, safeguards will need to be in place to address this concern.

During feature analysis, we found many features agreed with previous clinical mental health ML studies. Our findings for depression are in line with other work that found a higher PPT (Cannizzaro et al., 2004; Mundt et al., 2012; Bennabi et al., 2013) or an increase in total pause time associated with lower total speech duration (Albuquerque et al., 2021) for individuals with depression, as well as a correlation between clinical condition and standard deviation of F0 and PPT (Åsa Nilsonne, 1987). The lower standard deviations of the F0 for cases with a PHQ-9 $\geq$ 10 compared to controls indicate less variation in speech.

In agreement with our findings, previous studies have reported different voice quality in individuals with anxiety disorder compared with control participants, measured as harmonics-to-noise ratio (Murray and Arnott, 1993; Siegman and Boyle, 1993). The findings, however, show an irregular trend (increase vs. decrease). Moreover, as suggested by our results, a lower HNR score in controls seems to be counter-intuitive, as a low HNR is associated with a higher degree of hoarseness (Yumoto et al., 1982), which refers to abnormal voice quality (Feierabend and Shahram, 2009). Anxiety may be more intensely manifested in facial features, as our study suggests that (a) the facial modality performs better than speech and language features in classifying cases with anxiety disorder vs. controls and (b) twelve facial features are consistently selected across experiments compared to only one speech feature. Our results indicate that adults affected by this disorder show reduced facial behavior. However, anxiety disorders and especially facial features on this are understudied (Low et al., 2019), and more research is needed to investigate multimodal markers of the disease.

As in depression, a higher PPT for individuals with suicidal risk $\geq$ *Moderate* has been observed, which was also shown in clinician-patient interaction (Venek et al., 2015). Regarding the PPT, the largest effect between cases and controls is found when talking about hope, as can be seen in Table 7, suggesting individuals at moderate or high suicidal risk struggle more with this topic. In addition, decreased facial activity, as evidenced by lower eyebrow displacement and mouth opening in our study, has been associated with higher suicide risk in previous research studies (Galatzer-Levy et al., 2020). Cases show a higher downward, but not upward, velocity of the lower lip compared to controls, which may be interpreted as a more abrupt opening of the mouth compared to controls. However, future investigations are needed to provide a more thorough understanding of the observed behaviors.

The text features shown in Table 5 are the top 10 case and control features by weight of linear SVMs fit to the entire dataset for each condition, and represent a fraction of the total number of features. A full linguistic analysis is out of scope here, however, there are some noteworthy observations. First, other studies have found personal pronouns related to depression and suicide risk (Chung and Pennebaker, 2007), yet no personal pronouns appear in Table 5. For depression, the appearance of a name as a case feature is likely related to the limited number of depression cases in this study. For anxiety, the word “anxious” appears as a top feature for cases, while “depression” appears as a control feature. Interestingly, for suicide, some of the control features could be associated with stressors or protective factors related to suicide risk, depending on context. The use of n-grams (contiguous sequence of n nummber of words) or more advanced NLP techniques, such as word embeddings (Mikolov et al., 2013; Pennington et al., 2014), could capture more nuanced aspects of language. In a clinical setting, tools such as Local Interpretable Model-Agnostic Explanations (LIME; Ribeiro et al., 2016) could improve the interpretability of text features by considering their impact per prediction, as opposed to globally, and displaying the features within the context they were used.

For the classification tasks, we found that the combination of modalities typically improved model performance for all tasks. This is not altogether surprising as one might expect more information to help classification performance. However, an exception to this was observed for the depression classification task with facial features, which was not run due to the low number of significant features. It is likely that facial features would improve model performance with a larger, more balanced sample.

It is also worth noting that not all our single-modality models performed above the chance level, with five out of eight delivering near-chance results. We elected to report these lower-performing models as they are are also informative and contribute to a comprehensive understanding of the dataset and the performance characteristics of the different modalities. This approach underscores the importance of having a sufficient number of significant features for reliable classification performance.

Text features performed the best when identifying suicide risk and performed with near chance levels for the identification of depression and anxiety. The interview questions of the MHSAFE interview (hopes, secrets, anger, fear, and emotional pain) were originally developed to screen for suicide risk (Pestian et al., 2010), therefore, classification performance for depression and anxiety may improve if questions more relevant for those conditions are added. The nature of the interaction may also have influenced the semantic content shared by participants which may have affected classifier performance, as some indicated a pressure to speak long enough to fill the required amount of time.

As the prevalence of mental health conditions increases amidst greater health system strain, digital approaches to screen and monitor these conditions are emerging as promising avenues for research. Our interviews, which on average took <10 min, suggest the possibility of providing clinically useful information for three conditions, given that models have been appropriately validated. However, these are early findings and further research is needed to confirm and expand upon our results. The results of the interview could potentially offer a new perspective as clinical decision support for difficult cases, or direct appropriate resources or referrals to individuals when a mental health professional is not available.

4.1. Limitations and future directions

Although these findings align with the earlier-discussed studies with regards to the identification of important features and general model discriminative ability, some limitations should be noted. First, studies with small sample sizes face inherent limitations, such as limited representation of different genders and races, which may impact generalizability. Additionally, small sample size ML studies may lead to overly optimistic estimates of classification performance as it is difficult to eliminate information leakage across folds when both train and test sets are used for feature selection (Vabalas et al., 2019; Berisha et al., 2022). Our method to determine the number of features to include during the classification tasks was based on the number of features identified as statistically significant when fit on the entire dataset. Therefore, we acknowledge some information leakage across the folds, however, the specific features selected were determined during each CV fold, and as seen in Table 7, only a fraction of the statistically significant features from the entire dataset (Figure 3) appear in all of the CV folds. While this technique may have lead to more reasonable estimates of model performance, we intend to repeat our analysis with a larger sample size in future work and ultimately explore more advanced modeling techniques, including deep learning. Note that we did not explore deep learning methods in this work, for two important reasons—the primary one being the need to clearly interpret the results/performance of the system in order to be practically applicable in the healthcare setting, and the second being the limited sample size. Lastly, we tried oversampling techniques to account for our dataset's case imbalance, but did not see any improvements; we will continue to explore these techniques with a larger dataset.

The supervision of participants by CRCs during this study may have influenced participant responses. Previous research indicates that participants interacting with a computer reported lower fear of self-disclosure and displayed more intense sadness than when they believed they were interacting with a human (Gratch et al., 2007; Lucas et al., 2014; Rizzo et al., 2016). In our study, some participants even pointed out the potential advantage of system confidentiality, and none expressed negative feedback regarding the presence of CRCs. The identification of features consistent with the literature and the discriminative ability of the classifiers suggest that most participants expressed themselves at least as openly as in studies involving human interviewers. In future studies, we aim to remove direct CRC supervision to better reflect real-world scenarios of remote patient monitoring and to possibly elicit more authentic user responses. By doing so, we also hope to facilitate the collection of larger datasets, crucial for overcoming common machine learning challenges such as overfitting, generalizability, and bias.

Participants indicated several areas of improvement in the user feedback section that we have implemented and will test in future studies. First, we have added slight animation of the virtual agent with the aim of increasing human likeness. To improve the flow of the interview and aid in prompting participants, we have reduced the minimum amount of time required for each response to 30 s and included nudges specific to each question of the MHSAFE interview.

5. Conclusions

This study found that a multimodal dialog system (MDS) is a feasible, scalable, and interpretable solution for remote patient monitoring (RPM) in real-world clinical depression, anxiety and suicidal populations. A novelty of this study is that it investigates features derived from multiple modalities—speech, language, and facial behavior—to analyze and characterize three mental disorders—depression, anxiety, and suicide risk—simultaneously. An interesting finding to highlight here is that different modalities were found to be most effective at distinguishing controls from cases for each disorder considered: speech for depression, facial for anxiety, and text/language for suicidality. We also found that a combination of features from different modalities extracted during a brief, standardized MDS interview generally improved the discriminative ability of machine learning models for mental state characterization in *all three disorders*. Furthermore, both healthy participants and those affected by a mental disorder indicated acceptance of the technology. Finally, we presented several lessons learned from implementation, user experience, feature engineering and machine learning perspectives for future practitioners.

Data availability statement

The datasets presented in this article are not readily available because the dataset contains confidential health-related data that cannot be shared. These data will be made available for research purposes only to any researcher who meet criteria for access to confidential data based on relevant Institutional Review Boards. Requests to access the datasets should be directed to research@clarigenthealth.com.

Ethics statement

The studies involving humans were approved by Advarra's Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JC, VRi, MN, and VRa conceptualized the study and wrote the manuscript. VRi, MN, and DB performed feature analysis and model development/validation. AH and JW-B performed analysis of survey responses. JC and JW-B are principal investigators of the study Classification and Assessment of Mental Health Performance Using Semantics—Expanded. All authors contributed to the article and approved the submitted version.

Funding

This study was funded by Clarigent Health. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article or the decision to submit it for publication.

Acknowledgments

The authors thank Lesley Rohlfs, Devyn Thompson, and Marci Campbell for their support with study development and implementation. We also thank David Pautler, Andrew Cornish, Doug Habberstad, Hardik Kothare, Jackson Liscombe, and Oliver Roesler for their help in adapting the MDS system for this study. We further thank the participants of the study for their time and the community of scientists and researchers that provide open-source machine learning tools and education resources.

Conflict of interest

JC, DB, and AH were employed full-time by Clarigent Health. VRi, MN, and VRa were full-time employees of Modality.AI. JW-B is a part-time consultant and is sponsored by a grant from Clarigent Health. The above interests do not alter our adherence to Frontiers Media's policies. Clarigent Health and Modality did not influence or restrict the submission of this publication.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2023.1135469/full#supplementary-material

References

Albuquerque, L., Valente, R., Teixeira, A., Figueiredo, D., Sa-Couto, P., and Oliveira, C. (2021). Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PLoS ONE* 16, e0248842. doi: 10.1371/journal.pone.0248842

Arif, M., Basri, A., Melibari, G., Sindi, T., Alghamdi, N., Altalhi, N., et al. (2020). Classification of anxiety disorders using machine learning methods: a literature review. *Insights Biomed. Res.* 4, 95–110. doi: 10.36959/584/455

Åsa Nilsonne (1987). Acoustic analysis of speech variables during depression and after improvement. *Acta Psychiatr. Scand.* 76, tb02891. doi: 10.1111/j.1600-0447.1987.tb02891.x

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., and Grundmann, M. (2019). Blazeface: sub-millisecond neural face detection on mobile GPUs. *CoRR* abs/1907.05047.

Bennabi, D., Vandel, P., Papaxanthis, C., Pozzo, T., and Haffen, E. (2013). Psychomotor retardation in depression: a systematic review of diagnostic, pathophysiologic, and therapeutic implications. *BioMed Res. Int.* 2013, 158746. doi: 10.1155/2013/158746

Berisha, V., Krantsevich, C., Stegmann, G., Hahn, S., and Liss, J. (2022). "Are reported accuracies in the clinical speech machine learning literature overoptimistic?," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Vol. 2022* (Incheon), 2453–2457.

Bernert, R. A., Hilberg, A. M., Melia, R., Kim, J. P., Shah, N. H., and Abnousi, F. (2020). Artificial intelligence and suicide prevention: a systematic review of machine learning investigations. *Int. J. Environ. Res. Publ. Health* 17, 5929. doi: 10.3390/ijerph17165929

Boersma, P., and Van Heuven, V. (2001). Speak and unspeak with praat. *Glot. Int.* 5, 341–347.

Botelho, C., Schultz, T., Abad, A., and Trancoso, I. (2022). "Challenges of using longitudinal and cross-domain corpora on studies of pathological speech," in *Proc. Interspeech* (Incheon), 1921–1925.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., and Snyder, P. (2004). Voice acoustical measurement of the severity of major depression. *Brain Cogn.* 56, 30–35. doi: 10.1016/j.bandc.2004.05.003

Chung, C., and Pennebaker, J. W. (2007). The psychological functions of function words. *Soc. Commun.* 1, 343–359.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences, 2nd Edn)*. London: Routledge.

Cohen, J., Wright-Berryman, J., Rohlfs, L., Trocinski, D., Daniel, L., and Klatt, T. W. (2022). Integration and validation of a natural language processing machine learning suicide risk prediction model based on open-ended interview language in the emergency department. *Front. Digit. Health* 4, 818705. doi: 10.3389/fdgth.2022.818705

Cohen, J., Wright-Berryman, J., Rohlfs, L., Wright, D., Campbell, M., Gingrich, D., et al. (2020). A feasibility study using a machine learning suicide risk prediction model based on open-ended interview language in adolescent therapy sessions. *Int. J. Environ. Res. Public Health* 17, 21. doi: 10.3390/ijerph17218187

Cummins, N., Baird, A., and Schuller, B. W. (2018). Speech analysis for health: current state-of-the-art and the increasing impact of deep learning. *Methods* 151, 41–54. doi: 10.1016/j.ymeth.2018.07.007

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., and Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Commun.* 71, 10–49. doi: 10.1016/j.specom.2015.03.004

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., et al. (2014). "Simsensei kiosk: a virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 international Conference on Autonomous Agents and Multi-agent Systems* (Paris: International Foundation for Autonomous Agents and Multiagent Systems), 1061–1068.

Dimberg, U., and Lundquist, L.-O. (1990). Gender differences in facial reactions to facial expressions. *Biol. Psychol.* 30, 151–159. doi: 10.1016/0301-0511(90)90024-Q

Drimalla, H., Scheffer, T., Landwehr, N., Baskow, I., Roepke, S., Behnia, B., et al. (2020). Towards the automatic detection of social biomarkers in autism spectrum disorder: introducing the simulated interaction task (SIT). *NPJ Digit. Med.* 3, 25. doi: 10.1038/s41746-020-0227-5

Feierabend, R. H., and Shahram, M. N. (2009). Hoarseness in adults. *Am. Fam. Phys.* 80, 363–370.

France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., and Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Trans. Biomed. Eng.* 47, 829–837. doi: 10.1109/10.846676

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., et al. (2017). Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol. Bullet.* 143, 187. doi: 10.1037/bul0000084

Galatzer-Levy, I. R., Abbas, A., Ries, A., Homan, S., Sels, L., Koesmahargyo, V., et al. (2020). Validation of visual and auditory digital markers of suicidality in acutely suicidal psychiatric inpatients: proof-of-concept study. *J. Med. Internet Res.* 23, 25199. doi: 10.2196/preprints.25199

Gratch, J., Lucas, G. M., King, A. A., and Morency, L.-P. (2014). "It's only a computer: the impact of human-agent interaction in clinical interviews," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems* (Paris: International Foundation for Autonomous Agents and Multiagent Systems), 85–92.

Gratch, J., Wang, N., Okhmatovskaia, A., Lamothe, F., Morales, M., van der Werf, R. J., et al. (2007). "Can virtual humans be more engaging than real ones?," in *International Conference on Human-Computer Interaction* (Berlin: Springer), 286–297.

Hunter, J. (2007). Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* 9, 90–95. doi: 10.1109/MCSE.2007.55

Kartynnik, Y., Ablavatski, A., Grishchenko, I., and Grundmann, M. (2019). Real-time facial surface geometry from monocular video on mobile GPUs. *CoRR* abs/1907.06724.

Kothare, H., Neumann, M., Liscombe, J., Roesler, O., Burke, W., Exner, A., et al. (2022). "Statistical and clinical utility of multimodal dialogue-based speech and facial metrics for Parkinson's disease assessment," in *Proc. Interspeech 2022* (Incheon), 3658–3662.

Kothare, H., Ramanarayanan, V., Roesler, O., Neumann, M., Liscombe, J., Burke, W., et al. (2021). "Investigating the interplay between affective, phonatory and motoric subsystems in autism spectrum disorder using a multimodal dialogue agent," in *Proceedings of the 22nd Annual Conference of the International Speech Communication Association* (Brno: Interspeech).

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *J. Gen. Internal Med.* 16, 606–613. doi: 10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disord.* 114, 163–173. doi: 10.1016/j.jad.2008.06.026

Kusuma, K., Larsen, M., Quiroz, J. C., Gillies, M., Burnett, A., Qian, J., et al. (2022). The performance of machine learning models in predicting suicidal ideation, attempts, and deaths: a meta-analysis and systematic review. *J. Psychiatr. Res.* 9, 50. doi: 10.1016/j.jpsychires.2022.09.050

Laksana, E., Baltrušaitis, T., Morency, L.-P., and Pestian, J. P. (2017). "Investigating facial behavior indicators of suicidal ideation," in *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)* (Washington, DC: IEEE), 770–777.

Lisetti, C., Amini, R., and Yasavur, U. (2015). Now all together: overview of virtual health assistants emulating face-to-face health interview experience. *KI-Künstliche Intelligenz* 29, 161–172. doi: 10.1007/s13218-015-0357-0

Low, D. M., Bentley, K. H., and Ghosh, S. S. (2019). Automated assessment of psychiatric disorders using speech: a systematic review. *Laryngos. Investig. Otolaryngol.* 5, 96–116. doi: 10.1002/lio2.354

Lucas, G. M., Gratch, J., King, A., and Morency, L.-P. (2014). It's only a computer: virtual humans increase willingness to disclose. *Comput. Hum. Behav.* 37, 94–100. doi: 10.1016/j.chb.2014.04.043

McKight, P. E., and Najab, J. (2010). Kruskal-wallis test. *Corsini Encycl. Psychol.* 2010, 1. doi: 10.1002/9780470479216.corpsy0491

McKinney, W. (2010). "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference, Vol. 445* (Austin, TX), 51–56.

Mendoza, E., Valencia, N., Muñoz, J., and Trujillo, H. (1996). Differences in voice quality between men and women: use of the long-term average spectrum (LTAS). *J. Voice* 10, 59–66. doi: 10.1016/S0892-1997(96)80019-1

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. doi: 10.48550/arXiv.1301.3781

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., and Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J. Neurolinguist.* 20, 50–64. doi: 10.1016/j.jneuroling.2006.04.001

Mundt, J. C., Vogel, A. P., Feltner, D. E., and Lenderking, W. R. (2012). Vocal acoustic biomarkers of depression severity and treatment response. *Biol. Psychiatry* 72, 580–587. doi: 10.1016/j.biopsych.2012.03.015

Murray, I. R., and Arnott, J. L. (1993). Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion. *J. Acoust. Soc. Am.* 93, 1097–1108. doi: 10.1121/1.405558

Neumann, M., Roesler, O., Liscombe, J., Kothare, H., Suendermann-Oeft, D., Pautler, D., et al. (2021). "Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association* (Brno: ISCA), 4783–4787.

Neumann, M., Roessler, O., Suendermann-Oeft, D., and Ramanarayanan, V. (2020). "On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers," in *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, 47–52.

Oliphant, T. E. (2007). Python for scientific computing. *Comput. Sci. Eng.* 9, 10–20. doi: 10.1109/MCSE.2007.58

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Machine Learn. Res.* 12, 2825–2830.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha), 1532–1543.

Pestian, J. (2010). A conversation with edwin shneidman. *Suicide Life-Threat. Behav.* 40, 516–523. doi: 10.1521/suli.2010.40.5.516

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., and Leenaars, A. (2010). Suicide note classification using natural language processing: a content analysis. *Biomed. Informat. Insights* 3, BII.S4706. doi: 10.4137/BII.S4706

Pestian, J. P., Grupp-Phelan, J., Cohen, K. B., Meyers, G., Richey, L. A., Matykiewicz, P., et al. (2016). A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. *Suicide Life-Threat. Behav.* 46, 154–159. doi: 10.1111/sltb.12180

Pestian, J. P., Sorter, M., Connolly, B., Cohen, K. B., McCullumsmith, C., Gee, J. T., et al. (2017). A machine learning approach to identifying the thought markers of suicidal subjects: a prospective multicenter trial. *Suicide Life-Threat. Behav.* 47, 112–121. doi: 10.1111/sltb.12312

Posner, K., Brown, G. K., Stanley, B., Brent, D. A., Yershova, K. V., Oquendo, M. A., etal. (2011). The columbia-suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *Am. J. Psychiatry* 168, 1266–1277. doi: 10.1176/appi.ajp.2011.10111704

Rajaraman, A., and Ullman, J. D. (2011). *Data Mining* (Cambridge: Cambridge University Press), 1–17.

Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F., and Green, J. R. (2022). Speech as a biomarker: opportunities, interpretability, and challenges. *Perspect. ASHA Spec. Interest Groups* 7, 276–283. doi: 10.1044/2021_PERSP-21-00174

Ramanarayanan, V., Roesler, O., Neumann, M., Pautler, D., Habberstad, D., Cornish, A., et al. (2020). "Toward remote patient monitoring of speech, video, cognitive and respiratory biomarkers using multimodal dialog technology," in *INTERSPEECH* (Shanghai), 492–493.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ""why should I trust you?": explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. San Francisco, 1135–1144.

Rice, M. E., and Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's D, and R. *Law Hum. Behav.* 29, 615–620. doi: 10.1007/s10979-005-6832-7

Richter, V., Neumann, M., Kothare, H., Roesler, O., Liscombe, J., Suendermann-Oeft, D., et al. (2022). "Towards multimodal dialog-based speech and facial biomarkers of schizophrenia," in *Companion Publication of the 2022 International Conference on Multimodal Interaction* (Bangalore), 171–176.

Rizzo, A., Lucas, G., Gratch, J., Stratou, G., Morency, L., Shilling, R., et al. (2016). "Clinical interviewing by a virtual human agent with automatic behavior analysis," in *2016 Proceedings of the International Conference on Disability, Virtual Reality and Associated Technologies* (Reading: University of Reading), 57–64.

Roesler, O., Kothare, H., Burke, W., Neumann, M., Liscombe, J., Cornish, A., et al. (2022). "Exploring facial metric normalization for within- and between-subject comparisons in a multimodal health monitoring agent," in *Companion Publication of the 2022 International Conference on Multimodal Interaction, ICMI '22 Companion* (New York, NY: Association for Computing Machinery), 160–165.

Satiani, A., Niedermier, J., Satiani, B., and Svendsen, D. P. (2018). Projected workforce of psychiatrists in the united states: a population analysis. *Psychiatr. Serv.* 69, 710–713. doi: 10.1176/appi.ps.201700344

Siegman, A. W., and Boyle, S. (1993). Voices of fear and anxiety and sadness and depression: the effects of speech rate and loudness on fear and anxiety and sadness and depression. *J. Abnorm. Psychol.* 102, 430–437.

