LATENT REPRESENTATION ENCODING AND MULTIMODAL BIOMARKERS FOR POST-STROKE SPEECH ASSESSMENT

Giulia Sanguedolce
Department of Computing
Imperial College London, UK
g.sanguedolce@imperial.ac.uk

Dragos C. Gruia
Department of Brain Sciences
Imperial College London, UK
d.gruia@imperial.ac.uk

Fatemeh Geranmayeh
Department of Brain Sciences
Imperial College London, UK
Imperial College Healthcare NHS Trust, UK
f.geranmayeh@imperial.ac.uk

Patrick A. Naylor
Department of Electrical and
Electronic Engineering
Imperial College London, UK
p.naylor@imperial.ac.uk

ABSTRACT

Post-stroke language impairments affect speech and language production, leading to lexical, semantic, syntactic, and articulatory-prosodic deficits. These disruptions extend from impaired cognitive-motor planning to execution, manifesting as altered vocal fold dynamics that compromise speech fluency and intelligibility. The high-dimensional and multimodal nature of these impairments poses significant challenges to traditional assessment methods, necessitating automated solutions that can capture the heterogeneity of disfluencies. We present a multimodal framework that integrates foundation model embeddings with clinically-guided features for speech assessment. Leveraging SONIVA, our purpose-built database of $\approx$ 600 post-stroke patients, we fine-tune Whisper to extract encoder embeddings that capture pathological speech characteristics. These representations are integrated with linguistic complexity metrics, physiological glottal parameters, and acoustic features through neural networks. Our model achieves 92.4% classification accuracy in stroke detection, outperforming feature-based methods, with SHAP analysis validating the modality-specific importance. We further demonstrate real-world clinical utility through severity prediction on Comprehensive Aphasia Test (CAT) scores, achieving an N-RMSE of 0.1299. This framework establishes a clinically relevant approach for integrating speech representations with domain-specific biomarkers to potentially support diagnosis, severity tracking, and precision rehabilitation strategies.

1 INTRODUCTION

The speech signal conveys information across multiple levels, from the physical vibrations of the vocal folds to the lexical and semantic structures of language. This multidimensional nature makes speech a robust biomarker for the evaluation of physiological and cognitive health, particularly in neurological conditions (Ramanarayanan et al., 2022). However, speech analysis is inherently complex due to inter-individual variability in vocal tract anatomy, speaking patterns, and disease-specific manifestations (Stefaniak et al., 2022; Olafson et al, 2024). Effective speech analysis would require the integration of acoustic features (e.g., fundamental frequency, jitter, shimmer, formant trajectories) and linguistic measures (e.g., lexical diversity, syntactic complexity, semantic coherence). Traditional diagnostic tools, such as clinician-rated scales or standardised language tests (e.g., Boston Diagnostic Aphasia Examination; Roth 2011), though valuable, are limited by their reliance on subjective interpretation, lack of granularity and inability to capture subtle, real-time changes in speech. Additionally, these methods are often costly and resource-intensive, requiring specialised training and significant time investment, which restricts their scalability and utility in large-scale or remote clinical settings (Mahmoud et al., 2023; Le et al., 2018; Palmer & Enderby, 2012). These limitations

underscore the need for objective, automated, and data-driven speech analysis tools that can provide quantifiable, reproducible, and clinically actionable insights.

Recently, large-scale foundation models have emerged as a promising approach for medical analysis, particularly following the introduction of BioBERT (Lee et al., 2020) and ClinicalBERT (Alsentzer et al., 2019) in early 2019, both of which built upon BERT (Devlin, 2018). The advancements of these Large Language Models (LLMs) have been instrumental in enhancing foundation models, particularly in their capacity to process and generate domain-specific, context-rich language—an essential factor for accurate language analysis in healthcare settings. More recently, multimodal LLMs such as GPT-4V (Achiam et al., 2023) have demonstrated potential in supporting clinical decision-making and management, though challenges in regulation and validation remain (Qiu et al., 2024). Similarly, Me-LLaMA (Xie et al., 2024), a newly developed medical LLM family, has exhibited superior performance across general and domain-specific medical tasks compared to existing open-source medical LLMs. Building on these advancements, recent research has explored the application of Whisper (Radford et al., 2022) specifically on speech disorders, where its pre-trained representations have demonstrated strong generalisability across diverse clinical contexts (Leung et al., 2024; Jiang et al., 2024; Li & Zhang, 2024; Sanguedolce et al., 2023; 2024a; Lee et al., 2024; Rathod et al., 2023; Best et al., 2024). Unlike traditional machine learning methods that rely on task-specific training and extensive labeled datasets, foundation models leverage pre-training on diverse speech corpora to learn robust representations that generalize well in low-resource scenarios. This is particularly valuable for medical applications where labeled data is scarce and speech variability is high due to disease severity and individual differences. However, while these models provide a strong foundation, they still require domain-specific fine-tuning to capture the nuanced characteristics of post-stroke speech disorders. The primary challenge remains the scarcity of large, clinically annotated datasets suitable for fine-tuning, limiting generalization across diverse patient populations and clinical settings.