Simpson, A. (2009). Phonetic differences between male and female speech. *Lang. Linguist. Compass* 3, 621–640. doi: 10.1111/j.1749-818X.2009.00125.x

Spitzer, R. L., Kroenke, K., Williams, J. B. W., and Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archiv. Internal Med.* 166, 1092–1097. doi: 10.1001/archinte.166.10.1092

Suendermann-Oeft, D., Robinson, A., Cornish, A., Habberstad, D., Pautler, D., Schnelle-Walka, D., et al. (2019). "NEMSI: a multimodal dialog system for screening of neurological or mental conditions," in *Proceedings of ACM International Conference on Intelligent Virtual Agents (IVA)*. Paris.

The Pandas Development Team (2020). *Pandas-dev/Pandas: Pandas*.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *J. Acousti. Soc. Am.* 85, 1699–1707. doi: 10.1121/1.397959

Torchiano, M. (2020). *effsize: Efficient Effect Size Computation. R Package Version 0.7.6*.

Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE* 14, 1–20. doi: 10.1371/journal.pone.0224365

Van Der Walt, S., Colbert, S. C., and Varoquaux, G. (2011). The numpy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* 13, 22–30. doi: 10.1109/MCSE.2011.37

Van Rossum, G., and Drake, F. L. (1995). *Python Tutorial, Vol. 620*. Amsterdam: Centrum voor Wiskunde en Informatica.

Venek, V., Scherer, S., Morency, L.-P., Rizzo, A., and Pestian, J. (2015). "Adolescent suicidal risk assessment in clinician-patient interaction: a study of verbal and acoustic behaviors," in *2014 IEEE Workshop on Spoken Language Technology, SLT 2014—Proceedings* (South Lake Tahoe, CA: IEEE), 277–282.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-020-0772-5

World Health Organization (2021). *Suicide Worldwide in 2019: Global Health Estimates*. Geneva: World Health Organization.

World Health Organization (2022). *Mental Health and COVID-19: Early Evidence of the Pandemic's Impact. Technical Report*. Geneva: World Health Organization.

Wright-Berryman, J., Cohen, J., Haq, A., Black, D. P., and Pease, J. L. (2023). Virtually screening adults for depression, anxiety, and suicide risk using machine learning and language from an open-ended interview. *Front. Psychiatry* 14, 1143175. doi: 10.3389/fpsyt.2023.1143175

Youngstrom, E. A. (2013). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: we are ready to ROC. *J. Pediatr. Psychol.* 39, 204–221. doi: 10.1093/jpepsy/jst062

Yumoto, E., Gould, W., and Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *J. Acoust. Soc. Am.* 71, 1544–1549.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 27 February 2024
DOI 10.3389/fpsyg.2024.1155285

OPEN ACCESS

EDITED BY
Laura M. Morett,
University of Missouri, United States

REVIEWED BY
Korbinian Riedhammer,
Technische Hochschule Nürnberg Georg Simon Ohm, Germany
Torrey Loucks,
Jacksonville University, United States

*CORRESPONDENCE
Peter Howell
✉ p.howell@ucl.ac.uk

RECEIVED 31 January 2023
ACCEPTED 07 February 2024
PUBLISHED 27 February 2024

CITATION
Barrett L, Tang K and Howell P (2024) Comparison of performance of automatic recognizers for stutters in speech trained with event or interval markers.
Front. Psychol. 15:1155285.
doi: 10.3389/fpsyg.2024.1155285

Comparison of performance of automatic recognizers for stutters in speech trained with event or interval markers

Liam Barrett[1], Kevin Tang[2,3] and Peter Howell[1]*

[1]Department of Experimental Psychology, University College London, London, United Kingdom, [2]Department of English Language and Linguistics, Institute of English and American Studies, Faculty of Arts and Humanities, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, [3]Department of Linguistics, University of Florida, Gainesville, FL, United States

Introduction: Automatic recognition of stutters (ARS) from speech recordings can facilitate objective assessment and intervention for people who stutter. However, the performance of ARS systems may depend on how the speech data are segmented and labelled for training and testing. This study compared two segmentation methods: event-based, which delimits speech segments by their fluency status, and interval-based, which uses fixed-length segments regardless of fluency.

Methods: Machine learning models were trained and evaluated on interval-based and event-based stuttered speech corpora. The models used acoustic and linguistic features extracted from the speech signal and the transcriptions generated by a state-of-the-art automatic speech recognition system.

Results: The results showed that event-based segmentation led to better ARS performance than interval-based segmentation, as measured by the area under the curve (AUC) of the receiver operating characteristic. The results suggest differences in the quality and quantity of the data because of segmentation method. The inclusion of linguistic features improved the detection of whole-word repetitions, but not other types of stutters.

Discussion: The findings suggest that event-based segmentation is more suitable for ARS than interval-based segmentation, as it preserves the exact boundaries and types of stutters. The linguistic features provide useful information for separating supra-lexical disfluencies from fluent speech but may not capture the acoustic characteristics of stutters. Future work should explore more robust and diverse features, as well as larger and more representative datasets, for developing effective ARS systems.

KEYWORDS

stuttering, speech pathology, automatic speech recognition, machine learning, computational paralinguistics, language diversity, language model, whisper

1 Introduction

Human assessment of stuttering is time consuming and even trained observers give variable scores for the same materials (Kully and Boberg, 1988). If automatic recognition of stuttering (ARS) met acceptable performance standards, these assessments would save time and could standardize score reports. Practical applications other than reducing workload in clinics, include ease of inter-clinic comparisons and making voice-controlled online applications accessible to people who stutter, PWS (Barrett et al., 2022). Given these desirable goals, ARS work began in the late 1990s (Howell et al., 1997a,b). Initial progress was limited because few labs had appropriate training material. Matters improved after the release of the first online audio database of stuttered speech, the University College London Archive of Stuttered Speech (UCLASS) (Howell et al., 2009). UCLASS has time markers indicating where stutters start, and end and the types of stutters are coded (an example of an *'event-based'* procedure). Each event-based segment varies in duration. Databases that have been established subsequently have segmented speech into fixed length intervals (usually 3 s; 3-s) referred to here as *'interval-based'* procedures (Lea et al., 2021; Bayerl et al., 2022a). They provide labels for each interval ('fluent' or 'stuttered' or, in some cases, 'fluent' and the specific type of stutter). Whereas fluent intervals are fluent throughout, stuttered intervals with or without symptom-type annotations may not be delimited to these intervals and, when stutters are less than 3-s long, contain some fluent speech and, in some cases, additional stuttered symptoms (Howell et al., 1998). Generally speaking, the fact that intervals are made ambiguous with respect to fluency designation when a fixed duration is imposed onto speech segments, limits overall recognition accuracy of ARS. Surprisingly however, no comparison checks have been made between interval and event-based procedures to verify or disconfirm this prediction.

Additionally, interval-based methods usually report poor performance with respect to whole-word repetitions (WWR) (Lea et al., 2021; Bayerl et al., 2022a). In fact, this might be a correct outcome since there is a wider debate about whether WWR are indeed stutters (Howell, 2010) and because repetition of each constituent word has all phones and these are in their correct positions (implying each word is produced fluently). Whether or not WWR are stutters, it would be difficult to separate them from the same words in fluent speech for procedures that use short-window, acoustic inputs because the segments (events or intervals) may not extend long enough to include any repeated words. In summary, recognition of WWRs and separation of them from fluent speech may be improved if ARS procedures are trained on intervals long enough to include lexical and supra-lexical features (e.g., n-grams for spotting multi-phone repetition representing word and phrase repetition).

Identifying stuttering events may be more appropriate than identifying intervals since stuttering events dominate in clinical and research reports. For example, Stuttering Severity Index (SSI) measures (Riley, 2009) that is partly based on symptoms are always reported in research publications whereas reports that use intervals are rare (Ingham et al., 1993). Additionally, there seems to be little justification as to why a duration of 3-s was chosen as the interval-length other than saving assessment time [see Ingham et al. (1993) for rationale and Howell et al. (1998) for evaluation]. To validate whether 3-s intervals are the preferred length for best ARS model performance, intervals of 2-s and 4-s were also investigated in our study.

The UCLASS database and the Kassel State of Fluency (KSoF) 3-s interval dataset (Bayerl et al., 2022a) were used in the investigation. UCLASS data were also reformatted into the 3-s interval format (intervals of 2-s and 4-s were also computed). UCLASS and KSoF interval data were each used to train and test a shallow (Gaussian support vector machine) and a deep (multi-layered perceptron neural network) machine learning model to establish whether model performance was equivalent for the two datasets. The shallow and deep learning models were then used to determine how model performance was affected by segmentation method for the same (UCLASS) data.

We hypothesized that the distribution of speech types (stutters and fluent speech) should be similar across KSoF, and UCLASS, interval datasets. Also, models trained using these datasets should perform similarly. If these predictions hold, they confirm that UCLASS and KSoF interval data are comparable and validate the subsequent interval-event comparisons made using UCLASS data alone.

Second, performance was compared for models trained on interval-based, or event-based, UCLASS data. It was predicted that models trained using the event-based format would outperform the interval-based models because only the former delimits speech extracts exclusively associated with their fluency types. Area under the curve for the receiver operating characteristic (AUC-ROC) was used as the performance indicator.

Third, the distribution of fluency types for 2-s, 3-s, and 4-s intervals and the effects of using these different-length intervals on model performance were assessed. It was predicted that using shorter interval lengths should lead to a greater proportion of fluent speech intervals relative to disfluent intervals.

Fourth, the model inputs for the 3-s interval and event-based models were switched to investigate whether the features transferred across segmentation formats. Specifically, after training a model on features derived from the 3-s subset of UCLASS, the model was tested on features derived from the event-based subset and *vice-versa*. By switching the feature inputs, this tested whether the parameters learned by one method transferred to the other. This is the first time such a method has been used to investigate whether a trained ARS model is robust to changes in the feature extraction process. A difference was hypothesized in model performance due to this switch, but no direction was hypothesized.

Finally, we compared models with and without language-based features. We hypothesized that the inclusion of language-based features: (i) could lead to better recognition of WWRs, and possibly of fluent speech; (ii) longer intervals should perform better than shorter intervals because of the increased chance that the language-based features identify whole-word repetitions; (iii) using interval data should improve performance over using event data for these models because intervals usually have more scope for including supra-segmental features. Together these experiments should afford a clear and direct comparison of the effects of using the two training-material types on model performance.

2 Method

2.1 Datasets

UCLASS data are in British English that includes 249 speakers (Howell et al., 2009) of which, audio from 14 speakers with approximately 180 min of valid and labelled speech were used for the current study. The KSoF data are in German from 37 speakers with approximately 230 min of valid and labelled speech (Bayerl et al., 2022a). In both datasets, a hard split was used to keep the speakers in the training, validation and test splits distinct. Hard, separate speaker, splits are key to evaluating models in the ARS field (Bayerl et al., 2022c). As such, the UCLASS data were split into nine speakers for Training, two speakers for Validation and three speakers for Test.

TABLE 1 Absolute and relative frequencies of the five classes of speech fluency in the UCLASS event, UCLASS 3-s interval and KSoF 3-s interval subsets.

Type	UCLASS event		UCLASS 3-s interval		KSoF 3-s interval	
	Absolute frequency	Relative frequency (%)	Absolute frequency	Relative frequency (%)	Absolute frequency	Relative frequency (%)
Fluent	11,837	82.48	1,228	37.39	1,538	52.91
Prolongation	396	2.76	383	12.69	346	11.90
PWR	469	3.27	733	24.30	339	11.66
WWR	173	1.20	44	1.46	94	3.23
Block	1,476	10.29	729	24.16	590	20.30
Total	14,351	100.00	3,117	100.00	2,907	100.00

For KSoF, 23, six and eight speakers were assigned to Training, Validation and Test, respectively.

2.1.1 The UCLASS dataset

The UCLASS data used has transcriptions at word and syllabic levels aligned against the audio recordings. Annotations also separate fluent speech, prolongations, part-word repetitions (PWR), WWR and blocks. These UCLASS data were segmented at both the event-based and interval-based levels as described below. The breakdown of observations per split varied by segmentation method (Table 1).

2.1.1.1 UCLASS event-based subset

Speech can be annotated at different levels of precision with the event-based method. Here, syllables were the defined event. Applying the event-based scheme to UCLASS yielded 14,351 unique annotations, each of which had a single, valid label.

2.1.1.2 UCLASS interval-based subset

The interval-based format that applies annotations to fixed, 3-s intervals of speech was the main focus in comparisons with event-based methods since this is the only interval length used for ARS to date (Lea et al., 2021; Bayerl et al., 2022a). The continuous UCLASS speech recordings were split automatically into 3-s intervals and their corresponding transcriptions were examined to identify candidate intervals and their type. The interval designation scheme used by Bayerl et al. (2022a) was applied and generated 3,984 intervals. Of these, 3,117 had a single type of stutter or were fluent throughout (valid labels) and 867 were dropped which had either multiple disfluency types, contained interlocutor speech or had no transcription (were silent).

Additionally, interval datasets for 2-s and 4-s were created to investigate the effect of interval length. From the 2-s scheme, 5,985 intervals were extracted. Of these, 3,508 intervals had singular and valid labels. The 4-s scheme yielded 2,982 intervals, of which 2,020 had singular and valid labels. Comparison across the 2-, 3-, and 4-s UCLASS interval subsets is made in section 3.2.2.

2.1.2 The KSoF dataset

The KSoF dataset contains 4,601 3-s intervals of speech of which 2,907 had valid singular labels for fluent speech, prolongation, part-word repetition (PWR), whole word repetition (WWR), and blocks. Here, the data were split into training ($N = 1{,}545$), validation ($N = 662$), and test ($N = 700$) folds which was the split that Schuller et al. (2022) used. KSoF also has filler ($N = 390$), modified speech ($N = 1{,}203$), and garbage intervals ($N = 101$). However, since these classes were not available in UCLASS and some are specific to Kassel's stuttering treatment, these intervals were dropped to allow cross dataset comparisons.[1] Any intervals where there was more than one type of disfluency within the 3-s interval were dropped in KSoF.

Comparison of the distribution of speech annotations for the UCLASS-Event subset and both Interval sets (Table 1) revealed some marked differences. Fluent speech accounted for >80% of observations in the event subset whereas, the relative frequency of fluent observations in both 3-s interval sets were 37 and 53%. Note that the UCLASS-Event and UCLASS-Interval sets were obtained from the same audio files. The difference in fluency distribution was due to the interval method reducing the percentage of fluent observations by marking whole intervals with disfluent speech as stuttered whereas they often contained some fluent speech. The event-based scheme preserved all instances of fluent speech since stutter labels delimited the exact extent of the disfluent speech. Effectively, the interval method under-samples fluent speech and would lead to interval-based schemes over-estimating stuttering severity. The absolute and relative frequencies of each type of event for the training, validation and test sets are given in a link in section 10.

2.2 Feature extraction

Acoustic and linguistic features were extracted to separate stutters from fluent speech. Acoustic features were extracted directly from the audio signal. The linguistic features were derived from a separate speech recognition model's prediction from the audio signal. Acoustic and linguistic feature sets were generated for all the available audio data.

2.2.1 Acoustic features

The acoustic features should provide information concerning how temporal and spectral components change across 2/3/4-s intervals and events. Before acoustic feature extraction was performed, all audio data were normalized such that the oscillogram

[1] Since certain classes of speech were dropped from KSoF to allow comparison with UCLASS, the number of observations in each split differed from that reported in (Schuller et al., 2022).

had maximum and minimum amplitudes in each audio file of +1 dB and −1 dB. For the acoustic set, a set of classical acoustic features were defined. These were: zero-crossing rate, entropy and 13 Mel Frequency Cepstral Coefficients (MFCCs) that were extracted for successive 25 ms time-windows (15 ms overlap). Delta derivatives were calculated across adjacent windows to represent how the features change dynamically across time. Together this resulted in 32 acoustic features per time-frame (Figure 1). These acoustic features are commonly used in the ARS field (Barrett et al., 2022) and pick up on both static (Ifeachor and Jervus, 2002; Tyagi and Wellekens, 2005) and dynamic features of speech (Fredes et al., 2017).

Additionally, a pre-trained deep neural network for representing speech was used to further increase the information presented to the classification models. Here, we used wav2vec 2.0 XLSR-53 (Conneau et al., 2020), as it was trained to represent cross-lingual speech representations from the raw waveform. Note, wav2vec 2.0 XLSR-53 was used for acoustic feature extraction only. For linguistic features a different model, Whisper, was used (Section 2.2.2). The raw waveforms from the data used in this project were inputted to the system, with the resultant tensors of each transformer layer model being used to represent latent aspects of the speech in the signal. This was combined with the classical acoustic features mentioned previously.

The feature matrices were mean-normalized and scaled on the training and validation splits. The resulting feature extraction process produces many features including 1,024 features from the pre-trained network and an additional 32 features from the classic acoustic features. The dynamics of a given feature over the time course of each interval/event was reduced to a singular observation using principal component analysis (Wei, 2019). That is, there was one observation (row of features) for each interval/event. As there were 3,117 intervals in the 3-s UCLASS dataset, its feature set had 3,117 rows.

2.2.2 Linguistic features

The linguistic feature set should provide supra-lexical information that is not readily captured by acoustic features. Stuttered speech contains non-words/syllables that are not included in standard language model vocabularies. Also, disfluent syllables/words/phrases are likely to be infrequent in standard text corpora used to train language models. Furthermore, the audio records in stuttered speech corpora often contain background noise particularly when they are collected in clinical settings. For these reasons, the current state-of-the-art automatic speech recognition model, Whisper, was used (Radford et al., 2023b) first because its architecture can decode speech without a language model, thus enabling it to transcribe both fluent and disfluent speech. Second, it was trained with 680,000 h of audio speech from a wide range of datasets which allows it to be robust against background noise such as those present in stuttered speech audio samples. Finally, its performance on English is reportedly similar to professional human transcribers and it outperformed another state-of-the-art system Wav2Vec2.0 (Conneau et al., 2020) with an improvement of 55% across a range of English datasets.

Whisper comes with multiple pre-trained models. The multilingual model of medium size was chosen. The medium model has 769 million parameters and is capable of transcribing English and German. The medium model performed similarly on English and German with a Word Error Rate (WER) of 4.4 and 6.5%, respectively, on Fleurs (a multilingual dataset).

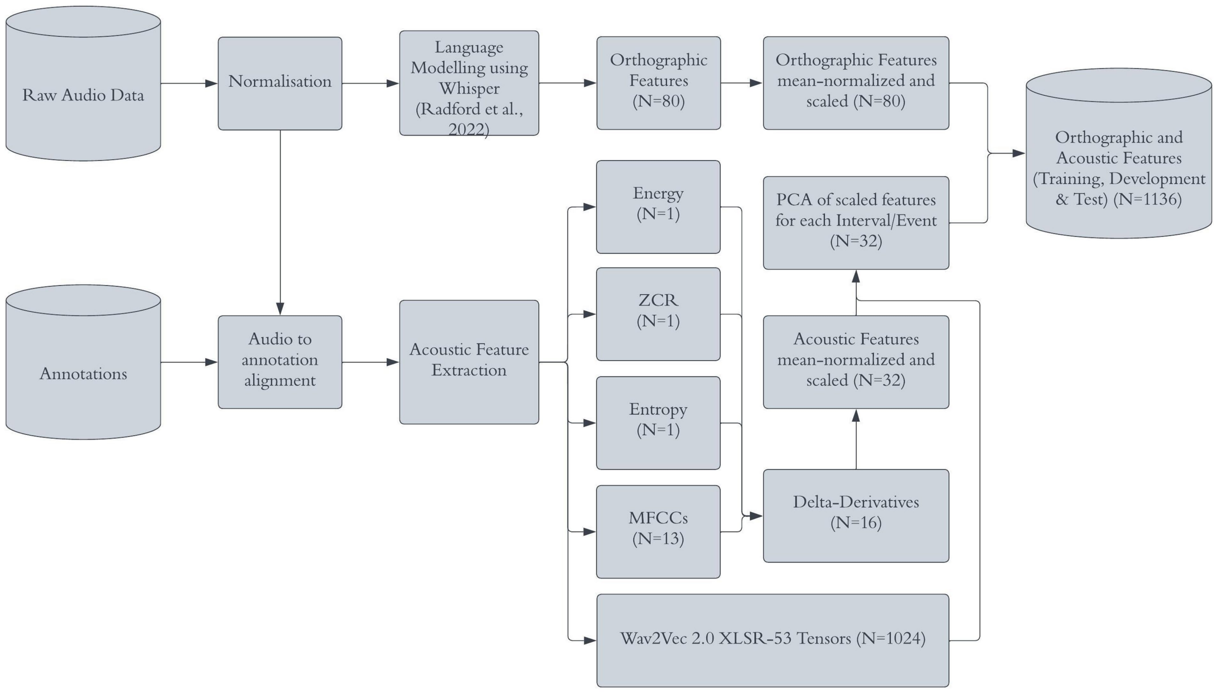

Raw Audio Data
Normalisation
Language Modelling using Whisper (Radford et al., 2022)
Orthographic Features (N=80)
Orthographic Features mean-normalized and scaled (N=80)
Orthographic and Acoustic Features (Training, Development & Test) (N=1136)
Energy (N=1)
PCA of scaled features for each Interval/Event (N=32)
ZCR (N=1)
Annotations
Audio to annotation alignment
Acoustic Feature Extraction
Acoustic Features mean-normalized and scaled (N=32)
Entropy (N=1)
MFCCs (N=13)
Delta-Derivatives (N=16)
Wav2Vec 2.0 XLSR-53 Tensors (N=1024)

FIGURE 1
Pipeline for acoustic feature extraction. Reproduced with permission from Barrett (2024).

The respective language identity information (English or German) was provided when transcribing the two datasets.[2] Given that Whisper was not established for its ability to transcribe stuttered speech including WWRs, we first explored what parameters would encourage a faithful transcription of stuttered speech in a set of small-scale experiments. We found that the default *temperature* parameter influenced its ability to transcribe stuttered speech, especially for WWRs. A model with a temperature of 0 always selected the candidate with the highest probability, and this often failed to generate any repeated syllables/words/phrases in our tests. We therefore experimented with raising the temperature parameter to encourage the model to generate more diverse transcriptions. In our experiments, we generated multiple top-ranked transcription candidates per audio sample. We found that stuttered speech samples were only sometimes faithfully transcribed as one of the candidates, whilst fluent speech samples were transcribed more consistently across transcriptions. We therefore opted to generate multiple possible transcriptions per audio sample following the procedures outlined in the GitHub discussion forum (Radford et al., 2023a). We did this by raising the *temperature* parameter to 0.1 and setting the *best of* parameter to 5 which selected from five independent random samples. Each audio sample was decoded three times, yielding three sets of transcriptions.

The model's decoding strategy generated transcription chunks (called *segments* in Whisper) which were similar to phrases. The three sets of information returned for each decoded stimulus were: (a) a sequential string of orthographic characters (including spaces and punctuation symbols) for each chunk; (b) the probability of the transcription and the non-speech probability of each decoded chunk; and (c) the timestamps of the acoustic signal that corresponded to each decoded chunk. Sub-sets of Orthographic, Probabilistic and Temporal ARS features were obtained using these respective outputs.

The orthographic features were computed over the entire transcription by concatenating the transcriptions from all chunks. Three types of orthographic features were computed: Sequential lexical n-gram repetition, non-sequential lexical n-gram repetition and non-sequential segmental-n-gram repetition. Sequential lexical n-gram repetition is the number of space-separated-word n-grams which are repeated sequentially. This feature was computed using unigrams to capture word/syllable repetitions, e.g., *das das Buch* "the the book," and an additional feature used bigrams to capture phrase repetitions, e.g., das Buch das Buch "the book the book."

Non-sequential lexical n-gram repetition is similar to sequential lexical n-gram repetition, but allows non-sequential repetitions, i.e., not immediately following the n-gram in question, For example, *das Buch nicht das Buch* ("the book no the book"). Two features were computed using unigrams and bigrams, respectively.

Non-sequential segmental n-gram repetition is, in turn, similar to non-sequential lexical n-gram repetition, but applies over characters rather than lexical units. This feature is required because the decoded lexical units had spaces that were not always correctly delimited such that the final instance of prefix repetitions were fragmented. In such cases segmental n-grams can tackle this issue. To avoid detecting repetitions that corresponded to the normal use of repeated syllables/inflectional morphemes in English and German, the repetitions of longer-grams (the length of the orthographic character string minus one) were computed first and. if no repetition was found, then the size of the character n-gram was successively decreased until trigrams were reached. The stopping rule was applied at trigram level to avoid picking up syllable/part-word repetitions. The algorithm stopped immediately at n-grams>3 when repetition was found.

The durations of all the decoded chunks were computed using the timestamps of each chunk. The following summary statistics for temporal features were computed over the durations: the sum, max, min, mean, median, standard deviation, lower quartile (25%), upper quartile (75%), and interquartile range. Two types of probability features were computed: the mean of the probability scores of the transcription and the non-speech probability scores of all decoded chunks.

Each of the above five orthographic features, two probabilistic features and nine temporal ARS features had three values, one from each of the three separately decoded transcriptions. Five summary statistic values (sum, mean, max, min and standard deviations) were computed over each of the three values. The final language-based feature set consisted of 80 feature values [(5 + 2 + 9) * 5 = 80] per audio sample.

As indicated, when applied to continuous speech, event-based segmentation delimits speech types exactly and they vary in duration whereas interval-based segmentation imposes fixed length durations irrespective of the type and extent of speech. Incorporation of language features into the interval-based segments occurs directly when long intervals are used (2-s, 3-s, and 4-s) where interval-length defined the language model's window. As the best way to provide comparability between event-based models and interval procedures that included language-based features, extracts of speech preceding the event were taken so that events were exactly 2-s, 3-s, or 4-s (as required). Two timeframes around an event were used. One where the *lookback* windows always ended at the end of the event defined for this interval (unlike what occurs in standard interval data). The other was where the event was in the middle of the timeframe. I.e., for a 500 ms event with a 3-s lookback, the linguistic features would be derived from 1.5-s before the end of the event and 1.5-s after the event. Note, the lookback could include other speech classes. Although the acoustic features from an event contain orthogonal information pertaining to the class of that event, the linguistic features contain information that pertains to other classes of speech in some cases. This cross-class information was allowed in the current experiments since this is allowed in standard interval datasets. Possible effects on the resultant models are discussed in section 4.3.

For the interval subsets, the ARS model was run for all interval lengths (2-s, 3-s, and 4-s). For the event-based subset, each event lasted approximately 450 ms on average (Table 2). Hence, the language model would have too short an extract to work with. Consequently, look-backs of 2–3- and 4-s were employed so that the ARS model had equivalent duration to the interval-lengths they were compared with (2-, 3-, and 4-s).

2.2.3 Summary

Thirty-two classic acoustic features were extracted directly from the audio signal. Additionally, the pre-trained acoustic model yielded 1,024 features. The linguistic procedure provided a further

[2] Note that we did not rely on Whisper's ability to automatically identify the language from speech because its reported performance is not competitive, and it was not an objective of the current study.

TABLE 2 Estimated mean, standard deviation and quartiles for the length of an event (in ms), split by fluency classes from UCLASS Event subset.

Class	Mean event length (ms)	Standard deviation (ms)	Lower quartile (ms)	Upper quartile (ms)	IQR (ms)
Fluent	222	208	102	270	168
Prolongation	521	311	313	660	347
PWR	763	418	467	980	513
WWR	237	155	142	302	160
Block	578	467	201	836	635

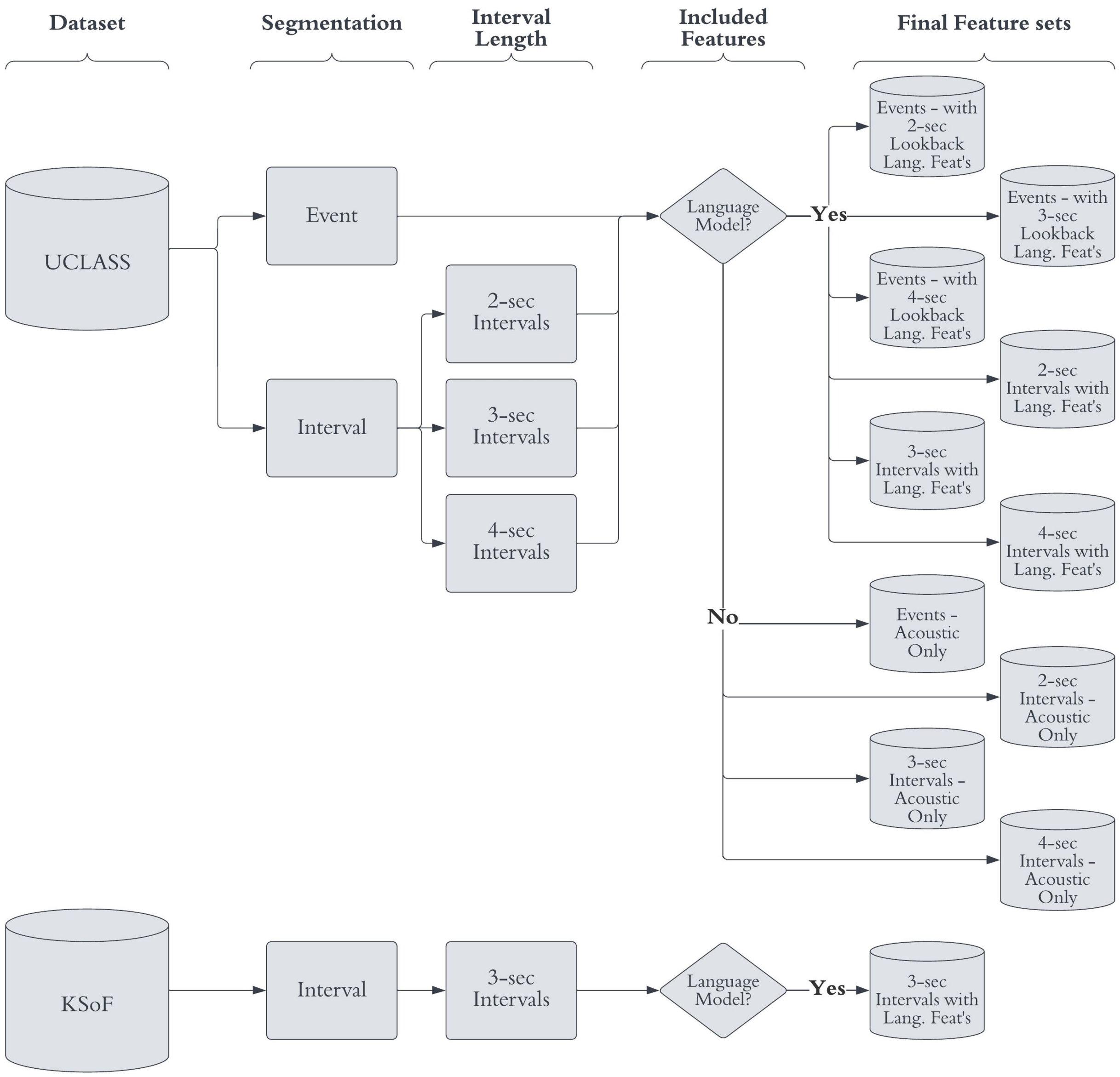

Dataset
Segmentation
Interval Length
Included Features
Final Feature sets
UCLASS
Event
Interval
2-sec Intervals
3-sec Intervals
4-sec Intervals
Language Model?
Yes
No
Events - with 2-sec Lookback Lang. Feat's
Events - with 3-sec Lookback Lang. Feat's
Events - with 4-sec Lookback Lang. Feat's
2-sec Intervals with Lang. Feat's
3-sec Intervals with Lang. Feat's
4-sec Intervals with Lang. Feat's
Events - Acoustic Only
2-sec Intervals - Acoustic Only
3-sec Intervals - Acoustic Only
4-sec Intervals - Acoustic Only
KSoF
Interval
3-sec Intervals
Language Model?
Yes
3-sec Intervals with Lang. Feat's

FIGURE 2
Flow diagram of feature extraction permutation. The final feature-sets used, the original dataset, segmentation method, interval length and included features are given. Reproduced with permission from Barrett (2024).

80 features. The two sets of features were concatenated, and z-score scaled (Obaid et al., 2019). This resulted in 11 feature sets (Figure 2) with 1,136 columns and the number of rows equaled the number of intervals/events. The feature sets were then split into training, validation, and test sets, using the same hard, speaker-independent split (Section 2.2).

2.3 Metrics

Classification reports are available for each model in the links in section 10. Here, AUC-ROC was the main metric of comparison. This provided an appropriate measure for unbalanced multiclass problems (Jeni et al., 2013) whilst also allowing for simple comparisons. While AUC-ROC provides a reasonable abstracted statistic of model performance, it can mask how the model performs for individual classes. AUC-ROC is used for brevity, but it is recommended to inspect the confusion matrices of all models (see Supplementary materials) for class-level comparisons.

2.4 Experimental models

Two types of model are reported in this paper: a Gaussian-kernel SVM (G-SVM) and a multi-layered perceptron neural network (MLP-NN). The Gaussian kernel of the SVM used a penalty term C of 1.15 and gamma varied as a function of the training sets (Equation 1):

$$\gamma = \frac{1}{n \times X_{\text{var}}} \tag{1}$$

Where, n is the number of classes (5) and X_{var} is the variance of the training set. Predictions were weighted by the class frequencies present in the training set, Equation 2.

$$\omega_i = \frac{N}{n \times N_i} \tag{2}$$

Where ω_i is the weight for the i$^{\text{th}}$ class, n was the number of classes (5), N was the total number of observations in the training set and N_i was the number of observations in the training set for the ith class.

For the MLP-NN, a sequential deep neural network was constructed with an input layer, five densely connected hidden layers, five drop-out layers and an output layer yielding probabilities for each speech class. The features were input to the first layer with an equal number of nodes. Then, node outputs were propagated through seven densely connected hidden layers, each with a normalization layer with 10% node drop-out. In each of the hidden layers, the outputs were passed through the Rectified Linear Unit activation function (Agarap, 2018), which returned the original input to the function if the input was positive. Finally, outputs from hidden layers were passed through the SoftMax activation function to yield the class probabilities for a given observation. This architecture yielded 819,205 trainable parameters.

The model was trained across 15 epochs with a batch size of 32. Loss was minimized using cross-categorical entropy, which permitted estimation of loss between multi-class probability densities, and was optimized with the solver 'Adam', a form of stochastic gradient descent (Kingma and Ba, 2014).

3 Results

The field of ARS lacks standards for comparing multiclass models making cross-model comparisons fallible (Barrett et al., 2022; Sheikh et al., 2022). Here, the unweighted AUC-ROC statistic was used as it provides a valid metric for model comparisons as it is virtually unaffected by skewness in datasets and can weight each class of speech equally (Jeni et al., 2013). If a weighted metric was used, it can lead to spuriously high performance due to over-learning fluent speech which is the most frequent class.

3.1 Distribution for the datasets

3.1.1 Distribution of the 3-s intervals and event-based subsets

Before report of the model performance on the UCLASS subsets, the differences in overall fluency/disfluency rates between datasets were reviewed. Table 3 gives the absolute and relative frequencies of

TABLE 3 Total and relative frequencies of intervals and events in the KSoF and UCLASS datasets with each datasets ratio of fluent speech to stuttered.

Sub-set	Class	Absolute frequency	Relative frequency (%)	Ratio to fluent speech
KSoF|3-s interval (N = 2,907)	**Fluent**	1,538	52.91	1
	Prolongation	346	11.90	0.22
	Part-word repetition	339	11.66	0.22
	Whole word repetition	94	3.23	0.06
	Block	590	20.30	0.38
UCLASS|3-s interval (N = 3,117)	**Fluent**	1,228	39.47	1
	Prolongation	383	12.31	0.31
	Part-word repetition	733	23.56	0.60
	Whole word repetition	44	1.41	0.04
	Block	723	23.24	0.59
UCLASS|Event (N = 14,351)	**Fluent**	11,837	82.48	1
	Prolongation	396	2.76	0.03
	Part-word repetition	469	3.27	0.04
	Whole word repetition	173	1.21	0.01
	Block	1,476	10.28	0.12

each class of speech per dataset. Additionally, the ratio of each class of speech relative to fluent speech is given.

When segmentation schemes applied to the same data were compared, drastic differences occurred in the relative frequencies of each speech class. In UCLASS-Interval, fluent speech accounted for less than half the labels whereas fluent speech accounted for over 80% of labels in the 3-s Event-based version of UCLASS. A Chi-square test for independence confirmed that the two distributions of speech classes differed significantly ($\chi_4^2 = 3031.80; p < 0.001$). As discussed in the introduction, this is due to under-sampling the occurrences of fluent intervals. As this paper is approaching stuttering and machine learning from a detection standpoint, fluent speech can be thought of as an absence of stuttering. When removing fluent speech from the distributions we again get a significant difference between UCLASS event and 3-s interval subsets, however with an appreciably smaller statistic ($\chi_3^2 = 305.39; p < 0.001$).

The distribution of stuttering classes for the 3-s interval types was compared across the KSoF and UCLASS Interval datasets. There was good agreement with respect to relative frequencies of event classes. Both estimated fluent speech to be the most frequent class, although the proportion in KSoF was higher. The higher relative frequency of fluent speech in KSoF was probably due to annotators knowing that a modified speech technique was used by participants (Euler et al., 2009). This would have led to some intervals which would have been categorized as one of the classes of stuttered speech being considered fluent. For example, modified KSoF speech allows intervals that are similar to prolongations to be designated fluent as Bayerl et al. (2022a) noted. Otherwise, the order of stuttering subtype by frequency was usually similar across the 3-s interval datasets. However, KSoF had more part-word repetitions than prolongations whereas the opposite was the case with the UCLASS-Interval subset. This was probably because some prolongation intervals were classified as modified intervals that reduced their incidence in KSoF. A Chi-square test showed that the distributions for the two datasets differed significantly across stuttered and fluent speech ($\chi_4^2 = 207.13; p < 0.001$). Hence, the hypothesis that the distribution of both the KSoF and UCLASS interval datasets would be homogenous was only partially supported. When fluent speech is dropped, this difference is further reduced ($\chi_3^2 = 98.98; p < 0.001$). However, the difference between the UCLASS-Event and UCLASS-3 s-Interval distributions ($\chi_4^2 = 3031.80$) was still larger than the difference between the UCLASS-3 s-Interval and KSoF-3 s-Interval distributions ($\chi_4^2 = 207.13$). This is explored further in section 4.1.

Unlike the interval subset, where the length of an interval was known *a priori*, the length of events varied. Since the events in the current subset were defined by syllable onsets and offsets, the event length was expected to be approximately 200 ms for fluent speech and 500 ms for disfluent speech (Howell, 2010). Table 2 provides further support for these estimates. This is the first time that the length of stuttered events, split by type, have been reported to our knowledge (Figure 3).

3.1.2 Distribution of fluency types in 2-s, 3-s, and 4-s interval subsets

When UCLASS datasets for different interval-lengths were compared, all had relatively small ratios of fluent to disfluent speech compared to the UCLASS event dataset apart from whole-word repetitions. For interval approaches, a high rate of fluent speech would be expected when shorter time windows (<3-s) were used and a low rate of fluent speech when longer time windows (>3-s) were used. The expected trends in the interval length permutations were confirmed; the shorter the interval, the greater the proportion of fluent speech (Table 4). Prolongations showed much the same relative frequency across the subsets while the proportion of PWR and blocks increased considerably as interval length increased.

3.1.3 Word error rates of the automatic transcriptions

As mentioned, the linguistic features were generated from the outputs of a pre-trained ASR model. While the performance of this

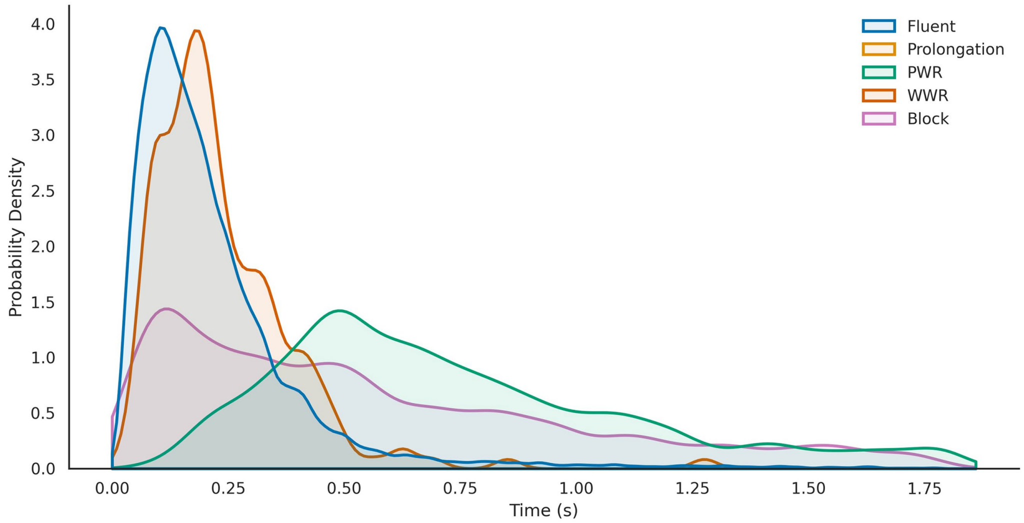


FIGURE 3
Gaussian kernel density estimates of the relative frequencies of event lengths split by speech class from the UCLASS Event subset. *X*-axis gives event length in seconds and the *Y*-axis shows probability. Reproduced with permission from Barrett (2024).

TABLE 4 Total and relative frequencies of intervals for 2-, 3-, and 4-s UCLASS interval subsets, split by class.

Type	2-s ($N = 3,508$)		3-s ($N = 3,117$)		4-s ($N = 2,020$)	
	Absolute	Relative (%)	Absolute	Relative (%)	Absolute	Relative (%)
Fluent	1,532	43.67	1,228	39.40	605	29.95
Prolongation	442	12.60	383	12.29	249	12.33
PWR	644	18.36	733	23.52	422	20.89
WWR	89	2.54	44	1.41	56	2.77
Block	801	22.83	729	23.39	688	34.06

TABLE 5 The word error rate (WER), number of substitutions, deletions, insertions, correct and total words from Whisper (Radford et al., 2023b) split by UCLASS file.

ID	Substitutions	Deletions	Insertions	Correct words	Total words	WER (%)
M_0030_16y4m_1	34	9	9	354	406	12.81%
M_0061_16y9m_1	36	30	9	272	347	21.61%
M_0078_16y5m_1	8	18	14	198	238	16.81%
M_0107_07y7m_1	16	29	30	145	220	34.09%
M_0121_11y1m_1	5	29	28	45	107	57.94%
M_0121_15y1m_1	10	26	18	38	92	58.70%
M_0553_10y0m_1	6	22	26	127	181	29.83%
M_0553_11y0m_1	11	14	23	116	164	29.27%
M_1064_47y0m_1	27	90	52	824	993	17.02%
M_1100_28y0m_1	28	88	38	889	1,043	14.77%
M_1101_35y0m_1	36	63	56	470	625	24.80%
M_1103_20y0m_1	35	78	41	555	709	21.72%
M_1104_40y0m_1	28	32	41	602	703	14.37%
M_1105_21y0m_1	63	324	203	765	1,355	43.54%
M_1106_25y0m_1	7	17	16	172	212	18.87%
Total	350	869	604	5,572	7,395	24.65%

model is well documented on reference speech corpora (Radford et al., 2023b), how the model performs with stuttered speech is not known. Here, manual transcripts of the selected UCLASS data were compared against automatically generated transcripts from Whisper. The same model settings were used as defined in section 2.2.2. Whisper's performance here was evaluated using word-error rate (Equation 3).

$$\text{WER} = \frac{S + D + I}{\text{N}} \qquad (3)$$

Where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of words in the veridical transcription. In the context of ASR transcriptions, substitutions are where the system replaces the reference word, for example "lose" with phonetically similar hypothesized word, for example "rouse." Deletions are where a reference word is removed completely from the hypothesis I.e., the "*it*" in the reference "has *it* gone missing" to the hypothesized "has gone missing." Finally, insertions are hypothesized words that are completely missing from the reference. As in the hypothesis "they wore *many* masks" from the reference "they wore masks."

Overall, Whisper yielded an average WER of 24.65% across all the UCLASS audio files (Table 5). For comparison, Radford et al. (2023b) reported an average WER of 12.8% across multiple speech corpuses. Stuttered speech presents an almost doubling of WER. This is one of the first investigations of how stuttered speech affects WER of state-of-the-art ASR models. How and why stuttered speech causes such decreases in performance remain unclear. While it is beyond the scope of the current work, this would be well worth further research.

3.2 Model performance on UCLASS datasets

3.2.1 Event subsets

3.2.1.1 Three second lookback

G-SVM and MLP used the principal components of the acoustic features, the outputs from a pre-trained deep neural net, along with the orthographic features. The ARS model was provided with a

TABLE 6 Classification reports of the G-SVM and MLP-NN tested on the UCLASS Event subset.