To address this, we introduce SONIVA — detailed below — a comprehensive in-house post-stroke speech dataset comprising approximately ≈ 1000 patients, developed over several years through close collaboration between multiple hospitals and our research facilities. This corpus, specifically designed to capture the inherent variability and complexity of post-stroke speech patterns, represents a significant step toward bridging the gap between general-purpose foundation models and the demands of clinical applications in speech pathology. Our approach leverages OpenAI Whisper (Radford et al., 2022), a state-of-the-art speech recognition model, to analyse pathological speech. Whisper was chosen since its architecture is particularly well-suited for this task, having demonstrated a near-human-level accuracy in low-resource settings and robust performance across diverse speaking conditions in healthy speech (Radford et al., 2022). The model's effectiveness stems from extensive pre-training on $680\,000$ hours of speech data encompassing varied acoustic conditions, speakers, and languages. For clinical deployment, we implement a secure fine-tuning pipeline that addresses the unique challenges of medical data handling. All model adaptations are performed within an isolated, on-site computing environment, ensuring compliance with privacy regulations while minimising external network dependencies. Such approach ensures sensitive patient data throughout the fine-tuning process. After adaptation, we extract embeddings from the model's weighted encoder, capturing high-level representations of pathological speech patterns. These learned representations are then integrated into our multimodal classification framework alongside clinically relevant features - including glottal, linguistic, and acoustic parameters - to enhance diagnostic precision while maintaining interpretability. This combination leverages both the strengths of foundation models and the domain expertise encoded in more established clinical metrics.

To validate our framework's potential as an automated clinical assessment tool, we evaluate its ability to replicate expert clinical judgment through regression analysis. Specifically, we predict the scores from Comprehensive Aphasia Test (CAT; Swinburn et al. 2004), a standardised metric traditionally assigned by speech therapists through time-intensive manual assessment. This regression task represents a step towards automated clinical decision support, as accurately predicting CAT scores would enable rapid, objective, and consistent evaluation of speech impairment severity. This capability, combined with our multimodal classification approach, provides a foundation for healthcare AI systems that could support the assessment and management of diverse speech and communication disorders in clinical settings.

2 METHODS

2.1 SONIVA: POST-STROKE SPEECH DATABASE

The database used in this study, SONIVA, is an in-house comprehensive corpus of post-stroke speech, developed for clinical and scientific research aimed at improving Automatic Speech Recognition (ASR) systems for disordered speech. It includes speech recordings from approximately 6000 healthy controls and 1000 individuals with a history of stroke, collected as part of two longitudinal studies, the Imperial Comprehensive Cognitive Assessment in Cerebrovascular Disease (IC3) study (Gruia et al., 2024; 2023) and the Predicting Language Outcome and Recovery after Stroke (PLORAS) (Seghier & et al., 2016). Both studies were approved by the UK's Health Research Authority[1] and all participants gave informed consent prior to data collection. The speech recordings include picture description tasks based on standard clinical assessments of the Comprehensive Aphasia Test (CAT) (Swinburn et al., 2004) and a beach scene picture stimulus we designed. Additionally, the dataset features detailed orthographic English transcriptions performed by trained speech pathologists, as well as phonetic transcriptions using the International Phonetic Alphabet. The labeled dataset used for this study consists of 794 audio recordings from 578 unique individuals, some of whom participated in multiple recording sessions to capture longitudinal recovery and individual speech pattern variations. In total, the dataset comprises approximately 15 hours of speech data. The dataset is predominantly male (70.79%), with an average age of 61.96 years at the time of testing, while female speakers had an average age of 58.52 years. This gender imbalance reflects global stroke incidence patterns especially in younger age ranges, in which males have higher occurrence rates (Appelros et al., 2009).

2.2 DATA PRE-PROCESSING

Before fine-tuning Whisper, data preparation was conducted to support the extraction of embeddings and linguistic features. To ensure consistency, trained speech therapists handled the transcription process, resulting in an inter-rater reliability of 73%. The transcriptions followed the standardised Codes for the Human Analysis of Transcripts (CHAT; MacWhinney 2014) and were further processed using the Computerised Language ANalysis software (CLAN;Conti-Ramsden 1996). A pre-processing step was applied to remove special symbols used for annotating linguistic errors, including those indicating semantic, phonological, and dysfluency-related issues. Additionally, when neologisms or vocalisations appeared, transcribers provided phonetic representations, which were later converted into Latin-alphabet phonemes while preserving their original sequence (Perez et al., 2020).

The transcriptions also included annotations for false starts, filler words (e.g., *er, erm*), and other interjections commonly found in dysfluent speech. Since filler words exhibit spelling variations between American and British English, which could affect Word Error Rate (WER) in automatic processing, these were normalised to match the conventions used in the Whisper training dataset. To align transcriptions with corresponding audio files, utterance segmentation was manually marked by expert transcribers following our established previous methodologies (Sanguedolce et al., 2023; 2024a;b). This ensured that each segment corresponded to a complete sentence rather than arbitrary fixed-length chunks, reducing the risk of overfitting during model training. Instances of assessor speech present in the recordings were also transcribed but subsequently excluded from the training dataset to ensure the model learns only patient speech.

Since Whisper by default does not process audio recordings exceeding 30 seconds, longer files were segmented into smaller units while maintaining alignment with their transcriptions. Additionally, files shorter than 3 seconds posed computational challenges in Fourier transform calculations for spectrogram generation (Torre & Romero, 2021). Instead of discarding these short recordings, they were merged with adjacent utterances from the same speaker to preserve valuable data and optimise training efficiency. All recordings were converted to WAV format, resampled to 16 kHz, encoded with 16-bit resolution, and downmixed to mono. After pre-processing, the final curated dataset contained approximately 13 hours of speech data. The data processing pipeline utilised several specialised tools, including SpeechBrain, Pydub (Robert et al., 2018), FFmpeg (Tomar, 2006), and SoX (Barras, 2012).

[1] IRAS numbers: 299333 and 133939 for IC3 and PLORAS respectively.

2.3 FINE-TUNING

Our fine-tuning approach builds on Whisper's encoder-decoder transformer architecture. The OpenAI model processes acoustic input through an encoder that converts 80-channel log-Mel spectrograms into rich representations via two convolutional layers and sinusoidal positional encoding. These initial features are then refined through transformer blocks that capture long-range dependencies, while the decoder architecture mirrors this structure using learned positional embeddings (Radford et al., 2022). To adapt this architecture for pathological speech recognition, we implemented a systematic fine-tuning protocol. The dataset was divided 70% (551 minutes) for training, 18% (141 minutes) for validation, and 12% (94 minutes) for testing. To ensure unbiased evaluation, the test set comprised recordings exclusively from speakers not represented in the training or validation sets. Whisper's medium-sized model has been selected and all trainable parameters were tuned, allowing both acoustic and linguistic layers to adapt uniformly to pathological speech patterns. This full model adaptation approach, with no frozen layers, maximised the model's capacity to learn disorder-specific features. Training was conducted with a batch size of 16 per device, employing gradient accumulation for memory efficiency. The optimisation process utilised the AdamW algorithm with cross-entropy loss minimisation, following a cosine learning rate schedule initialised at 1×10^{-5} and incorporating a 1000-step warm-up phase. Assessment was performed at 1000 steps intervals, with the checkpoint corresponding to the lowest WER on the validation set retained as the best-performing version. Training continued up to a maximum of 6000 steps. To improve computational efficiency, mixed-precision arithmetic (fp16) and gradient checkpointing were implemented, reducing memory overhead. The fine-tuning process required approximately 8 hours and was executed using PyTorch (Paszke & Gross, 2019) alongside the Hugging Face Transformers library (Wolf & Debut, 2019), utilizing a single NVIDIA RTX 6000 GPU. The final evaluation was conducted using the WER, quantifying the discrepancy between Whisper's transcriptions and human-annotated ground truth. To establish a performance benchmark, we first measured WER using the pre-trained Whisper model before comparing it against the fine-tuned variant. As shown in our latest work (Sanguedolce et al., 2024a;b) fine-tuning on the dataset significantly improved performance, reducing WER from the baseline 39.60% to 21.51% on the validation set and from 43.62% to 21.93% on the test set, showing generalisation capabilities also on external datasets like AphasiaBank (MacWhinney et al., 2011) or DementiaBank (Lanzi et al., 2023). Such fine-tuned model is then used in this work to extract embeddings and linguistic features.

2.4 FEATURE EXTRACTION

To ensure unbiased evaluation, Whisper-derived embeddings and linguistic metrics were extracted only from the unseen test set used during fine-tuning (94 minutes). This also prevents data leakage, as using the same speech samples from training could artificially inflate performance. Since embeddings encode speech characteristics learned during fine-tuning, reusing them from the training set would exploit prior exposure. Likewise, linguistic features from Whisper's ASR output might reflect learned transcription patterns rather than actual speech impairments. To maintain consistency, we applied the same test-set-only extraction across all modalities. The test set was complemented with age-matched healthy speakers ($\mu = 61.51$ years, $\sigma = 10.55$ years) with no history of neurological impairments, comprising 90 minutes of recordings.

Embeddings To extract embeddings we leverage the fine-tuned Whisper encoder, utilizing the final hidden state of its last transformer layer as a compact and informative representation of each input utterance. The extracted embeddings consist of a 1024-dimensional latent feature vector, encapsulating both acoustic and linguistic attributes of speech. In the earlier layers, the encoder captures low-level acoustic features (e.g., pitch, formants, energy), while deeper layers encode increasingly abstract linguistic structures, including phonetic, lexical, and semantic information. The HuggingFace implementation of Whisper was employed for this process (Wolf & Debut, 2019), ensuring reproducibility and consistency across feature extraction. These embeddings serve as a high-dimensional latent representation, providing a data-driven alternative to already established features. By preserving key speech characteristics in a structured feature space integrating acoustic and linguistic cues within a single representation, these embeddings constitute a core modality within our multimodal classification framework.

Linguistic Features Linguistic features were extracted from Whisper's ASR output and validated against manual transcriptions to ensure accuracy and reliability, as established in a previous work.

Using the NLTK library (Bird et al., 2009) for tokenization, part-of-speech tagging, and stopword identification, the extracted 10 features capture lexical diversity, grammatical complexity, disfluencies, filler word usage, part-of-speech patterns and content richness. Lexical diversity was assessed using metrics such as total word count, unique words, type-token ratio (TTR), lexical density, and noun-to-verb ratio, capturing vocabulary richness and accessibility. Grammatical complexity was measured through the ratio of function words to total words, indicating syntactic sophistication and cohesion, while part-of-speech (POS) transitions helped identify syntactic patterns and irregularities. Disfluencies were quantified by analysing filler words and their frequency, reflecting speech fluency. Content richness, evaluated through the diversity of content words (e.g., nouns, verbs, adjectives), provided insights into meaningful communication.