Class	Gaussian SVM			MLP-NN			Observations
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Fluent	92.89	75.30	83.18	82.29	17.06	28.83	1822
Prolongation	4.03	12.82	6.14	8.82	7.69	8.82	39
Part-word repetition	18.60	25.00	21.33	18.52	7.81	10.99	64
Whole word repetition	4.17	20.00	6.90	1.56	76.67	3.05	30
Block	48.37	70.18	57.27	51.11	58.55	54.57	275
Accuracy:			71.39			22.56	2,230
Unweighted average	33.61	40.66	34.96	32.26	33.56	21.03	2,230
Weighted average	82.52	71.39	75.83	74.77	22.56	30.36	2,230

TABLE 7 Summary of AUC-ROC scores for each event-based model from the UCLASS data, split by lookback duration (2-, 3-, and 4-s) and context of the language-based features.

Context	2-s ($N = 2{,}230$)		3-s ($N = 2{,}230$)		4-s ($N = 2{,}230$)	
	G-SVM	MLP	G-SVM	MLP	G-SVM	MLP
Before	0.82	0.73	0.82	0.73	0.82	0.69
Middle	0.82	0.75	0.82	0.74	0.82	0.71

'Before' is a lookback of N-seconds before the end of the event. Middle is ± $\frac{N}{2}$ seconds around the end of the event.

TABLE 8 Classification reports of the G-SVM and MLP-NN tested on the UCLASS 3-s interval subset.

Class	Gaussian SVM			MLP-NN			Observations
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Fluent	44.68	9.38	15.50	46.83	26.34	33.71	224
Prolongation	11.37	57.14	18.97	8.63	28.57	13.26	42
Part-word repetition	0.00	0.00	0.00	29.00	24.79	26.73	117
Whole word repetition	2.13	20.00	3.84	1.92	10.00	3.23	10
Block	23.81	34.48	28.17	26.98	19.54	22.67	84
Accuracy:			16.04			24.58	480
Unweighted Average	16.40	24.20	13.30	22.67	21.85	19.92	480
Weighted Average	26.21	16.04	14.08	34.61	24.58	27.58	480

maximum of 3-s of audio before the end of the speech event of interest. Where there were less than 3 s of speech available (i.e., within the first 3-s of the audio recording), the length was set to the longest duration available. The G-SVM yielded an average AUC-ROC of 0.83 in test. The MLP-NN performed less well with an AUC-ROC of 0.73. (Table 6 has the full classification report).

Given the large imbalance in class frequencies, accuracy should not be used as the sole metric for comparison (Barrett et al., 2022). How performance of these models compared to their interval-based counterparts is reviewed in section 3.3.2.

3.2.1.2 Varying lookback length and window length

Next, 2-s, 3-s, and 4-s lookback lengths were investigated to determine any effects they have on ARS trained on events. When the length of the window was varied with the window ending at the end of the event, there appears to be little effect of varying the duration of the lookback on the model's ability to classify the current event (Table 7). However, there was a drop off in performance for the NN-MLP when extending the lookback to 4-s (AUC-ROC = 0.69) as opposed to 2-s and 3- lookbacks (both AUC-ROC = 0.73). Additionally, it appears that allowing the linguistic features to represent both the preceding and succeeding speech improved performance with respect to AUC-ROC. This was the case with the NN-MLP models, where performance improved for all window lengths as a result of moving the window to include the preceding and succeeding signal.

3.2.2 Interval subsets

As mentioned, the reference interval length was 3-s. The G-SVM and MLP models were trained on the 3-s acoustic and linguistic features. The G-SVM yielded an AUC-ROC of 0.52 at test while the NN-MLP yielded AUC-ROC = 0.54 (Table 8 has the full classification report).

The hypothesis that event-based data should yield better performance than interval-based data was supported. Performance using AUC-ROC improved when models were trained and tested on data from event-based segmentation rather than from intervals.

The hypothesis that the smaller the interval length, the better the model performance was not supported. Rather the relationship between performance and interval length depended on the type of model used. For NN-MLP models, a quadratic relationship occurred with performance in terms of AUC-ROC peaking when a 3-s interval was used (AUC-ROC = 0.54) and dropping off with smaller (AUC-ROC = 0.48) and longer interval lengths (AUC-ROC = 0.49). In contrast, G-SVM models improved with increased interval length (AUC-ROC 2-s = 0.50; 3-s = 0.50; 4-s = 0.54). However, the variation across interval lengths for both types of model was minor throughout.

3.2.3 Input switching

To further investigate how event- and interval-based inputs influenced how models learned to separate classes of speech, the inputs to the trained models were switched. Thus, models trained and validated on event-based inputs were tested on interval-based inputs and *vice-versa*. This novel method allowed for investigation of a model's input-invariant properties. The audio data used was the same but the method of segmentation differed. Thus, if performance remained stable, models should be able to separate the classes of stuttering irrespective of segmentation method. When using a NN-MLP architecture, however, switching the input type between intervals and events resulted in models performing equally well, regardless of input (ROC-AUC = 0.54). A G-SVM, model trained on event inputs yielded a greater ROC-AUC (0.57) as compared to the G-SVM trained on intervals and tested on events (ROC-AUC = 0.51). Indeed, the model trained on events outperformed any model trained and tested on intervals (all ROC-AUC in section 3.2.2 < 0.57). This suggests that segmentation method is causal to a machine learning model's learnt class boundaries. Additionally, G-SVM models trained on event-based data can be used successfully to predict stutters in interval type data.

The hypothesis was made that switching input would yield different responses depending on what the models were trained on. However, it seems that regardless of how the data were segmented during training, if data were used from the other segmentation method, learning performance did not transfer. Therefore, deciding on segmentation *a priori* has lasting effects on their future utility for ARS. Given that event-based procedures are usually employed by speech-language pathologists, SLPs (Riley, 2009), this suggests a preference for training models on event-based data.

3.2.4 Effect of linguistic features

As reviewed in the introduction, classification of stutters has usually used acoustic features as input. Here, linguistic features were also used to help separate supra-lexical disfluencies from fluent speech (WWR) as these are reported to be difficult to separate when using acoustic features alone. Performance with the linguistic features has been reported in 3.1.1 and 3.2.2 for events and intervals, respectively. When these features were dropped from the models, using only the acoustic features, similar pattern of results were seen; models trained on events had AUC-ROC$_{SVM}$ = 0.82; AUC-ROC$_{MLP\text{-}NN}$ = 0.74 and these outperformed models trained on 3-s intervals (AUC-ROC$_{SVM}$ = 0.54; AUC-ROC$_{MLP\text{-}NN}$ = 0.55). For full classification report, visit the link in section 10. Using an AUC-ROC metric, it is not clear whether the linguistic features provided significant benefit to models trained on either Event- or Interval-based data. Indeed, the MLP models trained on events without the linguistic model features performed minorly worse than models with linguistic features, scoring an AUC-ROC of 0.74 as compared to a maximum of 0.75 on events with a 2-s lookback (Table 7). When considering the AUC-ROC, the addition of language-model features provided limited benefit. However, the changes at the class level for precision and recall showed some improvements as a result of language features (Figure 4).

Although the linguistic model features did not systematically improve performance with respect to AUC-ROC, the original purpose was to increase performance with respect to supra-lexical classifications (i.e., WWR). When linguistic features were included, only the disfluent classes of PWR and WWR showed an increase in F1-Score, however the nature of improvement was not the same for the two classes. For PWR, the linguistic features improved the models' recall while reducing the precision, while, for WWR, the opposite was true. Therefore, a tradeoff emerged between precision and recall, depending on whether PWR or WWR are considered. Another trade-off emerged but, in the identification of fluent speech. Fluent speech followed a similar pattern to PWR, with recall improving through inclusion of linguistic features while precision reduced. The implications of this trade-off are explored further in section 4.3.

When considering WWR's alone, events yielded better recall in 2-s and 3-s lookbacks (Figure 5). When the lookback was increased to 4-s, however, recall was worse in events than in 4-s intervals. Additionally, precision improved in all event-interval comparisons except when the linguistic features were input with 2-s of speech (Figure 5A).

3.2.5 KSoF interval dataset

The G-SVM yielded an AUC-ROC of 0.55 on the test set. The MLP-NN performed similarly with an AUC-ROC of 0.53 (Table 9 for the full classification report).

From AUC-ROC, the G-SVM outperformed the MLP-NN. However, there were substantial differences with respect to sub-class performance. When performance was compared with respect to precision, recall and F1-score, the G-SVM defined the classes of prolongation and part-word repetition better (Table 9), whereas the MLP described fluent speech, whole-word repetition and blocks better in most cases.

It was hypothesized that models trained on KSoF interval data and models trained on UCLASS interval data would perform differently with respect to AUC-ROC. KSoF models yielded AUC-ROC of 0.55 and 0.53 for the G-SVM and MLP-NN, respectively. UCLASS interval models yielded 0.52 for the G-SVM and MLP-NN, respectively. Although the deep learning model provided evidence for the hypothesized result, there does seem to be some non-negligible differences in performance due to the dataset when testing the shallow models.

4 Discussion

4.1 Summary of results

From the shape of the datasets, interval-based methods yielded a significantly lower proportion of fluent speech (KSoF$_{Fluent}$ = 52.97%; UCLASS|Interval$_{Fluent}$ = 39.47%; UCLASS | Event$_{Fluent}$ = 82.48%).

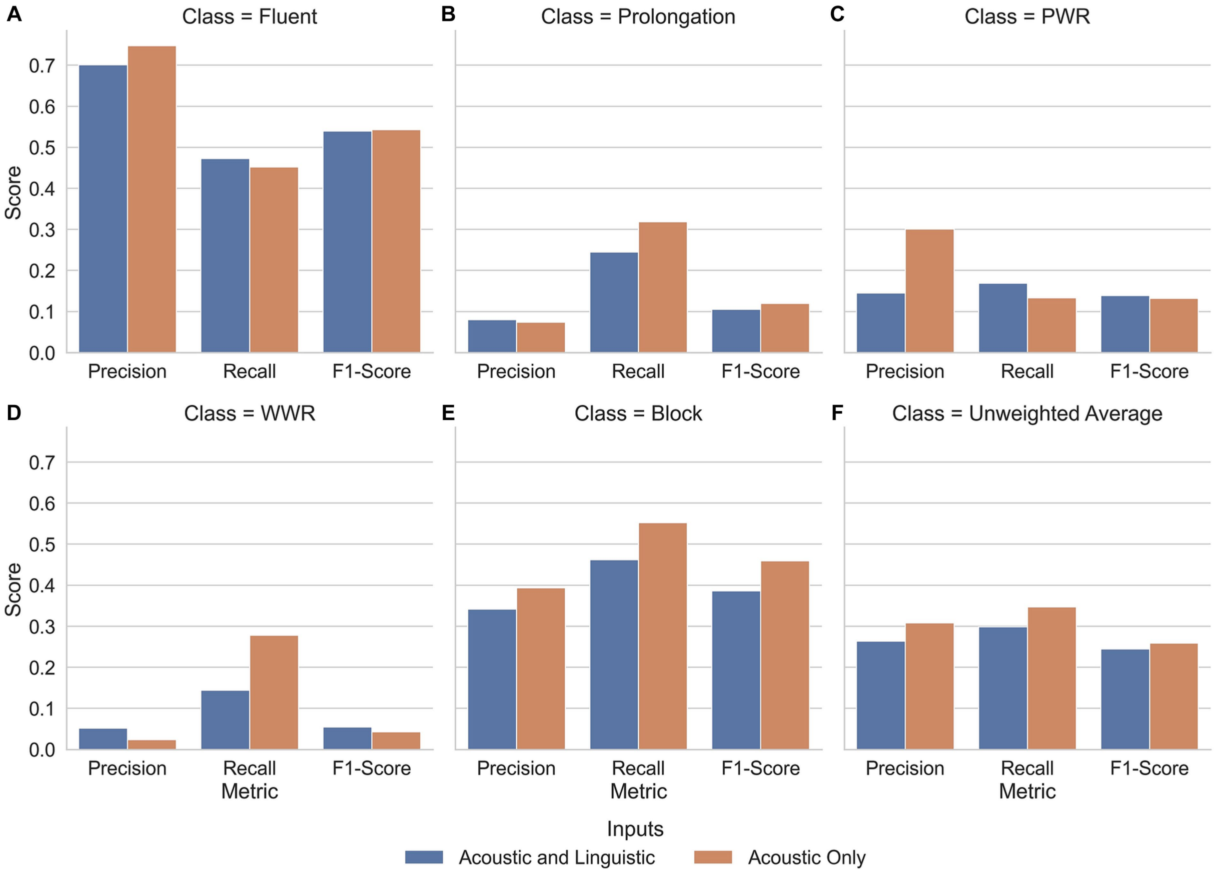


FIGURE 4
Average precision, recall and F1-score for each class of speech. Inputs to the model are split by inclusion (blue) and exclusion (orange) of linguistic features. Additionally, the unweighted average of each metric are plotted (**F**). Along the *x*-axis of each plot are the metrics precision, recall and F1-Score of each class (**A**. Fluent; **B**. Prolongation; **C**. PWR; **D**. WWR; **E**. Block) as well as the unweighted average across all classes (**F**). These are further split into models which input both acoustic and language features (blue) and models which input only acoustic features (orange). Here, the effect of language features on each class is apparent. Along the *y*-axis, the precision, recall and F1-Scores are measured. The scores are an average of G-SVM and NN-MLP models trained and tested on Event-based inputs with a 2-, 3-, and 4-s lookback. In all stuttering classes, language-based features improved recall and reduced precision as compared to their Acoustic-Only counterpart models. Whereas, in fluent speech, the reverse was true, with recall diminishing and precision improving as a result of language features. Reproduced with permission from Barrett (2024).

Due to the limited size of the datasets, it was not possible to specify which, if any, stuttering sub-types were over-sampled. The frequency of WWR, however, did not seem to alter drastically by segmentation approach.

Figure 6 shows the macro-average ROC curves for each model. The models trained and tested on both interval datasets performed poorly with respect to AUC-ROC. Interval models yielded an average macro-AUC-ROC of 0.51. This indicated that these models did not perform above chance when classifying stuttered speech. By contrast, the models trained and tested on event-based data performed reasonably well, with an average macro-AUC-ROC of 0.80.

Models trained on interval data from KSoF and UCLASS showed comparable performance. In both cases, the shallow G-SVM outperformed the deep-learning model on most metrics. It is interesting that the performance on the interval models performed similarly in terms of the ARS problem since they used completely different datasets collected for different purposes. The UCLASS data used here was solely from monologue or conversational speech recorded in the clinic. The KSoF dataset contained speech from monologues in the clinic but, also PWS reading aloud as well as when making phone calls. In KSoF, there were multiple additional sources of variance as compared to the UCLASS Interval subset. As mentioned, the speaking situations varied but also there were more speakers within the KSoF dataset ($N = 37$) than the subset used from UCLASS (N = 14). Additionally, the datasets were in different languages, German and English. The similar performance suggests that the features extracted for the class separation seem to be language independent, at least for those within the Germanic language family. The feature set may extract acoustic features that are universal to stuttered speech which allows fluent and disfluent speech to be separated regardless of the specific language. Future studies should extend examination to other language families to better examine the universality of our acoustic features.

Comparing the current models that used 3-s KSoF data with Schuller et al. (2022) showed that our models performed less well. Schuller et al. (2022) reported only unweighted average recall (UAR), achieving a 37.6 UAR in test using a set of one hundred principal components from a 6,373-feature set. In comparison, using a feature set of 1,136 on the KSoF intervals, the G-SVM yielded a UAR 25.47. Using the UCLASS intervals, a UAR of 24.20 again with a

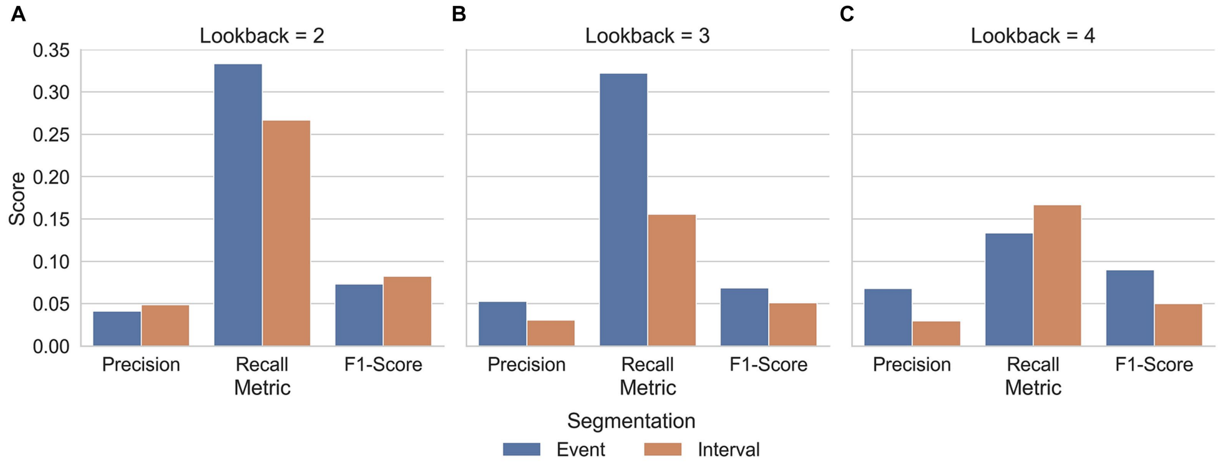


FIGURE 5
Along the x-axis of each plot are the metrics precision, recall and F1-Score of whole-word repetitions. The y-axis indicates the precision, recall and F1-Scores. Subplots are split by lookback (**A**. 2-s; **B**. 3-s; **C**. 4-s). The scores are an average of G-SVM and NN-MLP models trained and tested on Event- and Interval-based data. Sub-plots are separated by the lookback length given at the top. For Intervals, this is simply the length of the interval. For events, the lookback is the period of time before the end of an event that the linguistic features are derived from. That is, a 3-s lookback means that the linguistic features represent the speech 3-s prior to the end of the stuttered event. Reproduced with permission from Barrett (2024).

TABLE 9 Classification report for Gaussian SVM and MLP-NN models on the KSoF test data.

Class	Gaussian SVM			MLP-NN			Observations
	Precision	Recall	F1-score	Precision	Recall	F1-score	
Fluent	45.65	23.86	31.13	39.82	33.33	36.29	264
Prolongation	22.83	26.36	24.47	27.78	9.10	13.70	110
Part-word repetition	22.35	28.79	25.12	18.40	22.73	20.34	132
Whole word repetition	5.97	22.22	9.41	2.29	2.78	4.24	18
Block	23.23	26.14	24.60	16.13	5.68	8.40	176
Accuracy:			25.57			20.43	700
Unweighted average	24.01	24.47	23.00	20.88	19.72	16.59	700
Weighted average	31.02	25.71	26.84	26.97	20.43	21.90	700

Precision, recall and F1-score were split by each class. Additionally, overall model accuracy, unweighted and weighted average precision, recall and F1-score for each model are reported. Finally, the number of observations for each class is reported.

G-SVM. This suggests that performance can be boosted by using further feature dimension reduction techniques. This does not invalidate the conclusion that event-based approaches lead to better machine learning models since the UAR of the UCLASS event-based G-SVM (UAR = 40.66) outperformed Schuller's reference. Rather, models can be further improved by: (a) Supplying a richer feature set as demonstrated by Schuller et al. (2022); and (b) Using event-based segmentation methods.

The hypothesis that models trained on event-based inputs would outperform interval-based inputs was supported. Both shallow- and deep-learning models trained on events outperformed their interval-based counterparts in terms of AUC-ROC (Tables 7, 9). Indeed, for all aggregate metrics reported (accuracy, weighted and unweighted recall, precision, and F1-Score), the event-based UCLASS models outperformed the interval-based models UCLASS (Tables 6, 8).

4.2 Changes in performance due to segmentation approach

Considering the interval- and event-based segmentation method procedures, it was hypothesized that interval-based procedures would limit performance of machine learning models applied to the ARS problem. This hypothesis was confirmed. Interval-based methods led to sub-optimal performances across KSoF and UCLASS datasets compared to interval-based methods.

However, the hypothesis that as the interval length was shortened the performance would increase was not clearly supported. There was some evidence that lengthening the standard interval length from 3-s to 4-s was further detrimental to model performance. In the current study, the minimum interval-length was only reduced to 2-s. As seen in Figure 3, events were closer to 200 (fluent) and 500 ms (disfluent). It may be that further reductions

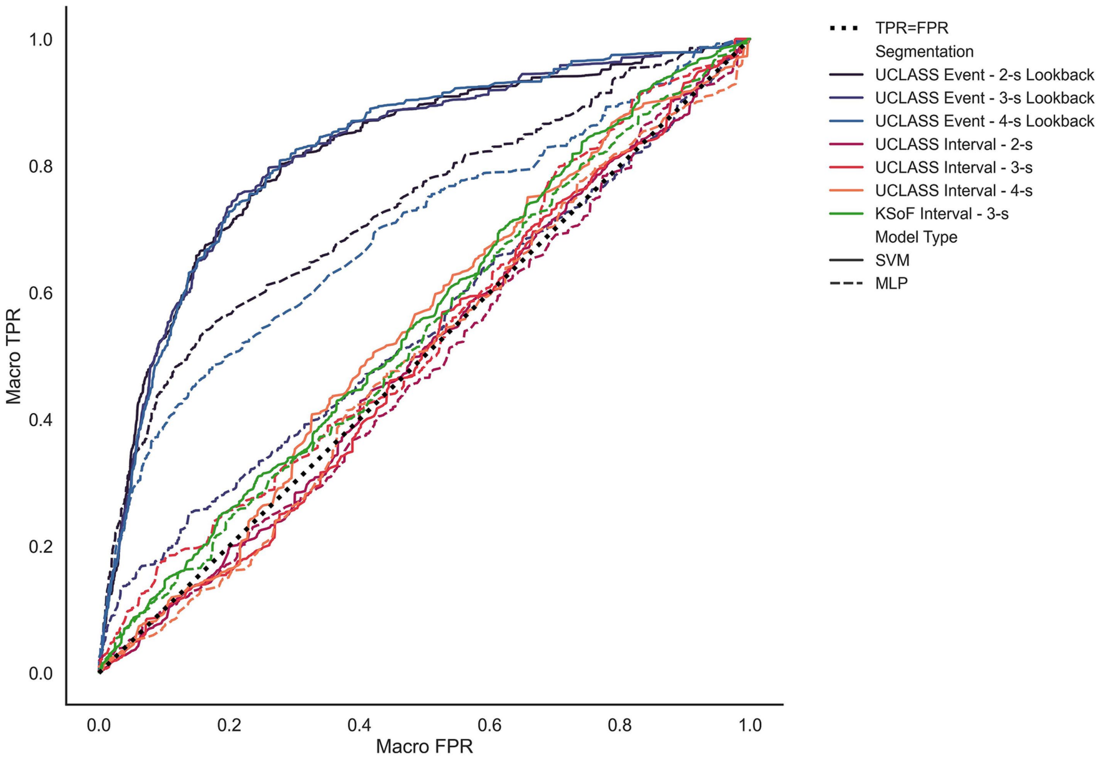


FIGURE 6
Unweighted average receiver operating characteristics for each model and segmentation method. The *X*-axis measures to unweighted average false positive rate. The *Y*-axis measures the unweighted true positive rate. The black dotted line shows where the true positive rate is equal to the false positive rate (i.e., chance). The solid lines represent performance by the G-SVM and dashed lines represent the MLP-NN. G-SVM models trained on the UCLASS event data yield the largest unweighted AUC-ROC with lookback length inseparable at this level. MLP-NN models on event data performing less well. Finally, all models trained on interval-based inputs vary around the TPR = FPR line. Reproduced with permission from Barrett (2024).

in interval length are necessary to observe the predicted effects on performance.

Additionally, input switching analyses revealed that G-SVM models trained on events and tested on intervals outperformed G-SVMs trained on intervals and tested on events. This suggests that the learnt parameters to identify stuttered speech are somewhat preserved when training on events. Indeed, the parameters learned from events seem to allow for a superior class separation to intervals even when a model trained on events is tested with intervals. As mentioned earlier, event-based approaches allow speech to be delimited such that each observation contains only one class of speech whereas stuttered intervals may contain some, or a majority of, fluent speech. As such, the acoustic features extracted from an event contain no concomitant information from another class, allowing models to learn fine grained differences which are likely to be removed when cross-segment orthogonality is disrupted, as in the interval method. Finally, the input switching analyses also suggest that models trained on events can be successfully used to predict speech fluency on intervals.

4.3 Effect of linguistic features

We were not able to find clear evidence that linguistic features increased separation of supra-lexical disfluencies from fluent speech. Despite the addition of features designed to highlight whole-word repetitions, the models overall performed worse when provided with these features. It is unclear why this was the case but, due to the multi-dimensionality of the problem, by increasing the complexity of the inputs to the model, the previously learnt patterns in the acoustic data that help separate sub-lexical disfluencies may become obscured when linguistic features are added. This may explain why linguistic features also reduced the F1 score in speech classes apart from WWR. Another possibility lies with a stopping parameter used to generate the non-sequential segmental n-gram repetition features. The algorithm started with a high n-gram size to find repetitions. The n-gram size decreased if a repetition was not found, whereas it stopped if a repetition was found or if it reached tri-grams to avoid picking up non-lexical repetitions (such as part-word repetitions and prolongation). This tri-gram parameter might be too small to start with and possibly should be increased.

Additionally, linguistic features may have had a detrimental effect on model performance due to possible cross-class correlations within the features. As mentioned, using event-based segmentation the acoustic features represented only the target class. However, the linguistic features incorporated information of up to 4 s before the end of the event. It is feasible, then, that a non-target stutter that precedes the target event influenced the linguistic features. For example, in a 4-s utterance ‘the cat sat sat on the mmmat,’ (target event the prolonged ‘mmm’), there is also the preceding WWR ‘sat.’ The language model would then flag a WWR in the resultant features, leading to contradictory inputs to the ARS model. Future uses of language model features should avoid this issue by ensuring features are relevant to the target event/interval only. Of course, this problem is avoided if event segmentation is employed. We consider that non-orthogonality in the inputs leads to a major limitation in the optimization of ARS modelling. It is proposed that this source of non-orthogonality is a significant factor concerning why the linguistic features did not improve overall performance.

Linguistic features improved the precision of the disfluent classes (including WWR), at the expense of their respective recall. This tradeoff suggests that the quality of the linguistic features has room for improvement. Our small-scale inspection of the ASR transcription suggested that the ASR model would not over-transcribe WWR, that is, the ASR would not transcribe WWR when the signal does not contain WWR (low false positive rate). Future work should inspect the precision and recall of WWR in terms of the ASR transcription. The tradeoff between the recall of the fluent class and the disfluent classes (including WWR) suggests that linguistic features might have helped the model to distinguish between WWR and fluent speech since they cannot be easily distinguished by acoustic features alone.

When considering the quality of the transcripts generated from the ASR system, there is a large error (Table 5). As mentioned, the reference WER of the ASR system used is between 5 and 13% (Radford et al., 2023b). Here, however, an average WER of 24.65% was found with a range of 12.81–58.70%. By far the most frequent type of error made by the ASR system was the deletion of spoken words, reducing the number of transcribed words as compared the number of words actually said by the speaker. Again, deletion here is the complete removal of a word present in a reference transcript in the ASR's hypothesized transcript. I.e., the "it" in the reference "has it gone missing" to the hypothesized "has gone missing." Note, this is the number of errors by the ASR system and not the number of errors (stutters) by the speaker. This may be due to stuttered speech being ignored by the ASR system, resulting in a loss of words transcribed. However, a dedicated analysis is required to confirm this hypothesis which is beyond the scope of the current paper. Additionally, deletions might result from the ASR system removing repetitions. Despite the current paper's attempt to reduce this through adjustment of the hyper-parameters (See section 2.2.2). This not only increases the estimated WER of the system but also removes information of interest for the current purposes. The relatively poor quality of the transcriptions may, therefore, contribute to the current linguistic features' limited effect on model performance. This poor performance of ASR on stuttered speech was also found in (Thomas et al., 2023) which examined the potential for enhancing automatic cognitive decline detection (ACDD) systems through the automatic extraction of disfluency features using ASR systems. The accuracy of ACDD systems was much lower (78.4%) when trained on automatic disfluency annotations than when trained on manual annotations (88.8%).

When model type and segmentation method were combined and the overall difference between models with and without linguistic features was compared (Figure 5) a consistent trade-off between precision and recall emerged. In all stuttering sub-classes, recall improved and precision worsened with the inclusion of features from a language model whereas in fluent speech, the opposite was true; recall worsened and precision improved.

Precision is the ratio of true positive predictions to all positive predictions. Hence, in the fluent speech class precision is the percentage of correct predictions for fluent speech out of all a model's predictions of fluent speech. Recall is the ratio of true positive predictions to all instances of the chosen class. In fluent speech, it is the percentage of correctly predicted cases of fluent speech out of all fluent observations.

Therefore, it seems as though language features improved the representation of stuttering classes at a population level. However, the features also lowered the confidence in an individual prediction being true. In contrast, for fluent speech, language features resulted in an increase in confidence of a prediction being true.

The precision-recall trade-off leads to a decision on the aims of the ARS model. In other fields of machine learning which focus on symptom detection, such as cancer, a high recall is preferred over a high precision since the cost of missing a case of cancer is greater than a false positive. In the field of ARS, however, it is not clear whether precision is preferred over recall or *vice-versa*. For Speech and Language Pathologists who may review the predictions, an emphasis on recall may be optimal since wrong predictions can be resolved later.

As Dinkar et al. (2023) noted, a language model's abstraction from audio input to textual output may result in critical loss of the information which makes the speech stuttered. State-of-the-art language models are often trained using highly fluent materials which are unrealistic in real world scenarios and indeed the audio used in the current work. A large increase in WER was reported here for transcriptions from PWS's speech. This may explain why the linguistic features were of limited benefit to the ARS models. The linguistic features often provided inaccurate information about the represented speech, reduce class separation. For linguistic features to be better utilized for the ARS classification problem, the ASR systems themselves need to be improved for stuttered speech. Work by Rohanian and Hough (2021) highlighted how the ASR outputs can be modified to better capture certain types of disfluent inputs. However, this was limited to fillers in Rohanian and Hough (2021)work. Further work is required to investigate: (a) which stutters are vulnerable to reduction in an ASR's outputs; and (b) how to improve the ASR's outputs for the aims of stuttering detection.

4.4 On event-based approach

The current study presented consistent evidence that the event-based procedure for segmenting stuttered speech allowed models to better classify stuttered and fluent speech than the interval procedure. Regardless of whether the models were shallow or deep, whether language features were included or not and irrespective of length of interval, all event-based models outperformed all interval-based models in AUC-ROC (Tables 7, 9), amongst other metrics. Therefore, it is highly recommended for future research to employ event-based data to train ARS models.

Beyond the practical implications, the results also highlight the importance of class orthogonality in training. A key difference in the features provided to the models by event- and interval-based schemes is the level of cross-class orthogonality. Although the interval-based scheme resulted in no cross-stutter correlation (no intervals contained more than one class of stutter), there was a significant level of fluent-stutter correlation. From the novel analyses on event lengths (Section 3.1.1.), a prolonged syllable is on average 521 ms. Given a 3-s interval labelled as prolonged, with a single prolonged syllable, the expected proportion of audio which pertains to the labelled class is only a sixth of the audio used for feature extraction. The other five-sixths audio provide information of non-target classes (fluent speech, silence, noise, etc.).

However, the event-based procedure is not without its limitations. First, event-based approaches require labelling events rather than intervals. This is time-consuming, with limited opportunity for automation. Syllabic levels of transcription and annotation, as used here, often require expertise in linguistics for reliable markers within the signal to be inserted. Also, unlike interval-based labelling, label permutation or over−/under-sampling methods are not feasible.

Second, as shown in 3.1.1, event-based segmentation resulted in a large class imbalance. The classes of interest (stutters) were dramatically skewed by the predominance of the fluent class. Although this is representative of fluency rates in PWS, this does lead to possible limits and pitfalls for machine learning approaches (Gosain and Sardana, 2017). Given this class imbalance, it is more surprising that event-based models outperformed interval-based models uniformly, as the latter allow for a more balanced dataset. From Table 3, the relative frequency of blocks in the event-based segmentation (10.28) was less than half the relative frequency of blocks in the 3-s interval-based segmentation (23.24%). Yet, the event-based models' ability to represent blocks outperform the interval-based models in every reported metric (Tables 6**,** 8). Again, event-based data provides superior materials for training ARS models. However, there may still be detrimental effects of this class imbalance. The same 3-s interval models outperformed the event-based models in certain metrics on prolongations and part-word repetitions. Therefore, when using event-based approaches, future research may benefit from using methods to counteract this class imbalance.

Third, the event-based approach assumes *a priori* knowledge of event onset and offset times. When given an unlabeled, purely continuous audio stream, a separate event onset-offset model would be required. This contrasts with the interval-based approach where the audio stream is automatically 'chunked' into the prior set time intervals. Also, how one segments events in speech in an online, real-time approach is a further limitation. In the interval-based scheme this problem is trivial. Buffer the Input by the length of the pre-set time window (e.g., 3-s), perform feature extraction and reduction over the signal in this timeframe and feed the resultant features to the model. As discussed above, this may inherently limit the speed of predictions of a model using the interval-based scheme since there is a preset buffer, in our case, 3 s.

Overall, the event-based procedure for speech segmentation provided the best training materials for ARS models.

4.5 On interval-based approach

There are several aspects of the interval-based approach that could be automated where the event-based one cannot. For example, the time duration of an interval is preset. Hence, extracting intervals from a file is easy to process whereas, events must usually be done manually. As traditional ASR models can automate word/phonemic boundary locations in speech, the events could feasibly be automated at this stage also. In a similar vein, the annotator does not need to be trained on separating linguistic components of speech (i.e., syllables, phonemes, etc.) in the interval-based method. This is another stage at which the event-based procedure is more time consuming and costly. However, an interval-based approach cannot ensure that an interval contains only one type of stuttering. Therefore, unless using a multi-label system, the interval approach is fallible to data loss where the event is not. Bayerl et al. (2022b) used a multi-label approach on interval-based data with positive results. Models were able to incorporate this more complex multi-label information without detriment to model performance relative to single-label methods as in Lea et al. (2021) and Bayerl et al. (2022a). Therefore, if using an interval-based dataset, a multi-label approach should be used to limit data drop-out.

Finally, given the inputs to an interval-based model are temporally inflexible, the interval procedure is highly applicable. In the event procedure, events would first need to be separated out in online speech classification, requiring a phoneme recognizer as an initial layer to the model. Whereas an interval method simply makes predictions about the interval provided. For example, a 3-s interval model would be able to make predictions on any 3-s input of audio signal. This does, however, also lead to a critique of the interval method in that the classification speed is, at minimum, the same lag as the interval speed. It therefore seems incompatible with real-time uses where latency is critical.

4.6 On whole word repetition

In both the current paper and a baseline model for kSoF (Schuller et al., 2022), WWR was the most difficult class of speech to recognize. Unlike blocks or prolongations, for example, WWR have no within-word disfluency. Rather, the perceived disfluency is only identified at the word or phrase level. For instance, the prolongation in "The cat ssssat on the mat" occurs on the "s," alone. Whereas, in the phrase repetition "The cat sat on the on the mat," the disfluency occurs across the two words "on the." Given that the models presented here were mainly based on acoustic features with no language model or decoder-encoder components, would WWRs be separable from, for example, fluent speech? It is proposed (a) that WWR are not separable at an acoustic level and (b) they should not be included in the same roster as sub-segmental disfluencies.

Point (a) is supported by the spread of model predictions when an instance of a WWR is input to the model. In Schuller et al.'s (2022) model and the 3-s interval MLP-NN model, WWRs were predominantly assigned to the 'Fluent' speech class. In the Gaussian SVM, the 'Fluent' class was the second-best predicted class for true instances of WWRs after Fillers.

It was hypothesized that the inclusion of language features would increase class separation for WWR. There is limited evidence for this hypothesis. Recall rate improved across all stuttering classes after inclusion of language features. Although language features were detrimental to precision, the theoretical motivation remains clear; if WWR cannot be separated from fluent speech at an acoustic level, information at the supra-segmental/lexical level is required. Continued exploration of features such as phone and word-level n-grams is suggested with special attention to sequence of words. Further investigation with more complex language models may help solve this class inseparability.

4.7 Clinical implications

The human assessment of stuttering, a significant bottleneck for clinical work, is costly in terms of valuable clinical time and often yields variable assessment outcomes (Kully and Boberg, 1988). Automated procedures have promised to lighten workloads (Howell et al., 1997b; Barrett et al., 2022), but they have yet to be implemented in clinical practice.

Despite 25 years of research into automatic stuttering detection and labelling, a significant trade-off remains between model flexibility and model performance. Models are either highly specific to a task within stuttering recognition and yield adequate performance for application in a clinic (Mahesha and Vinod, 2016), or they are flexible enough to better handle the complex nature of stuttering and its classification but do not meet the necessary standards for use in clinical settings. For example, Mahesha and Vinod (2016) present a Gaussian Mixture model with an approximate 95% accuracy. However, the model is only able to classify repetitions (it is unclear whether this includes PWR, WWR or both), prolongations, and interruptions. The models presented here, as well as those presented in Lea et al. (2021) and Mishra et al. (2021), among others (See Barrett et al., 2022 for review), all perform with less than 95% accuracy.[3] However, some works, such as that by Gupta et al. (2020), which achieve more than 95% accuracy, are trending towards a level of performance where use in a clinic should be considered. It is unclear whether the model was provided with an event- or interval-based segmentation scheme.

The study provides compelling evidence that employing event-based procedures enhances the capacity of machine learning models to address the ARS problem in comparison to interval-based procedures. This observation is congruent with common human assessment practices for stuttering (Riley, 2009), which frequently utilize event-based metrics like the percentage of syllables stuttered. Models generated through event-based procedures offer predictions based on events and seamlessly align with prevailing clinical practices, presenting an avenue for not only partially automating stuttering severity assessment but also achieving full automation. While the current models provide important insights for ARS research, they are not suitable for use in clinical scenarios 'out-of-the-box.' As mentioned earlier, the performance levels do not meet the necessary standards. This is demonstrated by a comparison of the true and predicted cases of stuttering in the test (see the Supplementary materials for confusion matrices). For example, the 3-s event-based G-SVM with linguistic features included (described in section 3.2.1.1) yielded a set of predictions (see Supplementary Data Sheet 10) which significantly changed the shape of the speech fluency distribution. Event-based segmentation types led to an approximate distribution of 83% fluent, 3% prolongation, 3% PWR, 1% WWR, and 10% block/break. This approximates the true distribution in the test set. However, if one were to implement automated labelling using the aforementioned model (arguably the best presented here), the shape of speech fluency changes drastically: 46% fluent, 12% prolongation, 3% PWR, 24% WWR, and 15% block/break. Clearly, the models presented are for research purposes only and not for use in the clinic.

While in-clinic work with ARS models has yet to take place, the current work contributes to a growing field providing proof-of-concept evidence that ML models could improve workflows in clinical assessments of stuttering. The current work strongly supports the use of event-based segmentation in the preparation of data for ARS models. Additionally, this form of segmentation fits well with commonly used stuttering assessments (Riley, 2009). Future work should seek to compare how partial and full automation of stuttering assessment performs in comparison to the current standard (no automation). Research should consider the trade-off between the time taken for the assessment and the error imparted due to automation.

5 Conclusion

The current work investigated methods of speech segmentation for machine learning classification. The two main methods of speech segmentation for stuttering classification have been employed: interval- and event-based. While interval-based methods are time and cost effective, event-based methods yield far superior models with less data. This is particularly pertinent given the lack of openly-available stuttering event data currently available (Barrett et al., 2022). Further research could make use of the additional interval databases (Lea et al., 2021) to provide further power to the current study's findings.[4] It is therefore highly recommended that future research uses event-based segmentation methods to build stuttering classifier models. Software to add annotations about stuttering events (onsets, offsets, and stuttering type) to continuous audio files has been provided in Howell and Huckvale (2004).

Data availability statement

The datasets generated and analyzed for this study can be found in the paper's Open Science Framework Page [https://www.doi.org/10.17605/OSF.IO/29K7Q] and KSoF [https://doi.org/10.5281/zenodo.6801844]. The features, saved instances of the models used and code used to generate these objects are available in OSF.

[3] 95% accuracy is chosen as a threshold as we consider that for adequate use in the clinic. In that, the probability that a predicted dysfluency is not actually present for a given prediction should be at least 0.05 or lower. Further work should seek to establish a set of thresholds across accuracy, precision and recall – amongst other metrics – both theoretically and empirically to guide application to clinical settings. While state-of-the-art performance of ARS models can vary freely within research, translation to in-clinic practice requires a separate set of baseline standards.

[4] We would like to acknowledge the reviewers for their suggestions and recommendations on including all datasets available in the future.

Author contributions

LB: conceptualization, investigation, methodology, software, visualization, data curation, writing- original draft preparation. KT: conceptualization, methodology, software, data curation, and writing. PH: conceptualization, supervision, writing- reviewing and editing.

Funding

The work is part of LB's PhD, we would like to thank the Engineering and Physical Sciences Research Council, UK (EP/R513143/1).

Acknowledgments

We would like to thank all those involved in the publication of the Kassel State of Fluency database (Bayerl et al., 2022a). Especially Sebastian Bayerl for his help in navigating and splitting the data.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1155285/full#supplementary-material

References

Agarap, A. F. (2018). *Deep learning using rectified linear units (relu),* arXiv e-prints.

Barrett, L., Hu, J., and Howell, P. (2022). Systematic review of machine learning approaches for detecting developmental stuttering. *IEEE/ACM Trans Audio Speech Lang Process* 30, 1160–1172. doi: 10.1109/TASLP.2022.3155295

Bayerl, S. P., von Gudenberg, A. W., Hönig, F., Nöth, E., and Riedhammer, K. (2022a). *KSoF: The Kassel state of fluency dataset--a therapy centered dataset of stuttering*, (European Language Resources Association).

Bayerl, S. P., Wagner, D., Hönig, F., Bocklet, T., Nöth, E., and Riedhammer, K. (2022b). *Dysfluencies seldom come alone--detection as a multi-label problem*, arXiv e-prints.

Bayerl, S. P., Wagner, D., Nöth, E., Bocklet, T., and Riedhammer, K. (2022c). *The influence of dataset partitioning on dysfluency detection systems*, Springer International Publishing

Barrett, L. (2024). Measurement of feedback in voice control and application in predicting and reducing stuttering using machine learning [Doctoral thesis] University College London.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). *Unsupervised cross-lingual representation learning for speech recognition*, arXiv e-prints.

Dinkar, T., Clavel, C., and Vasilescu, I. (2023). *Fillers in spoken language understanding: computational and psycholinguistic perspectives*, arXiv e-prints.

Euler, H. A., Gudenberg, A. W. V., Jung, K., and Neumann, K. (2009). Computergestützte therapie bei redeflussstörungen: die langfristige wirksamkeit der kasseler stottertherapie (KST). *Sprache·stimme·gehör* 33, 193–202. doi: 10.1055/s-0029-1242747

Fredes, J., Novoa, J., King, S., Stern, R. M., and Yoma, N. B. (2017). Locally normalized filter banks applied to deep neural-network-based robust speech recognition. *IEEE Signal Process Lett* 24, 377–381. doi: 10.1109/LSP.2017.2661699

Gosain, A., and Sardana, S. (2017). Handling class imbalance problem using oversampling techniques: a review. In 2017 international conference on advances in computing, communications and informatics (ICACCI) (IEEE), 79–85.

Gupta, S., Shukla, R. S., Shukla, R. K., and Verma, R. (2020). Deep learning bidirectional LSTM based detection of prolongation and repetition in stuttered speech using weighted MFCC. Available at: www.ijacsa.thesai.org

Howell, P. (2010). *Recovery from stuttering*. New York: Psychology Press.

Howell, P., Davis, S., and Bartrip, J. (2009). The University College London archive of stuttered speech (UCLASS). *J. Speech Lang. Hear. Res.* 52, 556–569. doi: 10.1044/1092-4388(2009/07-0129)

Howell, P., and Huckvale, M. (2004). Facilities to assist people to research into stammered speech. *Stammering Res* 1, 130–242. doi: 10.1145/3581783.3612835

Howell, P., Sackin, S., and Glenn, K. (1997a). Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: II. ANN recognition of repetitions and prolongations with supplied word segment markers. *J. Speech Lang. Hear. Res.* 40, 1085–1096. doi: 10.1044/jslhr.4005.1085

Howell, P., Sackin, S., and Glenn, K. (1997b). Development of a two-stage procedure for the automatic recognition of dysfluencies in the speech of children who stutter: I. Psychometric procedures appropriate for selection of training material for lexical dysfluency classifiers. *J. Speech Lang. Hear. Res.* 40, 1073–1084. doi: 10.1044/jslhr.4005.1073

Howell, P., Staveley, A., Sackin, S., and Rustin, L. (1998). Methods of interval selection, presence of noise and their effects on detectability of repetitions and prolongations. *J. Acoust. Soc. Am.* 104, 3558–3567. doi: 10.1121/1.423937

Ifeachor, E. C., and Jervus, B. W. (2002). *Digital signal processing: a practical approach*, 2nd edn. (Peason Education).