Glottal Features Glottal feature extraction provides a time-domain signal independent of vocal tract resonances, allowing precise assessment of laryngeal function. The process involved two steps: (1) isolating voiced segments using the PEFAC algorithm in MATLAB (Gonzalez & Brookes, 2014), ensuring that only phonation-related portions were analysed, and (2) detecting glottal closure (GCI) and opening instants (GOI) using the YAGA algorithm (Thomas et al., 2011). Once GCI and GOI were identified, a set of 80 well established (Kadiri & Alku, 2019; Narendra & Alku, 2018; 2019; 2020; Corcoran et al., 2019) glottal parameters were extracted, categorised into time-domain and frequency-domain features, through their summary statistics (mean, standard deviation, minimum, maximum, kurtosis, median, skewness, and range).

The Time-Domain features quantify vocal fold motion over time. The Opening Quotient (OQ) and Closing Quotient (CQ) measure the proportion of the glottal cycle in which the glottis remains open or closed, respectively. Speed Quotient (SQ) captures the asymmetry of glottal pulses with deviations indicating irregular vocal fold vibrations, while Amplitude Quotient (AQ) and Normalised Amplitude Quotient (NAQ; normalised by the fundamental frequency) characterise glottal closure intensity, aiding in the detection of breathy or pressed phonation. Frequency-Domain features assess spectral properties of glottal vibrations. Harmonic Richness Factor (HRF) evaluates harmonic energy relative to the fundamental frequency, reflecting the richness of voiced sounds. H1H2 and H2H4 quantify harmonic amplitude differences, offering insight into vocal fold tension and phonation type. Parabolic Spectrum Parameter (PSP) measures how closely the power spectrum around each GCI resembles a parabolic shape, providing insights into spectral energy distribution. Peak Slope (PS) assesses the sharpness of glottal closure, with steeper slopes indicating more abrupt vocal fold contact, often linked to voice disorders.

Acoustic Features Acoustic parameters were extracted using the openSMILE Python library (v. 3.0.1), employing the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPSv02; Eyben et al. 2015). This standardised feature set comprises 88 parameters, selected for their relevance in clinical and paralinguistic speech analysis already widely clinically proven for assessing pathological speech (Shahin et al., 2019; Barche et al., 2020; Liu et al., 2022; Mawalim et al., 2023; Kumar et al., 2024). The acoustic feature set encompasses prosodic measures (e.g., F_0 stability, jitter, shimmer) to assess pitch and phonatory control, spectral features (e.g., MFCCs, flux) for vocal resonance and articulatory precision, and energy-related metrics (e.g., loudness, harmonics-to-noise ratio, pauses) to evaluate fluency and rhythmic disturbances.

2.5 Multimodal Assessment and Feature Attribution

To systematically classify speech patterns associated with neurological impairments, we implement a multimodal classification neural network integrating acoustic, linguistic, and glottal biomarkers alongside learned fine-tuned Whisper embeddings. The proposed neural network is a feedforward model comprising three layers: (1) a fully connected layer with 64 units, Layer Normalization, LeakyReLU activation, and Dropout (0.4); (2) a second fully connected layer with 32 units, Batch Normalization, LeakyReLU activation, and Dropout (0.4); and (3) a final output layer with a single neuron and Sigmoid activation for binary classification between patient and healthy speech samples. For the loss, we incorporated a weighted variant of the Binary Cross-Entropy (BCE) loss, integrating Focal Loss and L1 regularization as follows:

$$\mathcal{L} = \alpha_t (1 - p_t)^\gamma \mathcal{L}_{\text{BCE}} + \lambda \|\theta\|_1 \tag{1}$$

where p_t represents the predicted probability for the true class, α_t is a dynamically computed weighting factor that compensates for class imbalance, and γ controls the focusing mechanism, reducing the loss contribution of easy-to-classify samples and emphasizing harder ones. The term $(1 - p_t)^\gamma$ adjusts the impact of each sample based on prediction confidence, effectively prioritizing difficult cases. To further regularize the model, we introduce an L1 norm penalty $\|\theta\|_1$, which promotes sparsity in the learned parameters, preventing overfitting and improving generalization. Data is trained using 5-fold stratified group cross-validation on the training data before defined, ensuring that speaker-level dependencies are maintained, followed by feature standardisation via `StandardScaler`. Training employs also here the AdamW optimiser (lr = 1×10^{-5}, weight decay = 0.01), gradient clipping ($\|\nabla\theta\| \leq 1.0$), and a `ReduceLROnPlateau` scheduler that adjusts the learning rate dynamically based on validation loss. Early stopping (patience $= 12$) is used to prevent overfitting. Performance is measured via accuracy, precision, recall, and F1-score. Next, SHapley Additive exPlanations (SHAP) have been employed to assess the contribution of different speech-derived features to patient classification. KernelSHAP was used to estimate feature attributions by measuring the impact of perturbing input variables on the model's predictions. SHAP values were computed across four feature modalities—acoustic, linguistic, glottal, and embeddings—to determine their relative importance on the best fold of the full multimodal model. We then calculated the mean absolute SHAP values for each modality, providing a quantitative measure of their influence on classification outcomes.

To assess the relationship between speech-derived biomarkers and the severity of impairment, we performed a regression analysis on the patient cohort, using the Comprehensive Aphasia Test (CAT) score as a measure of speech dysfunction tested by clinicians in hospitals. We trained a severity regression model, namely a feedforward neural network consisting of a fully connected layer with 32 units and LeakyReLU activation, followed by an output layer for continuous score prediction. The model was optimised using Huber loss with $\delta = 2.0$, which provides robustness to outliers while preserving sensitivity to small deviations. Training employed the AdamW optimiser with a learning rate of 0.02 and weight decay of 0.01, along with a `ReduceLROnPlateau` scheduler that adjusted the learning rate dynamically based on validation performance, with a patience of 11 epochs and a minimum learning rate of 5×10^{-5}. Performance was assessed using range normalised root mean squared error (N-RMSE) to account for variations in score distribution, as well as absolute error distribution. The model was trained via 5-fold cross-validation and the average over folds is reported. By creating a severity estimation through the different feature modalities, this analysis complements the classification framework in order to provide a more fine-grained assessment of speech deficits.

3 RESULTS

Table 1 show the results of the classification of individual and combined modalities. The multimodal model that combined all features (including embeddings) achieved the highest performance, with an accuracy of 92.4% and an F1-score of 0.924. Removing embeddings led to a drop in accuracy (88.6%) and F1-score (0.885), highlighting the importance of learned representations in improving classification accuracy. Among single modalities, acoustic features proved the most discriminative, achieving an accuracy of 90.4% and an F1-score of 0.903, followed by glottal modality with 84.6% of accuracy and a F1 score of 0.846. The embeddings-based classifier showed comparable results to glottal features (accuracy: 80.5%, F1: 0.803), indicating that even alone learned speech representations are able to decently capture key impairments. On the other hand, the linguistic modality exhibited the most limited performance (accuracy: 66.3%, F1: 0.666), with precision (0.662) and recall (0.657) suggesting a low power in distinguishing between post-stroke and healthy speech. Following classification, the bar chart (Fig. 1) illustrates the relative importance of each feature modality in the Multimodal with embeddings model, as measured by the mean absolute SHAP values. The plot reveals that embeddings have the highest mean absolute SHAP values, indicating they play a crucial role in model predictions. This aligns with their contribution to the performance boost observed in the classification setup, highlighting their ability to encode complex speech patterns and pathologies effectively. Following embeddings, acoustic features exhibit the second-highest SHAP values, underscoring their critical role in capturing temporal and spectral characteristics that are highly indicative of neurological impairments.

Regression analysis of CAT scores demonstrated predictive errors across modalities ranging from 12.99% to 20.23% relative to the CAT points range, showing high overall performance in severity

Table 1: Results for different feature modalities and multimodal combinations with (w/ Emb.) and without (w/o Emb.) embeddings, for classifying controls and patients (Classification) and predicting Comprehensive Aphasia Test (CAT) total scores (Regression).

Modality	Classification				Regression
	Accuracy	F1-score	Precision	Recall	N-RMSE
Embeddings	0.805	0.803	0.804	0.802	0.2023
Glottal	0.846	0.846	0.846	0.844	0.1883
Linguistic	0.663	0.666	0.662	0.657	0.1608
Acoustic	0.904	0.903	0.902	0.902	0.1540
Multimodal w/o Emb.	0.886	0.885	0.882	0.883	**0.1299**
Multimodal w/ Emb.	**0.925**	**0.925**	**0.923**	**0.923**	0.1801

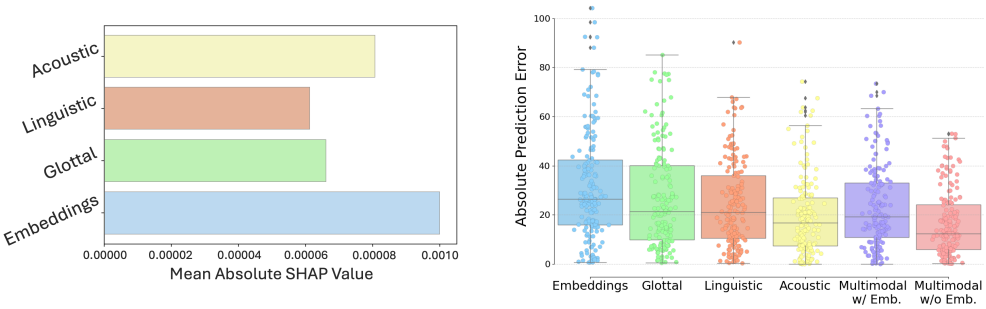

Figure 1: (Left) Classification task SHAP relative importance of each feature modality measured by mean absolute SHAP values. (Right) Regression task boxplot showing the distribution of absolute prediction errors across the modalities and multimodal combinations with (w/ Emb.) and without (w/o Emb.) embeddings.

prediction (Tab. 1). The box plot in Fig. 1 illustrates the distribution of absolute prediction errors across modalities, with individual models emonstrating greater variability in errors. The multimodal approach without embeddings achieved the best performance (Tab. 1) with the lowest N-RMSE (0.1299) and a compact error distribution, suggesting that established clinical features remain crucial for precise severity assessment. Among individual modalities, acoustic features exhibited the narrowest error range (N-RMSE: 0.1540), reflecting their reliability, followed by linguistic features (N-RMSE: 0.1608). Interestingly, while embeddings significantly enhanced classification performance, their integration in the regression task led to slightly reduced accuracy (N-RMSE: 0.1801 vs 0.1299 without embeddings - $t = 4.76,\ p = 0.008$). This underscores the importance of tailoring feature selection to the specific clinical objective, whether optimizing for diagnostic accuracy or enhancing severity estimation.