Ingham, R. I., Cordes, A. K., and Patrick, F. (1993). Time-interval measurement of stuttering. *J. Speech Lang. Hear. Res.* 36, 1168–1176. doi: 10.1044/jshr.3606.1168

Jeni, L. A., Cohn, J. F., and Torre, F. D.La (2013). Facing imbalanced data–recommendations for the use of performance metrics. In 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction

Kingma, D. P., and Ba, J. (2014). Adam: A method for stochastic optimization. *J. Fluen. Disord.* arXiv preprint arXiv:1412.6980

Kully, D., and Boberg, E. (1988). An investigation of interclinic agreement in the identification of fluent and stuttered syllables. *J. Fluen. Disord.* 13, 309–318. doi: 10.1016/0094-730X(88)90001-0

Lea, C., Mitra, V., Joshi, A., Kajarekar, S., and Bigham, J. P. (2021). Sep-28k: a dataset for stuttering event detection from podcasts with people who stutter. in ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 6798–6802.

Mahesha, P., and Vinod, D. S. (2016). Gaussian mixture model based classification of stuttering dysfluencies. *J. Intell. Syst.* 25, 387–399. doi: 10.1515/jisys-2014-0140

Mishra, N., Gupta, A., and Vathana, D. (2021). Optimization of stammering in speech recognition applications. *Int J Speech Technol* 24, 679–685. doi: 10.1007/s10772-021-09828-w

Obaid, H. S., Dheyab, S. A., and Sabry, S. S. (2019). The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In IEMECON 2019 – 9th Annual Information Technology, Electromechanical Engineering and Microelectronics Conference, 279–283

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023a). *Getting the top few transcription results.* GitHub. Available at: https://github.com/openai/whisper/discussions/478

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023b). Robust speech recognition via large-scale weak supervision. In International conference on machine learning (PMLR), 28492–28518.

Riley, G. (2009). *SSI-4: stuttering severity instrument. 1st* Edn. London, England: PRO-ED, an International Publisher.

Rohanian, M., and Hough, J. (2021). Best of both worlds: making high accuracy non-incremental transformer-based disfluency detection incremental. In Proceedings of the 59th Annual Meeting of the Association for Computational

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 3693–3703.

Schuller, B. W., Batliner, A., Amiriparian, S., Bergler, C., Gerczuk, M., Holz, N., et al. (2022). The ACM multimedia 2022 computational paralinguistics challenge: vocalisations, stuttering, activity, & mosquitoes 31, 9635–9639.

Sheikh, S. A., Sahidullah, M., Hirsch, F., and Ouni, S. (2022). Machine learning for stuttering identification: review, challenges and future directions. *Neurocomputing* 514, 385–402. doi: 10.1016/j.neucom.2022.10.015

Thomas, M., Hollands, S., Blackburn, S., and Christensen, H. (2023). Towards disfluency features for speech technology based automatic dementia classification. In Proceedings of the 20th International Congress of Phonetic Sciences (Prague), 3903–3907.

Tyagi, V., and Wellekens, C. (2005). On desensitizing the Mel-cepstrum to spurious spectral components for robust speech recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 1

Wei, W. W. S. (2019). “Principle component analysis of multivariate time series” in *Multivariate time series analysis and applications*, Tsay, R. S., & Weisberg, S. eds. Wiley Series in Probability and Statistics, (New Jersey, USA: John Wiley & Sons, Ltd), 139–161.

frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 06 June 2024
DOI 10.3389/fpsyg.2024.1114811

OPEN ACCESS

EDITED BY
Ratree Wayland,
University of Florida, United States

REVIEWED BY
Gerard H. Poll,
Miami University, United States
Douglas Owen Cheyne,
University of Toronto, Canada

*CORRESPONDENCE
Jun Wang
✉ jun.wang@austin.utexas.edu

RECEIVED 02 December 2022
ACCEPTED 24 May 2024
PUBLISHED 06 June 2024

CITATION
Dash D, Teplansky K, Ferrari P, Babajani-Feremi A, Calley CS, Heitzman D, Austin SG and Wang J (2024) Automatic detection of ALS from single-trial MEG signals during speech tasks: a pilot study.
Front. Psychol. 15:1114811.
doi: 10.3389/fpsyg.2024.1114811

Automatic detection of ALS from single-trial MEG signals during speech tasks: a pilot study

Debadatta Dash[1], Kristin Teplansky[2], Paul Ferrari[3], Abbas Babajani-Feremi[4], Clifford S. Calley[1], Daragh Heitzman[5], Sara G. Austin[1] and Jun Wang[1,2]*

[1]Department of Neurology, Dell Medical School, University of Texas at Austin, Austin, TX, United States, [2]Department of Speech, Language, and Hearing Sciences, University of Texas at Austin, Austin, TX, United States, [3]Helen DeVos Children's Hospital, Corewell Health, Grand Rapids, MI, United States, [4]Department of Neurology, University of Florida, Gainesville, FL, United States, [5]MDA/ALS Center, Texas Neurology, Austin, TX, United States

Amyotrophic lateral sclerosis (ALS) is an idiopathic, fatal, and fast-progressive neurodegenerative disease characterized by the degeneration of motor neurons. ALS patients often experience an initial misdiagnosis or a diagnostic delay due to the current unavailability of an efficient biomarker. Since impaired speech is typical in ALS, we hypothesized that functional differences between healthy and ALS participants during speech tasks can be explained by cortical pattern changes, thereby leading to the identification of a neural biomarker for ALS. In this pilot study, we collected magnetoencephalography (MEG) recordings from three early-diagnosed patients with ALS and three healthy controls during imagined (covert) and overt speech tasks. First, we computed sensor correlations, which showed greater correlations for speakers with ALS than healthy controls. Second, we compared the power of the MEG signals in canonical bands between the two groups, which showed greater dissimilarity in the beta band for ALS participants. Third, we assessed differences in functional connectivity, which showed greater beta band connectivity for ALS than healthy controls. Finally, we performed single-trial classification, which resulted in highest performance with beta band features (~ 98%). These findings were consistent across trials, phrases, and participants for both imagined and overt speech tasks. Our preliminary results indicate that speech-evoked beta oscillations could be a potential neural biomarker for diagnosing ALS. To our knowledge, this is the first demonstration of the detection of ALS from single-trial neural signals.

KEYWORDS

amyotrophic lateral sclerosis, beta oscillation, functional connectivity, magnetoencephalography, speech

1 Introduction

Amyotrophic lateral sclerosis (ALS), also known as Lou Gehrig's disease, causes rapidly progressive upper and lower motor neuron degeneration, thereby disrupting the ability of the brain to control voluntary motor function leading to dysphagia (disordered swallowing), dysarthria (disordered speech), impaired limb function, poor respiratory function, and ultimately fatality (Kiernan et al., 2011). The disease is categorized by significant across-patient heterogeneity in onset region, pattern, and rate of progression (Ravits et al., 2007). There is currently no universal standard for early detection or for monitoring the progression of ALS (Nzwalo et al., 2014; Malekzadeh, 2021). Due to the lack of a biomarker, patients with ALS are often initially misdiagnosed (up to 45% of the time) and their diagnosis can be delayed up to 12 months (Iwasaki et al., 2001).

Regardless of the focality of motor neuron degeneration at clinical onset, progressive bulbar motor deterioration is common in most patients with ALS, which leads to dysarthria (Green et al., 2013). Thus, the identification of a speech-motor biomarker for early detection of ALS has been an active area of research recently (An et al., 2018; Vieira et al., 2019; Stegmann et al., 2020). The degree to which clinicians can identify speech impairments in ALS using perceptual characteristics of speech (e.g., listening for deviations in articulation, voice quality, resonance, and prosody) is only moderately reliable (Allison et al., 2017). Early detection and monitoring of the progression of bulbar symptoms based on behavioral observations remain limited because oral-motor functional changes may not occur until muscle weakness progresses to a critical level (DePaul and Brooks, 1993; Green et al., 2013). However, physiologically, these subtle symptoms could be identified earlier by quantifying the neural activity pattern changes during speech tasks.

There have been intense investigations for diagnostic and prognostic biomarkers in the brain that can provide evidence for ALS mechanisms and thus novel targets for therapeutic intervention. Studies using functional magnetic resonance imaging (fMRI) have shown evidence of increased functional connectivity in ALS patients (Konrad et al., 2002; Lulé et al., 2007; Verstraete et al., 2010; Agosta et al., 2013). Similar findings have been reported using resting-state electroencephalography (EEG) (Iyer et al., 2015; Fraschini et al., 2016; Dukic et al., 2021) and magnetoencephalography (MEG) (Proudfoot et al., 2018; Sorrentino et al., 2018). Using MEG during a spinal motor task, another study demonstrated intensified cortical beta desynchronization followed by a delayed rebound for participants with ALS (Proudfoot et al., 2017), which hinted that beta-band oscillation may be used as an early distinguishing cortical feature for ALS. Such prior neuroimaging studies have provided tremendously impactful insights toward a better understanding of the major mechanisms of neurodegeneration due to ALS in the attempt to identify a neural biomarker. However, most neuroimaging studies have focused on group-level connectivity analyses during resting-state or spinal motor tasks. How cortical activation is impacted by ALS during speech-motor tasks has not been investigated. In addition, it is unknown if single-trial detection of ALS from neural signals is possible. In theory, single-trial ALS detection could instantly diagnose ALS in real-time thereby strengthening medical treatments for ALS. Classifying ALS on a single-trial basis involves training a machine learning model with multiple samples/trials of a quantifiable objective marker that can efficiently predict a sample/trial as ALS or healthy after proper training. Single-trial detection using machine learning has shown great potential in several neural disorders including major depressive disorder (MDD) (Liu et al., 2022), autism spectrum disorder (ASD) (Ezabadi and Moradi, 2021), post-traumatic stress disorder (PTSD) (Georgopoulos et al., 2010), schizophrenia (Xu et al., 2013), amongst other neurologic disorders (Aoe et al., 2019).

In this study, we investigated cortical differences between healthy and ALS brain signals during overt (involving bulbar motor coordination) and imagined speech (without motor involvement). The assumption is that there is a cortical disturbance during motor functions in the early stage of ALS which has been shown in previous studies (Kew et al., 1993; Mills and Nithi, 1997; Geevasinga et al., 2016; Shibuya et al., 2016; Eisen et al., 2017). Here, we used speech-motor tasks to trigger the disturbance and then detect the presence of ALS using machine learning. To our knowledge, this is the first study to use functional neuroimaging data during speech tasks for ALS detection. We examined cortical differences between healthy controls and patients with ALS using the following approaches: (1) signal correlation across sensors, (2) band power distance estimation for individual neural oscillations, (3) functional connectivity analysis, and (4) single-trial classification of ALS and healthy samples using machine learning. These approaches have been widely used in the literature to examine cortical differences between neurotypical controls and patients with neural disorders (Bob et al., 2010; Proudfoot et al., 2017, 2018; Aoe et al., 2019). Using these approaches, we found significant cortical differences between patients with ALS and healthy controls, particularly in the beta band MEG activity, which we detected at the single-trial level.

2 Materials and methods

2.1 Data collection

This study included data collected from three healthy volunteers (1 female; 52 ± 14 years) and three patients with ALS (1 female; 52 ± 12 years); see Table 1. Informed consent in accordance with the ethical committee of the participating institutions was collected from all the participants prior to data collection. The patients with ALS were in the early to mid-stage of the disease. A certified neurologist confirmed the diagnosis of ALS (one bulbar onset, one spinal onset, and one had generalized ALS symptoms). All the patients had a mild, but noticeable speech impairment (Table 1). Speech intelligibility was auditorily evaluated by a speech-language pathologist trainee who is not familiar with these patients. A commonly used software, Sentence Intelligibility Test (SIT), was used in this procedure. SIT first generated a randomized list of sentences with an increasing length from 5 to 15 words (Yorkston et al., 1996). The listener typed down what they heard from the patient's recording in the SIT software. The software then automatically calculated the percentage of correct words (speech intelligibility) as well as speaking rate.

MEG (Neuromag TRIUX; MEGIN, LCC) was used to collect the neuromagnetic signals from the participants (Figure 1). This device has 306 SQUID sensors (204 gradiometers and 102 magnetometers). A magnetically shielded room (MSIR) housed the MEG machine to restrict external magnetic noise. A digital light processing projector was used to present the visual stimuli approximately 90 cm from the subjects on a back projection screen. The stimuli were generated by a computer running the STIM2 software (Compumedics, Ltd.). Two pairs of bipolar EEG electrodes were used to record the electrocardiogram (EKG) and the electrooculogram (EOG) signals. A custom air-pressure transducer located outside the MSR and connected to the analog input of the MEG system was used to measure jaw displacement during the tasks. An air-bladder was fixed under the subjects' chin and relayed jaw movement (via pressure on the bladder) to the transducer via tubing connected to the air-inlet on the sensor. Voice data was recorded using a standard built-in microphone

TABLE 1 Demographics of ALS patients.

Participant	Gender	Age (years)	Speech intelligibility (%)	Speaking rate (words/min)
A1	M	56	71.81	116.83
A2	F	39	100.00	179.45
A3	M	61	92.00	132.53

SI: Speech Intelligibility; SR: Speaking rate; wpm: words per minute.

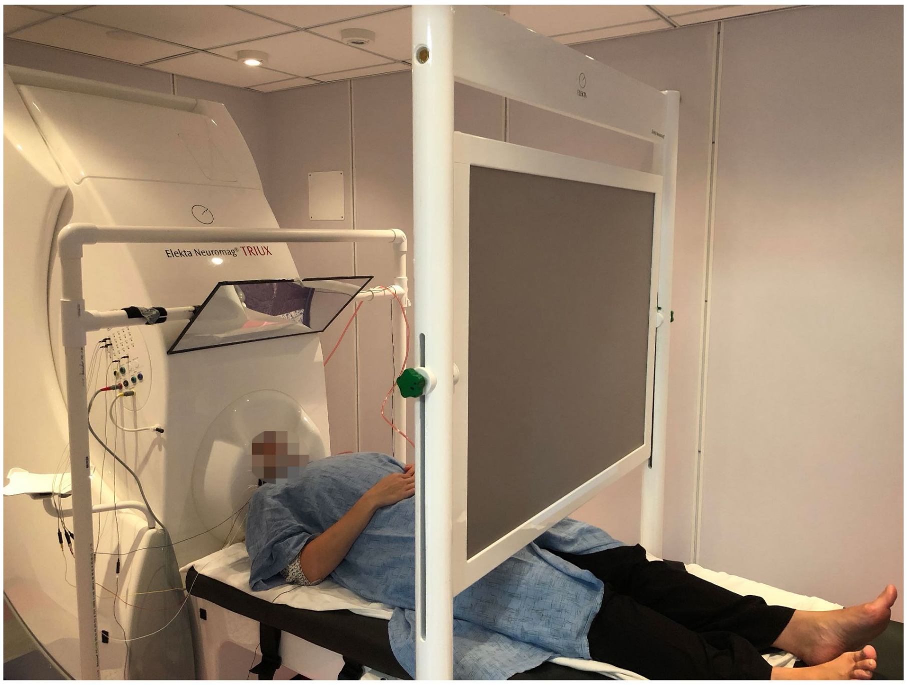


FIGURE 1
The MEG scanner and a subject with ALS.

connected to a transducer placed outside the MSR. Both voice and jaw movement analog signals were then digitized by feeding into the MEG ADC in real-time as separate channels. Five commonly used phrases were used as stimuli for the speech tasks: *1. Do you understand me; 2. That's perfect; 3. How are you? Good-bye; 5. I need help.* The task phrases came from phrase lists commonly used in alternative augmented communication (AAC) devices and were selected to be more familiar to the patients and easier to recite than novel speech (Beukelman et al., 1984; Dash et al., 2020). The experiment was designed as a time-locked delayed overt reading task where each trial was time-locked to stimulus onset (display of phrases on the screen). The phrases were individually presented for 1 s in a pseudorandomized order followed by a 1 s fixation cross. The subjects were previously instructed to think of speaking the phrase without mouthing during the fixation and to overtly articulate the phrase at their normal speaking rate and loudness when the fixation disappeared. The subjects had 3 s to perform the articulation before the next stimulus trial. Each participant completed 100 trials per phrase. To overcome potential difficulties verifying the timing of imagined speech (Cooney et al., 2018), we designed our protocol to collect both speech imagination and speech production consecutively, in the same trial and under time constraints.

The MEG data were recorded with 4 kHz sampling frequency with an online filter of 0.3–1,330 Hz. The data were low pass filtered to 250 Hz with a 4th order Butterworth filter and resampled to 1 kHz. Power line noise (60 Hz) and harmonics were removed with a 2nd order infinite impulse response (IIR) notch filter. Only gradiometer sensors were used for analysis. From the 204 gradiometer sensors, it was observed that four sensors exhibited substantial channel noise during the data collection process from various participants. Additionally, in certain cases, one or two additional sensors displayed irregularities resembling artifacts. Consequently, a total of eight sensors were deemed unsuitable and excluded from the analysis. The discarded sensors were the same for both ALS and healthy data. Therefore, the analysis was conducted using data exclusively from 196 sensors. Independent component analysis (ICA) was used to remove artifacts (cardiac activity, eye blinks, and saccades) from the data. The continuous MEG signals were epoched into trials from −0.5 to +4.5 s centered at stimulus onset. Covert speech segment was parsed as the data from 1 s to 2 s and overt speech segment was parsed as the data from 2 s to 4.5 s of each trial. By visually inspecting the data, trials were discarded if they contained high-amplitude artifacts or if the participant did not comply with the paradigm timing (e.g., the participant spoke before being provided the cue to articulate). Jaw movement data during the covert speech segment was used to verify that the participants were not moving their articulators during the covert speech task. Jaw movement data were not used for analysis in this study. Following preprocessing, a single participant's dataset contained only 63 valid trials for a particular phrase. Therefore, to ensure an impartial comparison, we exclusively considered the initial 60 trials per phrase per participant. The preprocessing of the raw MEG data was conducted using FieldTrip (Oostenveld et al., 2011) in MATLAB 2021b.

2.2 Data analysis

2.2.1 Sensor correlation

Sensor correlation has been often used to characterize neurological disorders (Schindler et al., 2007; Bob et al., 2010). Here, we computed Pearson's correlation between each pair of gradiometer sensor signals. Analyses were performed for both speech imagination and speech production for each stimulus (phrase) and participant separately. Correlation values were computed at the single-trial level and then averaged across all trials. For this analysis, we used all the spectral information (0.3–250 Hz) in the signals. Statistical 2-sample t-tests were used to compare the ALS and healthy groups (N = 15: 3 participants × 5 phrases) based on number of sensors showing larger absolute correlation coefficients ($r > 0.5$) and correlation density (sum of all absolute correlation values over total number of sensor-pairs) for both imagined and overt speech separately.

2.2.2 Band power distance

Each neural oscillation is associated with a key functional role in the brain and could potentially carry a neural biomarker of a disorder. Beta-band power has traditionally been associated with motor function in the brain (Fisher et al., 2012; Khanna and Carmena, 2015). Thus, for the speech-motor task (overt speech) and the speech-motor imagination task (imagined speech), we compared power in this band and other canonical bands between the two groups. We computed the average power of the neuromagnetic signals for each frequency range of interest: delta (1–4 Hz), theta (4–8 Hz), alpha (8–16 Hz), beta (16–30 Hz), gamma (30–59 Hz), and high gamma (61–119 Hz). We then averaged the band powers (this was completed separately for each band) across both trials and participants. The pairwise Euclidian distances between healthy and ALS band powers were calculated across all sensors for each phrase and after averaging across the 5 phrases. 1-way analysis of variance (ANOVA) and post-hoc Tukey test was conducted with the six bands as independent groups and 5 phrases as different samples for both imagination and articulation.

2.2.3 Functional connectivity

Functional connectivity is defined as the statistical dependence among measured neural signals which explains the temporal coincidence of spatially distant neurophysiological events (Friston, 1994). Functional connectivity analysis has become the conventional choice for a better understanding of the *in vivo* pathology of ALS. In this study, we used amplitude envelope correlation (AEC) (O'Neill et al., 2015) to measure the functional connectivity for each frequency band. For single-trial functional connectivity analysis, we used a 4th order Butterworth bandpass filter to first bandpass the gradiometer signals from all 196 sensors at each frequency range of interest, obtained the amplitude envelopes using Hilbert transform, and then computed the pairwise linear correlation of the amplitude envelopes across all sensors for each frequency range of interest separately. Connectivity was defined as the averaged pair-wise correlation across trials. For the individual subject analysis, first, we temporally concatenated all bandpass-filtered single trials, extracted the envelope, and then computed the correlations. We performed the AEC-based functional connectivity analysis for each phrase separately during both imagined and overt speech. A 2-sample one-sided t-test was conducted between healthy and ALS samples (3 subjects × 5 phrases—for each group) of functional connectivity density (sum of AEC values over total number of sensor pairs) to check for the hypothesis of whether patients with ALS show greater beta band connectivity than healthy controls.

2.2.4 Single-trial classification

We used power in the six canonical frequency bands of the MEG signals as features to train a linear discriminant analysis (LDA) algorithm and classified ALS and healthy data during both speech imagination and overt speech. We trained the model separately for each frequency range of interest and separately using a wide frequency range (0.3–250 Hz) which contained spectral information from all the neural oscillations. The choice of the LDA model was inspired by our previous work on speech decoding for ALS where the LDA model performed equivalently to both support vector machines and multilayer perceptron classifiers (Dash et al., 2020) at classifying 5 phrases. The *fitcdiscr* function in the Statistical and Machine Learning Toolbox of MATLAB was used for classification. The lower sample size than the feature dimension motivated for a linear type of discriminant. The linear coefficient threshold ('Delta') and the amount of regularization ('Gamma') of the model were tuned as the hyperparameters of the model, computed based on the Bayesian optimization search using a 10-fold cross-validation on the training data. All other parameters were set to the default values of the toolbox. We used a leave one-pair out cross-validation strategy where we trained the model with all trials from 2 healthy and 2 ALS participants and tested using the remaining data from 1 healthy and 1 ALS participant, irrespective of the phrase. This was repeated until each healthy-ALS pair was tested. This led to a training data size of 1,200 trials (4 participants (2 healthy +2 ALS) × 5 phrases × 60 trials) and a test data size of 600 trials (2 participants (1 healthy +1 ALS) × 5 phrases × 60 trials) for each fold. In this manner, the trained decoder was tested with completely unseen new participant data.