4 Discussion

Our analysis demonstrates the potential of foundation model-derived representations in conjunction with clinical metrics in post-stroke speech assessment. By combining embeddings that capture high-level representations, acoustic features that represent fine-grained spectral details, and glottal features that contribute unique motor-related insights, we achieved high classification accuracy. The importance of embeddings in classification is in keeping with recent work (Syed et al., 2020; Venugopalan et al., 2021; Bartelds et al., 2022; Neumann et al., 2024), who reported that learned representations can capture subtle speech biomarkers overlooked by traditional features. Interestingly, while embeddings enhanced classification performance, their reduced accuracy in severity prediction indicates that foundation models might excel at detecting pathological patterns but require complementary clinical features for precise severity assessment. This suggests that, though effective for capturing discriminative patterns in pathology detection, embeddings may struggle to preserve the fine-grained relationships needed for severity scoring.

Compared to prior advancements in automated speech assessment, our work, leveraging SONIVA, demonstrates notable improvements in performance and methodology in classification. While Venu-

gopalan et al. (2021) achieved 82% accuracy using ASR embeddings for disordered speech classification, we advance this approach by integrating clinical features for better pathology detection. In the context of stroke detection, Ou et al. (2025) validated the superiority of multimodal approaches, reaching 82.6% accuracy and providing early evidence that integrated features outperform single modalities. Similarly, Soltau et al. (2023) achieved 83% accuracy using a Perceiver-based sequence classifier for neurological speech abnormalities. Particularly relevant to our approach, Zusag et al. (2023) leveraged Whisper with the AphasiaBank database to differentiate various types of aphasia from healthy speech, reaching an F1 score of 90.6%. Our framework's superior performance (92.4%) builds on their findings while demonstrating the additional value of integrating clinical features with foundation model representations.

To the best of our knowledge, the prediction of CAT scores from speech, as well as the use of our metrics, has not been widely explored yet, precluding direct performance comparisons with existing approaches. Nevertheless, such a multimodal analysis with clinically established features shows high predictive ability, achieving a 12% error rate across the full CAT score range (0-216 points). This result is particularly noteworthy given that the CAT assessment encompasses multiple linguistic domains including comprehension, expression, reading, naming, repetition and writing. Indeed, by predicting such a score we were able to derive predictions for this comprehensive score using solely audio recordings from the picture description task, suggesting that this focused speech sample contains rich information about overall language abilities. These results suggest potential clinical utility of these models built upon SONIVA in supporting patient monitoring decisions.

5 Future Work and Limitations

Future work will focus on evaluating the model's predictive capabilities for cognitive decline using the Montreal Cognitive Assessment (MoCA; Nasreddine et al. 2005), collected alongside CAT scores. This evaluation will provide deeper insights into the model's clinical relevance and its potential to capture broader neurological deficits beyond speech impairment. Several promising research directions emerge from our findings. Given that stroke-induced speech impairments often overlap with symptoms of other neurological conditions (e.g., Parkinson's disease, multiple sclerosis), our framework could potentially be adapted to additional disorder profiles. Indeed, a limitation of our study is that the model was trained and tested on a single dataset, but generalisability on different clinical environments is important for real-world scenarios. Expanding validation to publicly available datasets such as AphasiaBank (MacWhinney et al., 2011) or DementiaBank (Lanzi et al., 2023) would enable cross-pathology evaluation and strengthen the model's robustness. Technical improvements could also enhance model performance. We did not implement speech enhancement techniques, which could have improved a possible resilience to background noise. Furthermore, hyperparameter optimisation could refine the model architecture, potentially enhancing predictive accuracy and efficiency, alongside exploration of alternative architectures and dimensionality reduction approaches. Long-term clinical validation remains essential to ensure model outputs align with expert transcriptions and provide meaningful insights for assessment and monitoring. Our ongoing collaboration with an interdisciplinary team—including speech-language pathologists and neurologists—remains central to this process, ensuring technical advancements maintain clinical relevance.

6 Conclusion

Our work demonstrates that combining foundation model embeddings with established clinical metrics creates a powerful framework for detecting post-stroke speech impairments, with traditional clinical features that remain crucial for predicting overall language abilities across diverse linguistic domains. The strong performance of this tailored approach, using SONIVA, highlights the model's potential to transform clinical assessment — shifting away from traditional, labor-intensive methods toward an intelligent integration of multimodal data streams and foundation models. Just as BERT transformed clinical text analysis, Whisper-based models may represent a step change in speech-based diagnostics, particularly for early screening and severity assessment. By integrating computational models with clinical expertise in this targeted way, this research accelerates the adoption of automated speech analysis as a possible tool for diagnosing and managing language disorders post-stroke.

ACKNOWLEDGMENTS

The authors would like to thank A. Coghlan, O. Burton and N. Parkinson for assistance with the IC3 study, and J. Friedland, K. Stephenson, G. Gvero, Z. Zaskorska and C. Leong for labelling the speech data. We are grateful to the patients, clinical teams, and the PLORAS team, that contributed data to the SONIVA database. We also thank F. N. for technical support with the data processing and modelling. Infrastructure support was provided by the NIHR Imperial Biomedical Research Centre and the NIHR Imperial Clinical Research Facility. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. Funding support was provided to G.S. (UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare:EP/S023283/1]), F.G. (Medical Research Council P79100) and S.B. (EPSRC IAA -PSP415 and MRC IAA -PSP518).

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Peter Appelros, Birgitta Stegmayr, and Andreas Terént. Sex differences in stroke epidemiology: a systematic review. *Stroke*, 40(4):1082–1090, 2009.

Purva Barche, Krishna Gurugubelli, and Anil Kumar Vuppala. Towards automatic assessment of voice disorders: A clinical approach. In *INTERSPEECH*, pp. 2537–2541, 2020.

Benjamin Barras. Sox : Sound exchange. 01 2012.

Martijn Bartelds, Wietse de Vries, Faraz Sanal, Caitlin Richter, Mark Liberman, and Martijn Wieling. Neural representations for modeling variation in speech. *Journal of Phonetics*, 92:101137, 2022.

Paul Best, Santiago Cuervo, and Ricard Marxer. Transfer learning from Whisper for microscopic intelligibility prediction. In *Interspeech 2024*, pp. 3839–3843, 2024. doi: 10.21437/Interspeech.2024-2258.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

Gina Conti-Ramsden. CLAN (Computerized Language Analysis). *Child Language Teaching and Therapy*, 12(3):345–349, 1996.

Patrick Corcoran, Arnold Hensman, and Barry Kirkpatrick. Glottal flow analysis in parkinsonian speech. In *Proc. of the 12th Int. Joint Conf. on Biomedical Eng. Syst. and Technol. (BIOSTEC)*, pp. 116–123, 2019.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015.

Sira Gonzalez and Mike Brookes. Pefac-a pitch estimation algorithm robust to high levels of noise. *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, 22(2):518–530, 2014.

Dragos C Gruia, Valentina Giunchiglia, Aoife Coghlan, Sophie Brook, Soma Banerjee, Joseph Kwan, Peter J Hellyer, Adam Hampshire, and Fatemeh Geranmayeh. Online monitoring technology for deep phenotyping of cognitive impairment after stroke. *medRxiv*, 2024. doi: 10.1101/2024.09.06.24313173.

Dragos-Cristian Gruia, William Trender, Peter Hellyer, Soma Banerjee, Joseph Kwan, Henrik Zetterberg, Adam Hampshire, and Fatemeh Geranmayeh. Ic3 protocol: a longitudinal observational study of cognition after stroke using novel digital health technology. *BMJ open*, 13(11): e076653, 2023.

Yicong Jiang, Tianzi Wang, Xurong Xie, Juan Liu, Wei Sun, Nan Yan, Hui Chen, Lan Wang, Xunying Liu, and Feng Tian. Perceiver-prompt: Flexible speaker adaptation in Whisper for chinese disordered speech recognition. In *Interspeech 2024*, pp. 2025–2029, 2024. doi: 10.21437/Interspeech.2024-852.

Sudarsana Reddy Kadiri and Paavo Alku. Analysis and detection of pathological voice using glottal source features. *IEEE J. of Selected Topics in Signal Process.*, 14(2):367–379, 2019.

Deepak Kumar, Udit Satija, and Preetam Kumar. Pathological speech and electroglottography signals analysis using invariance scattering network. *Circuits, Systems, and Signal Processing*, pp. 1–18, 2024.

Alyssa M Lanzi, Anna K Saylor, Davida Fromm, Houjun Liu, Brian MacWhinney, and Matthew L Cohen. Dementiabank: Theoretical rationale, protocol, and illustrative analyses. *American Journal of Speech-Language Pathology*, 32(2):426–438, 2023.

Duc Le, Keli Licata, and Emily Mower Provost. Automatic quantitative analysis of spontaneous aphasic speech. *Speech Communication*, 100:1–12, 2018.

Jeehyun Lee, Yerin Choi, Tae-Jin Song, and Myoung-Wan Koo. Inappropriate pause detection in dysarthric speech using large-scale speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12486–12490. IEEE, 2024.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze. Training data augmentation for dysarthric automatic speech recognition by text-to-dysarthric-speech synthesis. *arXiv preprint arXiv:2406.08568*, 2024.

Jinpeng Li and Wei-Qiang Zhang. Whisper-based transfer learning for alzheimer disease classification: Leveraging speech segments with full transcripts as prompts. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11211–11215. IEEE, 2024.

Yuanyuan Liu, Mittapalle Kiran Reddy, Nelly Penttilä, Tiina Ihalainen, Paavo Alku, and Okko Räsänen. Automatic assessment of parkinson's disease using speech representations of phonation and articulation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31: 242–255, 2022.

Brian MacWhinney. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press, 2014.

Brian MacWhinney, Davida Fromm, Margaret Forbes, and Audrey Holland. Aphasiabank: Methods for studying discourse. *Aphasiology*, 25(11):1286–1307, 2011.

Seedahmed S Mahmoud, Raphael F Pallaud, Akshay Kumar, Serri Faisal, Yin Wang, and Qiang Fang. A comparative investigation of automatic speech recognition platforms for aphasia assessment batteries. *Sensors*, 23(2):857, 2023.

Candy Olivia Mawalim, Benita Angela Titalim, Shogo Okada, and Masashi Unoki. Non-intrusive speech intelligibility prediction using an auditory periphery model with hearing loss. *Applied Acoustics*, 214:109663, 2023.

NP Narendra and Paavo Alku. Dysarthric speech classification using glottal features computed from non-words, words and sentences. In *Interspeech*, pp. 3403–3407. Int. Speech Commun. Assoc. (ISCA), 2018.

NP Narendra and Paavo Alku. Dysarthric speech classification from coded telephone speech using glottal features. *Speech Commun.*, 110:47–55, 2019.

NP Narendra and Paavo Alku. Glottal source information for pathological voice detection. *IEEE Access*, 8:67745–67755, 2020.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

Michael Neumann, Hardik Kothare, and Vikram Ramanarayanan. Multimodal speech biomarkers for remote monitoring of als disease progression. *Computers in Biology and Medicine*, 180: 108949, 2024.

Emily R Olafson et al. Data-driven biomarkers better associate with stroke motor outcomes than theory-based biomarkers. *Brain Commun.*, 2024.

Zijun Ou, Haitao Wang, Bin Zhang, Haobang Liang, Bei Hu, Longlong Ren, Yanjuan Liu, Yuhu Zhang, Chengbo Dai, Hejun Wu, et al. Early identification of stroke through deep learning with multi-modal human speech and movement data. *Neural Regeneration Research*, 20(1):234–241, 2025.

Rebecca Palmer and Pam et al. Enderby. Computer therapy compared with usual care for people with long-standing aphasia poststroke: a pilot randomized controlled trial. *Stroke*, 43(7):1904–1911, 2012.

Adam Paszke and Sam et al. Gross. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Matthew Perez, Zakaria Aldeneh, and Emily Mower Provost. Aphasic speech recognition using a mixture of speech intelligibility experts. *arXiv preprint arXiv:2008.10788*, 2020.

Jianing Qiu, Wu Yuan, and Kyle Lam. The application of multimodal large language models in medicine. *The Lancet Regional Health–Western Pacific*, 45, 2024.

Alec Radford, Jong Wook Kim, and Tao et al. Xu. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Vikram Ramanarayanan, Adam C Lammert, Hannah P Rowe, Thomas F Quatieri, and Jordan R Green. Speech as a biomarker: opportunities, interpretability, and challenges. *Perspectives of the ASHA Special Interest Groups*, 7(1):276–283, 2022.

Siddharth Rathod, Monil Charola, Akshat Vora, Yash Jogi, and Hemant A. Patil. Whisper features for dysarthric severity-level classification. In *Interspeech 2023*, pp. 1523–1527, 2023. doi: 10. 21437/Interspeech.2023-1891.

James Robert, Marc Webbie, et al. Pydub, 2018. URL `http://pydub.com/`.

Carole Roth. *Boston Diagnostic Aphasia Examination*, pp. 428–430. Springer New York, New York, NY, 2011. ISBN 978-0-387-79948-3. doi: 10.1007/978-0-387-79948-3_868. URL `https://doi.org/10.1007/978-0-387-79948-3_868`.

Giulia Sanguedolce, Patrick A Naylor, and Fatemeh Geranmayeh. Uncovering the potential for a weakly supervised end-to-end model in recognising speech from patient with post-stroke aphasia. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pp. 182–190, 2023.

Giulia Sanguedolce, Sophie Brook, Dragos C Gruia, Patrick A Naylor, and Fatemeh Geranmayeh. When whisper listens to aphasia: Advancing robust post-stroke speech recognition. In *Interspeech*, 2024a.

Giulia Sanguedolce, Dragos-Cristian Gruia, Sophie Brook, Patrick Naylor, and Fatemeh Geranmayeh. Universal speech disorder recognition: Towards a foundation model for cross-pathology generalisation. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024b.

Mohamed L Seghier and Patel et al. The ploras database: a data repository for predicting language outcome and recovery after stroke. *Neuroimage*, 124:1208–1212, 2016.

Mostafa Shahin, Beena Ahmed, Daniel V Smith, Andreas Duenser, and Julien Epps. Automatic screening of children with speech sound disorders using paralinguistic features. In *2019 ieee 29th international workshop on machine learning for signal processing (mlsp)*, pp. 1–5. IEEE, 2019.

Hagen Soltau, Izhak Shafran, Alex Ottenwess, R Joseph Jr, Rene L Utianski, Leland R Barnard, John L Stricker, Daniela Wiepert, David T Jones, and Hugo Botha. Detecting speech abnormalities with a perceiver-based sequence classifier that leverages a universal speech model. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2023.

James D Stefaniak, Fatemeh Geranmayeh, and Matthew A Lambon Ralph. The multidimensional nature of aphasia recovery post-stroke. *Brain*, 145(4):1354–1367, 2022.

Kate Swinburn, Gillian Porter, and David Howard. Comprehensive aphasia test. *APA PsycTests*, 2004.

Zafi Sherhan Syed, Sajjad Ali Memon, and Abdul Latif Memon. Deep acoustic embeddings for identifying parkinsonian speech. *International Journal of Advanced Computer Science and Applications*, 11(10):726–734, 2020.

Mark RP Thomas, Jon Gudnason, and Patrick A Naylor. Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm. *IEEE Trans. on Audio, Speech, and Lang. Process.*, 20(1):82–91, 2011.

Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006.

Iván G Torre and Mónica et al. Romero. Improving aphasic speech recognition by using novel semi-supervised learning methods on Aphasiabank for English and Spanish. *Applied Sciences*, 11(19): 8872, 2021.

Subhashini Venugopalan, Joel Shor, Manoj Plakal, Jimmy Tobin, Katrin Tomanek, Jordan R Green, and Michael P Brenner. Comparing supervised models and learned speech representations for classifying intelligibility of disordered speech on selected phrases. *arXiv preprint arXiv:2107.03985*, 2021.

Thomas Wolf and Lysandre et al. Debut. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, et al. Me llama: Foundation large language models for medical applications. *arXiv preprint arXiv:2402.12749*, 2024.

Mario Zusag, Laurin Wagner, and Theresa Bloder. Careful Whisper - leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification. In *Interspeech 2023*, pp. 3013–3017, 2023. doi: 10.21437/Interspeech.2023-1653.