3 Results

Figure 2 shows the comparative histogram distribution of sensor-level signal correlations for ALS and healthy controls for each phrase (top for imagined speech and bottom for overt speech). A significantly larger number of sensors showed greater correlations for ALS compared to healthy controls across all phrases during both overt (one-sided, 2-sample t-test: $t = 3.76$, $df = 28$, $p < 0.001$) and imagined speech (one-sided, 2-sample t-test: $t = 6.01$, $df = 28$, $p < 0.001$). This is also evident by the higher variance in the distribution of correlations for ALS compared to healthy controls for both imagined and overt speech across all phrases. In other words, the majority of the correlations were near mean (i.e., zero correlation) for the controls compared to ALS. For imagined speech, 95% (Bayesian analysis based on Monte Carlo simulations) of the correlation values were in a range of −0.5 to 0.5 for healthy participants. The range was between −0.8 to 0.8 for ALS participants. For overt speech, the correlation range for

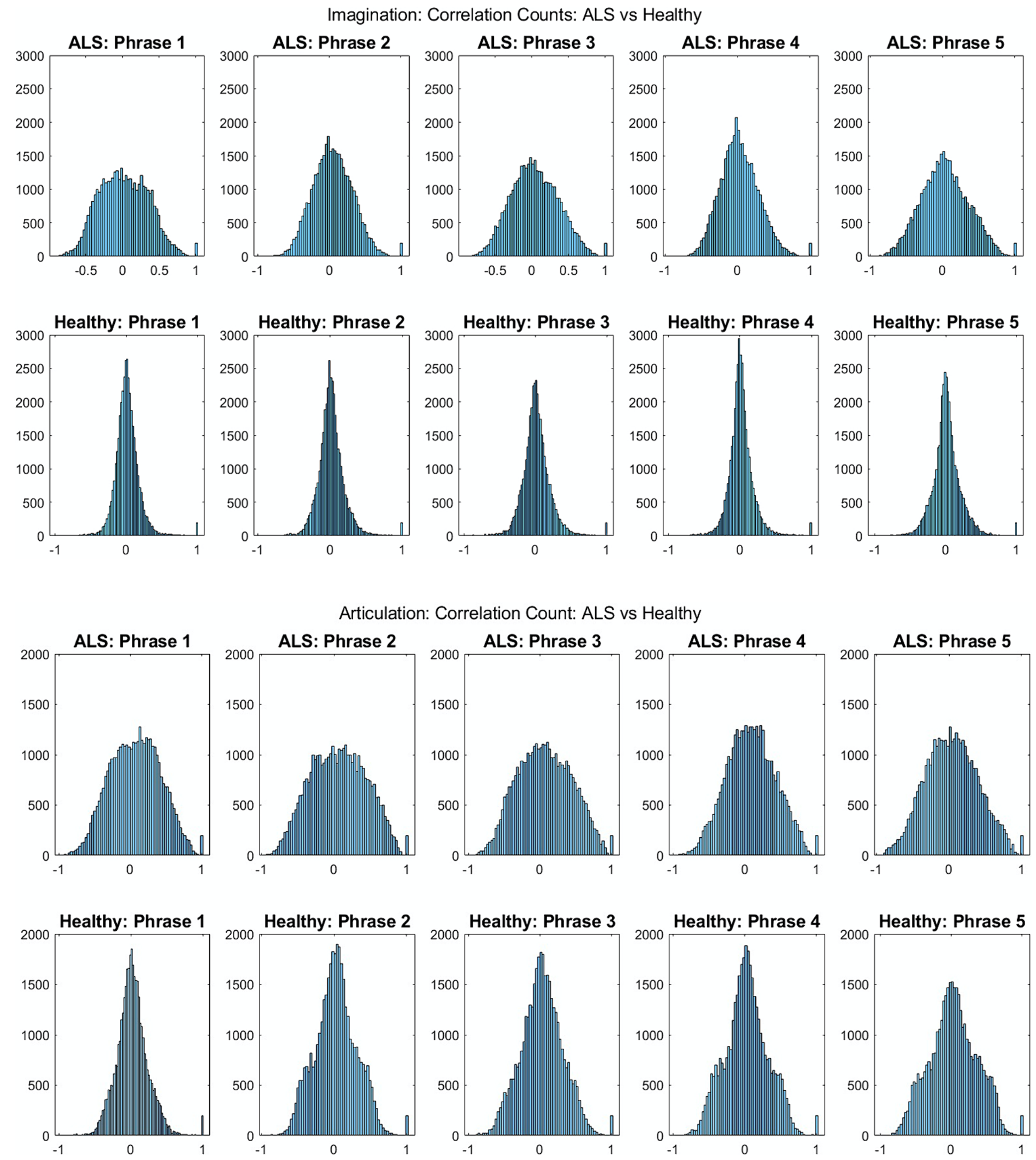


FIGURE 2
Histogram distribution of all pair-wise sensor correlations of healthy and ALS participants for each phrase during imagined speech (top) and overt speech (articulation) (bottom) for patients with ALS and healthy controls, respectively.

healthy controls was approximately within the range − 0.8 to 0.8, which was greater for the ALS ranging from −1 to 1. A heatmap plot of the correlation distribution for each subject is shown in Figure 3 (Top for imagined speech and middle for overt speech), which depicts stronger correlations across the whole brain for participants with ALS compared to the healthy controls, especially for the first participant with ALS (A1) who also had the lowest speech intelligibility and speaking rate scores (Table 1). To interpret these correlation heatmaps, correlation density was calculated as the sum of all absolute correlation values over total number of sensor-pairs and shown for each participant in Figure 3—Bottom panel. Mean correlation density was higher for participants with ALS (Overt: 0.428; Imagination: 0.326) compared to healthy subjects (Overt: 0.292; Imagination: 0.192) averaged across trials, phrases, and participants as well as statistically across all phrases and participants (one-sided, 2 sample *t*-tests: overt: $t = 6.13$, $df = 28$, $p < 0.001$; imagined: $t = 6.68$, $df = 28$, $p < 0.001$). As expected, a stronger correlation for overt speech was observed compared to speech imagination, irrespective of healthy or ALS data.

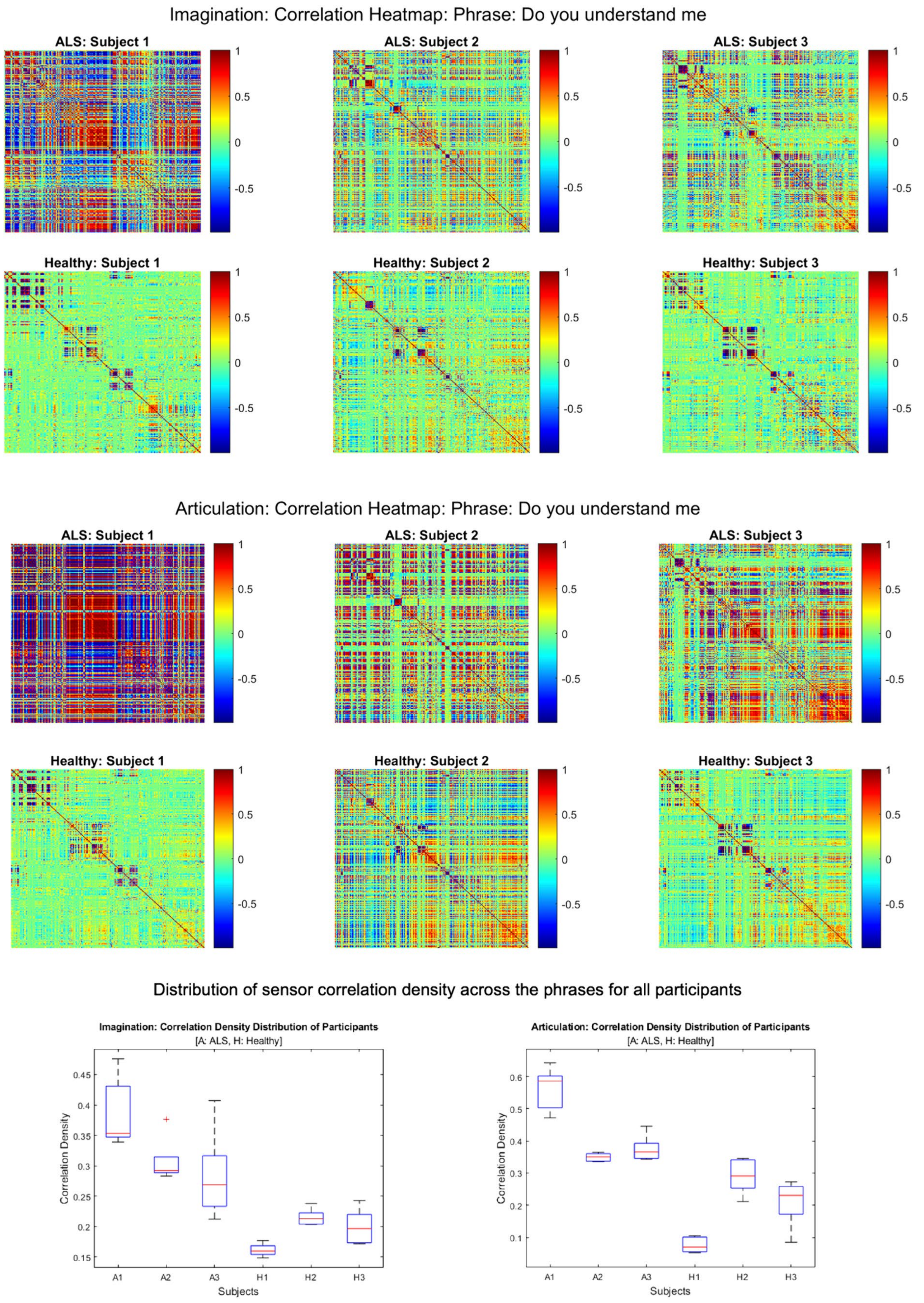


FIGURE 3
Heatmap of pair-wise sensor correlations for each subject on phrase “Do you understand me” for imagination (top) and articulation (middle), and the distribution of sensor density across the phrase for all participants (bottom). In the heatmaps, each colorful dot/point represents the correlation

(Continued)

FIGURE 3 (Continued)
between a pair of the 196 gradiometers. In the bottom panel, correlation density was calculated as the sum of all absolute correlation values over total number of sensor-pairs across five phrases for each participant.

Figure 4 shows the mean band-power distances between healthy and ALS for each during overt speech task (articulation), where panels (A) to (F) are for delta, theta, alpha, beta, gamma, and high gamma bands, respectively. For better visualization, the distances are shown as heatmaps where the color range from blue to red indicates minimum to maximum range of the normalized distance values. Each cell in the heatmap represents a pairwise distance between the band powers of a healthy sensor (y-axis) and an ALS sensor (x-axis) across all the phrases. The distances were significantly greater for the beta band powers than the other canonical bands for both imagination (1-way ANOVA: $F = 208.55$, $p < 0.001$; post-hoc Tukey tests: beta vs. rest: $p < 0.0019$) and articulation (1-way ANOVA: $F = 206.59$, $p < 0.001$; post-hoc Tukey tests: beta vs. rest: $p < 0.0021$, see Figure 4—Panel G and H). Also, a larger number of pairwise (ALS—healthy) dissimilarities were observed in the beta band. These oscillatory patterns were similar for each phrase and across all phrases for individual and group subject analysis irrespective of speech task, i.e., imagination or production. A couple of sensors showed the highest distance (solid red lines in the heatmap) which could be because those sensors were noisy.

Figure 5 shows the AEC-based beta-band functional connectivity for both groups during the production of the phrase '*Do you understand me?*' in the form of heatmaps; showing the correlation range of −1 to 1 (from blue [minimum] to red [maximum]). Greater beta band connectivity was significant for ALS patients compared to healthy subjects (one sided, 2-sample *t*-test: $p < 0.05$; see Supplementary Figure S3 for distributions of connectivity strengths for both patients with ALS and healthy controls). This is notably apparent in the first patient (A1), who had more severe bulbar impairment than the other two (A2 and A3). Interestingly, similar patterns of increased connectivity were also prominent during speech imagination (Supplementary Figure S2). A more diverse connectivity pattern among the 3 patients with ALS compared to the healthy participants can be observed by visualizing the connectivity strengths.

Figure 6 shows the median single-trial classification accuracy for the healthy versus ALS group for both speech imagination and overt speech tasks. The best performance (median accuracy ~98%) was obtained using beta bands and was similar for both speech tasks. The performance using each individual frequency range of interest (excluding delta) was significantly higher than chance level (50%) and was also higher when compared to performance using all frequency information (all: 0.3–250 Hz). The distribution of the test performance for different folds (i.e., for each pair of ALS-healthy single-trial test accuracy) is shown in Supplementary Figure S4. The performance accuracy was lowest for the first ALS patient (A1) (mean across folds = 65% for overt speech; 83% for imagined speech), likely because this participant's speech symptoms were severe compared to the other two participants with ALS. Although median performance was highest for beta band, statistically, 1-way ANOVA based comparison did not show a significant difference between the performances of different bands ($F = 1.33$, $p = 0.05$), possibly due to the low sample size. For the case of imagined speech, performances obtained with theta and gamma band were comparable to beta band performance.

4 Discussion

The evidence of greater inter-sensor correlation for ALS compared to healthy participants is a clear distinguishable marker between the two groups. This has been previously observed with M/EEG resting state (Proudfoot et al., 2019) and motor imagery studies (Yang et al., 2018). This difference in sensor correlations was apparent across all phrases and participants which further illustrates that this feature is independent of stimuli and an across-subject observation. A stronger correlation during the overt speech task compared to the imagined speech task indicated greater cortical activity for producing overt speech compared to speech imagination, which was true for both healthy and ALS groups and was expected. The signal artifacts introduced by movement during the production of speech gestures could have also contributed to the higher sensor correlation during overt speech production (Dash et al., 2018). Participant (A1) with the most severe symptoms (lowest speech intelligibility and speaking rate scores: Table 1) showed the strongest correlation (Figure 3) suggesting that the proposed approach may be useful as a marker of disease progression. We must note that a larger sample size is needed to statistically validate this observation. Further, sensor correlation differences could also arise from the differences in the head positions inside the scanner. Mapping the sensor data into source space and performing the correlations across parcels/voxels would be a better way to remove these confounds, as planned for future studies.

Beta band has been traditionally associated with motor function, and the observed differences in the beta band power during overt speech are consistent with the hypotheses that ALS is associated with cortical hyperexcitability, possibly due to the loss of inhibitory interneuron (Proudfoot et al., 2017). Our results reproduced the importance of beta band for identifying ALS during a bulbar motor task (speech). The prominent beta band differences during the speech imagination task (which does not involve motor execution) (Supplementary Figure S1) suggest that the beta band during speech tasks that involve motor planning could be a potential neural biomarker of ALS. Crucially, the pattern of band-power differences was similar for both imagination and overt speech in the beta band, possibly indicating a functional similarity between the two speech tasks. Clear differences in band power were also observed in the theta and the gamma band; however, they were less prominent compared to the beta band differences.

This finding of significantly greater beta band connectivity in the ALS group compared to healthy controls was expected since beta band functional connectivity changes have been previously shown (Verstraete et al., 2010; Agosta et al., 2013; Proudfoot et al., 2017) both during resting state as well as for spinal motor tasks. An increase in beta band functional connectivity has been hypothesized as the result from loss of intracortical inhibitory influence supported *in vivo* by neurophysiology findings of accentuated cortical beta-desynchronization during movement preparation and diminished post-movement beta-rebound (Proudfoot et al., 2017). This inhibitory influence may lead to compensatory mechanisms in early-stage ALS resulting in higher functional connectivity. Behaviorally, it may be explained as a

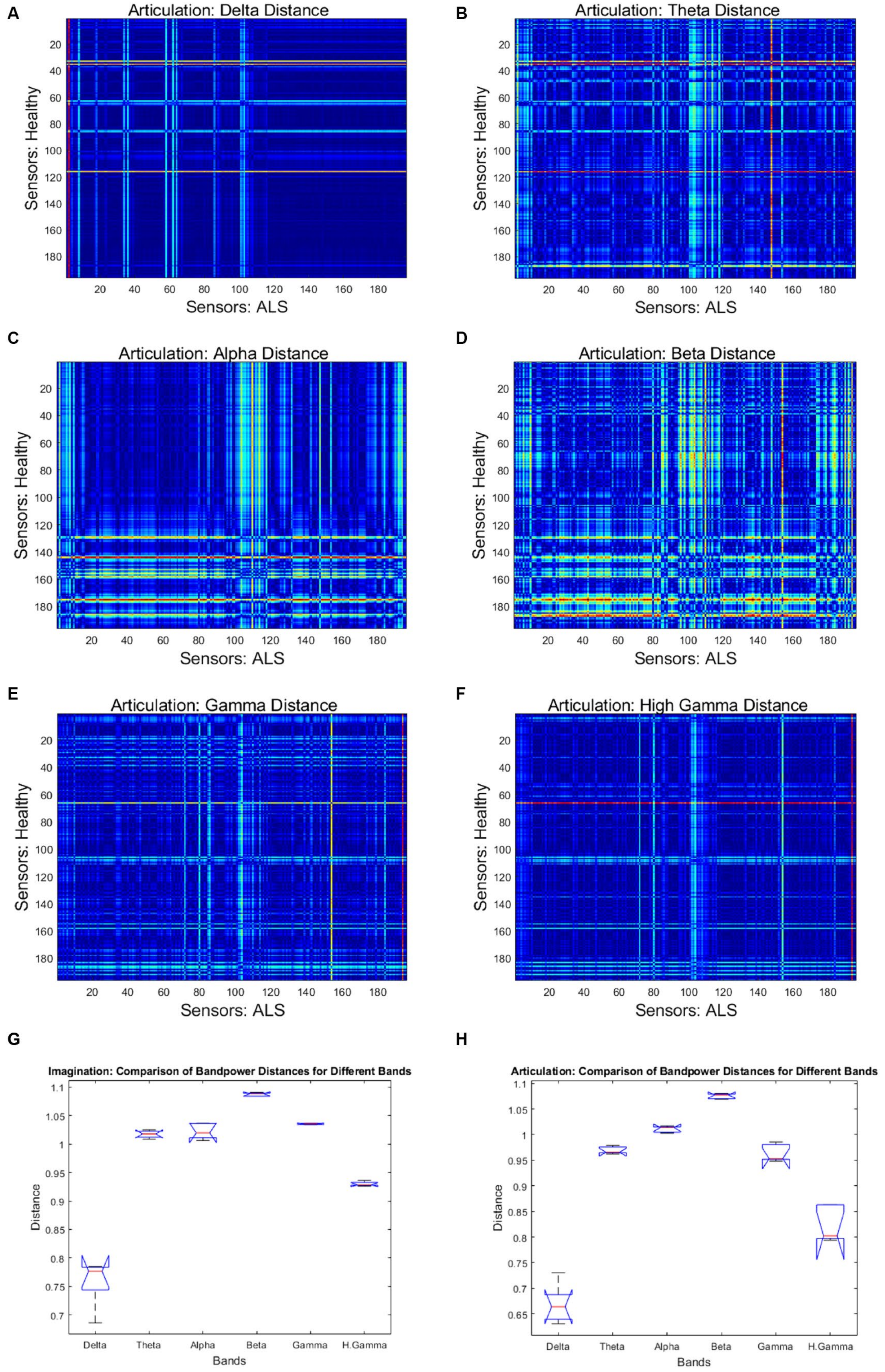


FIGURE 4
Heatmap of pairwise band-power distances between healthy and ALS sensors for each band during articulation. **(A)** Delta; **(B)** Theta; **(C)** Alpha; **(D)** Beta; **(E)** Gamma; **(F)** High Gamma. In the heatmaps, each colorful dot/point represents the bandpower distance between a pair of the 196 gradiometers. The bottom panels provide the bandpower distances for all frequency bands in speech imagination **(G)** and articulation task **(H)**, respectively.

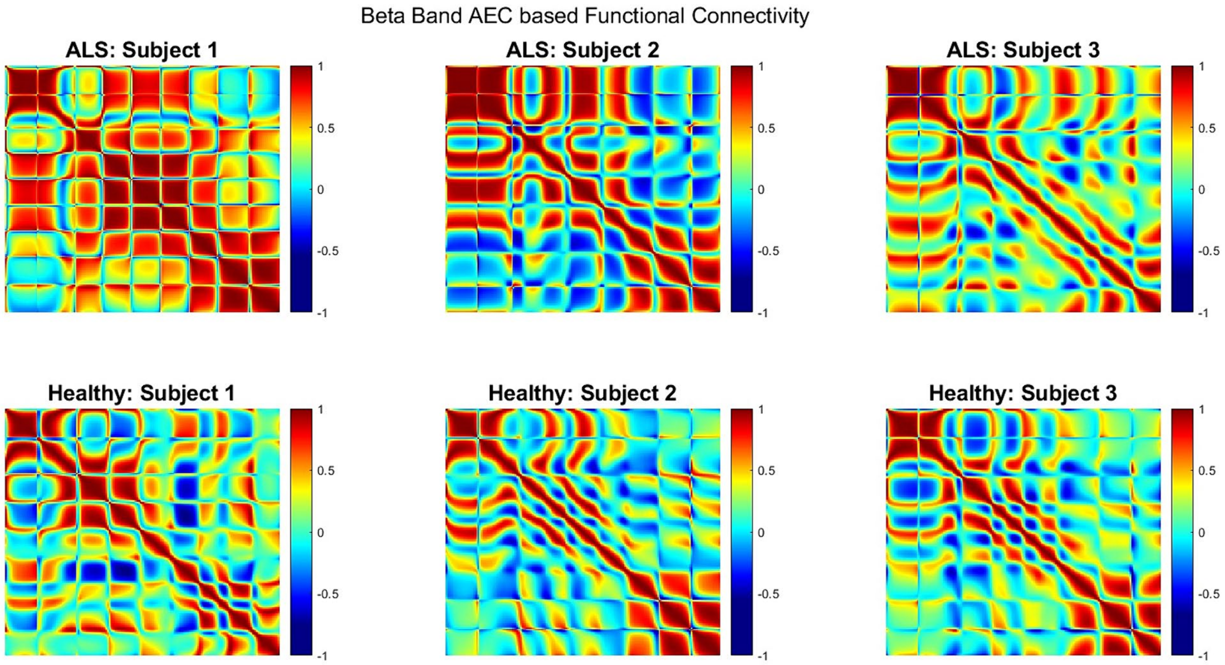


FIGURE 5
Heatmap of beta band AEC based functional connectivity across all sensors for participants with ALS (top row) and healthy controls (bottom row) during the overt speech task.

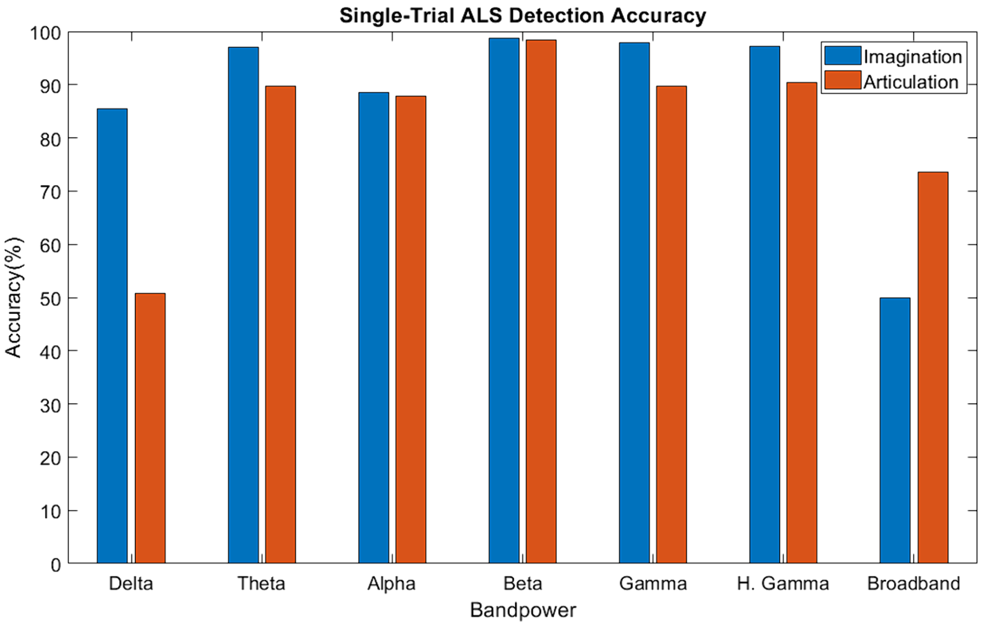


FIGURE 6
Single-trial ALS detection accuracy using band power [Delta: 1–4 Hz; Theta: 4–8 Hz; Alpha: 8–16 Hz; Beta: 16–30 Hz; Gamma: 30–59 Hz; High Gamma: 61–119 Hz; Broadband: 1–119 Hz].

compensatory mechanism for speech tasks in early-stage ALS due to the recruitment of larger neural networks, supported by tongue kinematic studies (Green et al., 2013; Kuruvilla-Dugdale and Mefferd, 2017; Teplansky et al., 2019). Additionally, disruption of efficient motor control networks in ALS may lead to higher cognitive control demands and attention increases, both of which are known to modulate beta-band oscillation power and connectivity (Cheyne and Ferrari, 2013; Riddle et al., 2021). This study provides the first evidence of increased beta band connectivity during a speech-motor task. Interestingly, increased connectivity was also prominent during speech imagination

(Supplementary Figure S2) indicating higher beta band connectivity during speech planning, thereby strengthening the role of beta band as a neural biomarker for ALS. Similar to the observations with band-power differences, beta-band functional connectivity was greatest for our most severe patient (participant A1), providing additional confidence in the specificity of this marker. Accumulating the connectivity strength (i.e., correlation values) of all three subjects and for all 5 phrases showed greater connectivity strength for the ALS group than the control group (Supplementary Figure S2) indicating beta-band connectivity to be an across-subject marker.

Evidence of a high single-trial ALS detection accuracy with beta-band suggests that the neural mechanisms for ALS could be specific to spectral content, particularly to the beta band during speech tasks. From the previous qualitative analyses (sensor correlation, band power difference, and functional connectivity) beta-band was expected to perform the best for single-trial classification. The median accuracy with beta band was superior when both overt and imagined phrases were considered, although theta and gamma band also showed comparable performance with beta band for the case of imagined phrases. A recent study suggested covert speech emphasizes both beta and gamma band (Moon et al., 2022), which may explain why gamma band also obtained high accuracies. In short, greater performance accuracies during the speech imagination task suggest that neural signals derived while imagining speech may be optimal for diagnosing early-onset ALS, whereas overt speech may be more appropriate for evaluating the rate of disease progression. Crucially, this is the first demonstration of ALS detection from single-trial neural signals.

In terms of behavior, individuals with ALS exhibited larger onset latency and duration in overt speech tasks when compared to their healthy counterparts (2-sample t-tests, t = 3.09, $p = 0.002$, N = 900 [3 participants × 5 phrases × 60 trials]; Supplementary Figure S5), as one would expect the patients to take longer time to complete the task. It is plausible that these behavioral effects manifest in elevated sensor correlation and functional connectivity strength for ALS patients as opposed to healthy controls. However, there is notable convergence in these behaviors at the single-trial level between the two population groups, with more than 44% overlap in onset time and over 22% overlap in duration. The behavioral difference was mostly driven by the first ALS participant (A1) with the lowest speaking rate and speech intelligibility. Consequently, relying solely on behavioral indicators for single-trial detection proves to be inefficient. In addition, similar cortical differences were also observed during the covert speech task, a scenario where these behavioral markers are absent. Further, covert speech segments are immune to movement artifacts that can be present during overt speech and bias the results. Hence, the optimal approach for single-trial ALS detection involves analyzing neural activity during covert speech tasks.

Although these results are encouraging, this study suffers from a very small sample size and the omission of a non-ALS clinical control group. Future studies should include larger cohorts and include another patient population with a movement disorder, e.g., Parkinson's disease and other motor neural diseases, in order to reveal the specificity of these detection methods. If validated, neuromagnetic signals during speech tasks with machine learning would open a new direction for assisting the diagnosis of ALS. As bulbar onset of ALS represents about 30% of the total and spinal onset accounts for about 70% (Van Es et al., 2017), we plan to combine neural signals during speech and spinal motor tasks (e.g., finger tapping) in future studies. A further step is to combine neuromagnetic signals with (speech) audio (An et al., 2018). Finally, individuals with ALS and other neurological diseases that show some similar symptoms such as Parkinson's disease will be included for differential analysis.

5 Conclusion

In this study, we investigated the neuromagnetic pattern differences between individuals with ALS and healthy subjects during imagined and overt speech tasks, towards identifying a potential neural biomarker. Our preliminary results showed a greater number of sensors with larger correlations, a higher dissimilarity in the beta band power, and a larger beta band connectivity for ALS patients compared to healthy controls. Single-trial ALS detection analysis resulted in the highest median classification accuracy using beta band features, which were significant across trials, phrases, and participants for both speech imagination and articulation. The preliminary results of this study provide a proof of concept for the use of beta band as a potential neural biomarker during speech tasks and machine learning for early detection of ALS.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by IRBs of University of Texas at Austin and University of Texas at Dallas. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

DD implemented the algorithms and drafted the manuscript. PF and JW designed the experimental paradigm for data collection. DH performed ALS diagnosis. DD, KT, PF, AB-F, CC, DH, SA, and JW interpreted the results and performed subsequent editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the UT System Brain Initiative under award 362221 and partly supported by the National Institutes of Health (NIH) under award numbers R03DC013990 and R01DC016621.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The handling editor RW declared a shared affiliation with the author AB-F at the time of review.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1114811/full#supplementary-material

References

Agosta, F., Canu, E., Valsasina, P., Riva, N., Prelle, A., Comi, G., et al. (2013). Divergent brain network connectivity in amyotrophic lateral sclerosis. *Neurobiol. Aging* 34, 419–427. doi: 10.1016/j.neurobiolaging.2012.04.015

Allison, K. M., Yunusova, Y., Campbell, T. F., Wang, J., Berry, J. D., and Green, J. R. (2017). The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to ALS. *Amyotroph Lateral Scler Frontotemporal Degener* 18, 358–366. doi: 10.1080/21678421.2017.1303515

An, K., Kim, M., Teplansky, K., Green, J., Campbell, T., Yunusova, Y., et al. (2018). Automatic Early Detection of Amyotrophic Lateral Sclerosis from Intelligible Speech Using Convolutional Neural Networks. *Proc. Interspeech* 2018, 1913–1917, doi: 10.21437/Interspeech.2018-2496

Aoe, J., Fukuma, R., Yanagisawa, T., Harada, T., Tanaka, M., Kobayashi, M., et al. (2019). Automatic diagnosis of neurological diseases using MEG signals with a deep neural network. *Sci. Rep.* 9:5057. doi: 10.1038/s41598-019-41500-x

Beukelman, D. R., Yorkston, K. M., Poblete, M., and Naranjo, C. (1984). Frequency of word occurbence in communication samples produced by adult communication aid users. *J. Speech hear. disord.* 49, 360–367.

Bob, P., Susta, M., Glaslova, K., and Boutros, N. N. (2010). Dissociative symptoms and interregional EEG cross-correlations in paranoid schizophrenia. *Psychiatry Res.* 177, 37–40. doi: 10.1016/j.psychres.2009.08.015

Cheyne, D., and Ferrari, P. (2013). MEG studies of motor cortex gamma oscillations: evidence for a gamma "fingerprint" in the brain? *Front. Hum. Neurosci.* 7:575. doi: 10.3389/fnhum.2013.00575

Cooney, C., Folli, R., and Coyle, D. (2018). Neurolinguistics research advancing development of a direct-speech brain-computer interface. *iScience* 8, 103–125. doi: 10.1016/j.isci.2018.09.016

Dash, D., Ferrari, P., Malik, S., and Wang, J.. (2018) "Overt speech retrieval from neuromagnetic signals using wavelets and artificial neural networks," *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Anaheim, CA, USA, pp. 489–493.

Dash, D., Ferrari, P., Hernandez, A., Heitzman, D., Austin, S. G., and Wang, J. (2020). Neural Speech Decoding for Amyotrophic Lateral Sclerosis. *Proc. Interspeech* 2020, 2782–2786. doi: 10.21437/Interspeech.2020-3071

DePaul, R., and Brooks, B. R. (1993). Multiple orofacial indices in amyotrophic lateral sclerosis. *J. Speech Lang. Hear. Res.* 36, 1158–1167. doi: 10.1044/jshr.3606.1158

Dukic, S., McMackin, R., Costello, E., Metzger, M., Buxo, T., Fasano, A., et al. (2021). Resting-state EEG reveals four subphenotypes of amyotrophic lateral sclerosis. *Brain* 145, 621–631. doi: 10.1093/brain/awab322

Eisen, A., Braak, H., del Tredici, K., Lemon, R., Ludolph, A. C., and Kiernan, M. C. (2017). Cortical influences drive amyotrophic lateral sclerosis. *J Neurol Neurosurg Psychiatry* 88, 917–924. doi: 10.1136/jnnp-2017-315573

Ezabadi, M. G., and Moradi, M. H. (2021) 'A novel algorithm for detection of social joint attention from single-trial EEG signals of autistic Spectrum disorder (ASD)'. In 2021 28th National and 6th International Iranian Conference on Biomedical Engineering (ICBME). IEEE, pp. 288–293.

Fisher, K. M., Zaaimi, B., Williams, T. L., Baker, S. N., and Baker, M. R. (2012). Beta-band intermuscular coherence: a novel biomarker of upper motor neuron dysfunction in motor neuron disease. *Brain* 135, 2849–2864. doi: 10.1093/brain/aws150

Fraschini, M., Demuru, M., Hillebrand, A., Cuccu, L., Porcu, S., di Stefano, F., et al. (2016). EEG functional network topology is associated with disability in patients with amyotrophic lateral sclerosis. *Sci. Rep.* 6:38653. doi: 10.1038/srep38653

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: a synthesis. *Hum. Brain Mapp.* 2, 56–78. doi: 10.1002/hbm.460020107

Geevasinga, N., Menon, P., Özdinler, P. H., Kiernan, M. C., and Vucic, S. (2016). Pathophysiological and diagnostic implications of cortical dysfunction in ALS. *Nat. Rev. Neurol.* 12, 651–661. doi: 10.1038/nrneurol.2016.140

Georgopoulos, A. P., Tan, H. R. M., Lewis, S. M., Leuthold, A. C., Winskowski, A. M., Lynch, J. K., et al. (2010). The synchronous neural interactions test as a functional neuromarker for post-traumatic stress disorder (PTSD): a robust classification method based on the bootstrap. *J. Neural Eng.* 7:016011. doi: 10.1088/1741-2560/7/1/016011

Green, J. R., Yunusova, Y., Kuruvilla, M. S., Wang, J., Pattee, G. L., Synhorst, L., et al. (2013). Bulbar and speech motor assessment in ALS: challenges and future directions. *Amyotroph Lateral Scler Frontotemporal Degener* 14, 494–500. doi: 10.3109/21678421.2013.817585

Iwasaki, Y., Ikeda, K., and Kinoshita, M. (2001). The diagnostic pathway in amyotrophic lateral sclerosis. *Amyotroph Lateral Scler Other Motor Neuron Disord* 2, 123–126. doi: 10.1080/146608201753275571

Iyer, P. M., Egan, C., Pinto-Grau, M., Burke, T., Elamin, M., Nasseroleslami, B., et al. (2015). Functional connectivity changes in resting-state EEG as potential biomarker for amyotrophic lateral sclerosis. *PLoS One* 10:e0128682. doi: 10.1371/journal.pone.0128682

Kew, J. J. M., Leigh, P. N., Playford, E. D., Passingham, R. E., Goldstein, L. H., Frackowiak, R. S. J., et al. (1993). Cortical function in amyotrophic lateral sclerosis: a positron emission tomography study. *Brain* 116, 655–680. doi: 10.1093/brain/116.3.655

Khanna, P., and Carmena, J. M. (2015). Neural oscillations: beta band activity across motor networks. *Curr. Opin. Neurobiol.* 32, 60–67. doi: 10.1016/j.conb.2014.11.010

Kiernan, M. C., Vucic, S., Cheah, B. C., Turner, M. R., Eisen, A., Hardiman, O., et al. (2011). Amyotrophic lateral sclerosis. *Lancet (London, England)* 377, 942–955. doi: 10.1016/S0140-6736(10)61156-7

Konrad, C., Henningsen, H., Bremer, J., Mock, B., Deppe, M., Buchinger, C., et al. (2002). Pattern of cortical reorganization in amyotrophic lateral sclerosis: a functional magnetic resonance imaging study. *Exp. Brain Res.* 143, 51–56. doi: 10.1007/s00221-001-0981-9

Kuruvilla-Dugdale, M., and Mefferd, A. (2017). Spatiotemporal movement variability in ALS: speaking rate effects on tongue, lower lip, and jaw motor control. *J. Commun. Disord.* 67, 22–34. doi: 10.1016/j.jcomdis.2017.05.002

Liu, M., Tan, J., Jiang, Y., and Tian, Y. (2022). Using deep learning to decode abnormal brain neural activity in MDD from single-trial EEG signals. *Brain-Appar. Commun. J. Bacomics* 1, 28–37. doi: 10.1080/27706710.2022.2075242

Lulé, D., Diekmann, V., Kassubek, J., Kurt, A., Birbaumer, N., Ludolph, A. C., et al. (2007). Cortical plasticity in amyotrophic lateral sclerosis: motor imagery and function. *Neurorehabil. Neural Repair* 21, 518–526. doi: 10.1177/1545968307300698

Malekzadeh, N. (2021). A comprehensive review of amyotrophic lateral sclerosis including: prevalence, pathogenesis, biomarkers diagnosis, and current treatment options. *Rev. Clin. Med.* 8, 180–184. doi: 10.22038/rcm.2022.57207.1365

Mills, K. R., and Nithi, K. A. (1997). Corticomotor threshold is reduced in early sporadic amyotrophic lateral sclerosis. *Muscle Nerve* 20, 1137–1141. doi: 10.1002/(SICI)1097-4598(199709)20:9<1137::AID-MUS7>3.0.CO;2-9

Moon, J., Orlandi, S., and Chau, T. (2022). A comparison and classification of oscillatory characteristics in speech perception and covert speech. *Brain Res.* 1781:147778. doi: 10.1016/j.brainres.2022.147778

Nzwalo, H., de Abreu, D., Swash, M., Pinto, S., and de Carvalho, M. (2014). Delayed diagnosis in ALS: the problem continues. *J. Neurol. Sci.* 343, 173–175. doi: 10.1016/j.jns.2014.06.003

O'Neill, G. C., Barratt, E. L., Hunt, B. A. E., Tewarie, P. K., and Brookes, M. J. (2015). Measuring electrophysiological connectivity by power envelope correlation: a technical review on MEG methods. *Phys. Med. Biol.* 60, R271–R295. doi: 10.1088/0031-9155/60/21/R271

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011, 1–9. doi: 10.1155/2011/156869

Proudfoot, M., Bede, P., and Turner, M. R. (2019). Imaging cerebral activity in amyotrophic lateral sclerosis. *Front. Neurol.* 9:1148. doi: 10.3389/fneur.2018.01148

Proudfoot, M., Colclough, G. L., Quinn, A., Wuu, J., Talbot, K., Benatar, M., et al. (2018). Increased cerebral functional connectivity in ALS: a resting-state magnetoencephalography study. *Neurology* 90, e1418–e1424. doi: 10.1212/WNL.0000000000005333

Proudfoot, M., Rohenkohl, G., Quinn, A., Colclough, G. L., Wuu, J., Talbot, K., et al. (2017). Altered cortical beta-band oscillations reflect motor system degeneration in amyotrophic lateral sclerosis. *Hum. Brain Mapp.* 38, 237–254. doi: 10.1002/hbm.23357

Ravits, J., Paul, P., and Jorg, C. (2007). Focality of upper and lower motor neuron degeneration at the clinical onset of ALS. *Neurology* 68, 1571–1575. doi: 10.1212/01.wnl.0000260965.20021.47

Riddle, J., McFerren, A., and Frohlich, F. (2021). Causal role of cross-frequency coupling in distinct components of cognitive control. *Prog. Neurobiol.* 202:102033. doi: 10.1016/j.pneurobio.2021.102033

Schindler, K., Leung, H., Elger, C. E., and Lehnertz, K. (2007). Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial EEG. *Brain* 130, 65–77. doi: 10.1093/brain/awl304

Shibuya, K., Park, S. B., Geevasinga, N., Menon, P., Howells, J., Simon, N. G., et al. (2016). Motor cortical function determines prognosis in sporadic ALS. *Neurology* 87, 513–520. doi: 10.1212/WNL.0000000000002912

Sorrentino, P., Rucco, R., Jacini, F., Trojsi, F., Lardone, A., Baselice, F., et al. (2018). Brain functional networks become more connected as amyotrophic lateral sclerosis progresses: a source level magnetoencephalographic study. *NeuroImage: Clinical* 20, 564–571. doi: 10.1016/j.nicl.2018.08.001

Stegmann, G. M., Hahn, S., Liss, J., Shefner, J., Rutkove, S., Shelton, K., et al. (2020). Early detection and tracking of bulbar changes in ALS via frequent and remote speech analysis. *NPJ Digit. Med.* 3:132. doi: 10.1038/s41746-020-00335-x

Teplansky, K. J., Tsang, B. Y., and Wang, J. (2019). Tongue and lip motion patterns in voiced, whispered, and silent vowel production. In *Proc. International Congress of Phonetic Sciences* (pp. 1–5).

van Es, M. A., Hardiman, O., Chio, A., al-Chalabi, A., Pasterkamp, R. J., Veldink, J. H., et al. (2017). Amyotrophic lateral sclerosis. *Lancet* 390, 2084–2098. doi: 10.1016/S0140-6736(17)31287-4

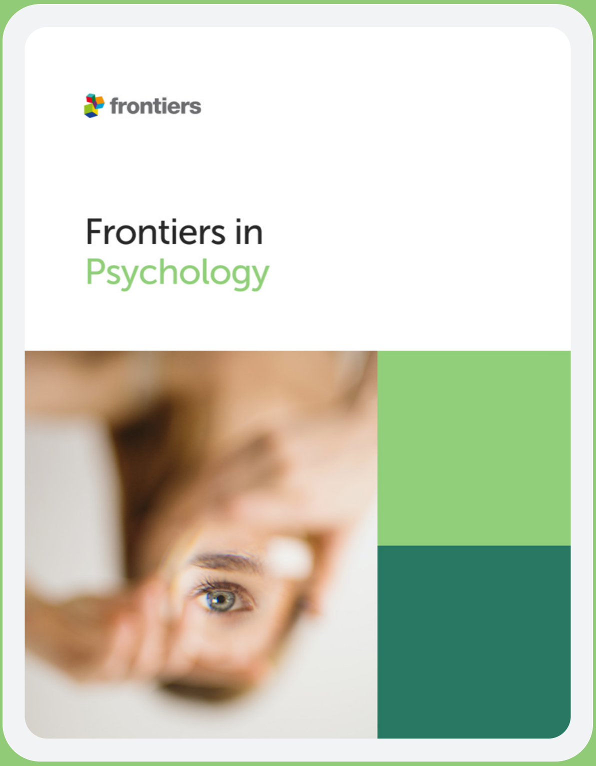
Verstraete, E., van den Heuvel, M. P., Veldink, J. H., Blanken, N., Mandl, R. C., Hulshoff Pol, H. E., et al. (2010). Motor network degeneration in amyotrophic lateral sclerosis: a structural and functional connectivity study. *PLoS One* 5:e13664. doi: 10.1371/journal.pone.0013664

Vieira, H., Costa, N., Sousa, T., Reis, S., and Coelho, L. (2019). Voice-based classification of amyotrophic lateral sclerosis: where are we and where are we going? A systematic review. *Neurodegener. Dis.* 19, 163–170. doi: 10.1159/000506259

Xu, T., Stephane, M., and Parhi, K. K. (2013) 'Classification of single-trial MEG during sentence processing for automated schizophrenia screening'. In 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE, pp. 363–366.

Yang, T., Ang, K., Phua, K. S., Yu, J., Toh, V., Ng, W., et al. (2018) 'EEG channel selection based on correlation coefficient for motor imagery classification: a study on healthy subjects and ALS patient'. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, pp. 1996–1999.

Yorkston, K. M., Strand, E. A., and Kennedy, M. R. (1996). Comprehensibility of dysarthric speech: Implications for assessment and treatment planning. *Am J Speech Lang Pathol*, 5, 55–66.



Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact