Transferable Lesion-Supervised Speech Representations for Post-Stroke Modelling

Anonymous Authors[1]

Abstract

Characterising post-stroke brain injury typically relies on structural neuroimaging, which is costly, infrastructure-dependent, and poorly suited to repeated or large-scale monitoring; this also creates a data-efficiency problem, because supervised targets in clinical cohorts are often sparsely observed across modalities, limiting the amount of paired speech and clinical data available for downstream modelling. We investigate a transfer learning approach in which representations learned through lesion inference from speech are reused for downstream structural and clinical phenotypes absent from the original training signal. These lesion-supervised representations are derived from a multi-head speech-to-lesion (S2L) model as patient-level aggregations of out-of-fold predicted lesion probabilities, stacked learner outputs, and uncertainty estimates across speech samples. These S2L representations are benchmarked against clinically interpretable speech features, Whisper embeddings, clinical covariates, and demographic baselines under nested cross-validation. S2L representations are the strongest direct predictors of focal lesion bounding-box extent ($R^2 = 0.241$, Pearson $r = 0.503$), despite spatial extent being unseen during S2L training, while high-dimensional Whisper encoder embeddings perform best directly for cognitive outcomes ($R^2 = 0.486$, $r = 0.702$) with further improvement after residual adaptation using S2L representations ($R^2 = 0.501$). These findings suggest that representation transfer can partially offset sparse clinical annotations, offering a scalable route for modelling post-stroke brain–behaviour relationships.

1. Introduction

Post-stroke speech and language impairment affect roughly 40% of stroke survivors, strongly contributing to reduced functional outcome and increased long-term care needs (Lazar & Boehme, 2017; Flowers et al., 2013). Patterns of impairment may vary depending on factors such as lesion size, hemispheric lateralization, anatomical location, and subsequent network-level reorganization (Geranmayeh et al., 2016; Butler et al., 2014; Forkel et al., 2014). These same factors also predict long-term cognitive and language outcomes following stroke (Bowren Jr et al., 2022; Ernst et al., 2018; Salvalaggio et al., 2020; Weigel et al., 2025). At the same time, the consequent language impairment presents highly heterogeneous symptoms, varying substantially between patients and within the same patient over time (Stefaniak et al., 2022). This heterogeneity makes accurate outcome modelling difficult and limits the usefulness of rigid diagnostic categories alone.

In current practice, such characterization still relies heavily on structural neuroimaging and clinician-administered standardized face-to-face assessments. These tools are clinically valuable, but MRI-based stroke workflows carry substantial cost and infrastructure requirements (Wardlaw et al., 2014), while standardized assessment batteries face implementation barriers in routine care, including administration time and variable adherence (Bland et al., 2013). Speech is therefore an attractive complementary modality. It is inexpensive, can be obtained remotely, and directly reflects neurological impairment (Bowden et al., 2023; Yang et al., 2026; Chen et al., 2024; Mirheidari et al., 2024). Structural brain imaging studies indicate that language performance covaries with brain structure: fluency and sentence generation relate to cortical morphometry (Roehrich-Gascon et al., 2015), while Western Aphasia Battery (Kertesz, 2007) and Philadelphia Naming Test (Roach et al., 1996) scores correlate with lesion location and white matter damage (den Ouden et al., 2019).

Although promising, modelling lesion characteristics directly from speech remains largely unexplored. In the existing literature cited above, speech measures are mainly presented as behavioural consequences of neural damage rather than predictors for anatomical inference, or analyses have only focused on statistical associations and not direct prediction. Two fundamental reasons account for this gap. First, the mapping between symptoms and brain damage is not one-to-one: lesion patterns are shaped by vascular structure, and behavioural deficits emerge from

distributed network-level disruption. Second, and as a direct consequence of this complexity, large-scale datasets jointly containing richly annotated speech and matched structural and clinical measures remain scarce, and even within those that exist, annotations are frequently incomplete or inconsistently collected across clinical populations, compounding the challenge for model development and validation (Li et al., 2026).

The present work addresses this gap directly. We release the dataset used in this work [XXX][1]. Then, building on our prior speech-to-lesion model (S2L)[2], which showed that post-stroke speech contains lesion-relevant information, we investigate whether S2L representations generalise to downstream targets absent from the original training signal, benchmarking against clinically interpretable features, raw speech embeddings from Whisper (Radford et al., 2023), clinical, and demographic baselines under a residual adaptation framework (Section 4).

2. Data collection

The speech corpus comes from an ongoing multimodal study of two post-stroke cohorts with speech, language, cognitive, and imaging measures [XXX], containing recordings from 639 stroke survivors ($\mu = 60.78$, $\sigma = 13.06$ years, 68% male) and 104 age-matched healthy controls ($\mu = 59.65$, $\sigma = 11.49$ years, 46% male). Both cohorts use a picture-description task elicited from in-house stimuli and the Comprehensive Aphasia Test (CAT; Swinburn et al. 2004), recorded at 16 kHz/16-bit. Verbatim transcripts followed CHAT guidelines (MacWhinney, 2014), were produced by speech therapists and trained annotators (73% inter-rater reliability), and were processed with CLAN (Conti-Ramsden, 1996). The S2L model was trained to predict shared lesion targets from speech across both cohorts, including lobe-level involvement and lesion volume. Richer lesion geometry and cognitive measures were available only for one cohort, which we use here to test transfer to targets outside the original S2L supervision (110 stroke survivors, 68.2% male; age $\mu = 62.24$, $\sigma = 13.81$ years).

The downstream targets span cognitive and structural lesion characteristics. Cognitive outcomes included MoCA (Nasreddine et al., 2005) and a multi-task composite summarising orientation, attention, memory, calculation, comprehension, semantic judgement, visuospatial processing, and daily living activities [XXX]. Individual task scores were used as candidate predictors where allowed, and the composite was treated as a separate target, computed as the mean of z-scored task measures. Structural phenotypes capture focal lesion geometry from MRI data processed in standard space using `FSL` (Jenkinson et al., 2012) and `SPM` (Friston et al., 1994), with masks manually delineated and verified by a neurologist, then represented as binary voxel-wise maps in Montreal Neurological Institute (MNI) space (Mazziotta et al., 2001). Low-dimensional descriptors were derived including component count, bounding-box dimensions, and anterior-posterior and superior-inferior spatial indices. All variables are summarised in Table 2 in the appendix.

[1]DOI and details released after review.

[2]Work revealed after anonymity period.

3. Speech Representations Extraction

Speech was represented at two levels: speech features extracted directly from audio recordings and their matched transcript and patient-level prediction from the S2L model. The recording-level features included interpretable raw speech measures grouped in five feature domains: glottal-source descriptors derived via `YAGA` (Thomas et al., 2011); acoustic-prosodic descriptors extracted with `openSMILE` using eGeMAPS (Eyben et al., 2015); fluency features from `pyannote` (Bredin, 2023) segmentation; linguistic features from CHAT-formatted manual transcripts (MacWhinney, 2014) computed with `spaCy` (Honnibal, 2017) and `NLTK` (Bird et al., 2009), together yielding 324 interpretable features; and 1024-dimensional Whisper-medium encoder embeddings (Radford et al., 2023) extracted via HuggingFace (Wolf & Debut, 2019). Manual timestamps from transcriptions were used to remove interviewer speech before feature extraction, then the features were aggregated to patient-level via mean and standard deviation.

The second representation family was derived from the trained S2L model, a multi-head inverse lesion predictor (Figure 1) organised around clinically interpretable tasks: stroke detection, lesion laterality, left-hemisphere lobe involvement, and lesion burden. For each patient i, we constructed a frozen S2L output representation u_i by averaging cross-fitted S2L outputs across that patient's speech samples,

$$u_i = \frac{1}{|S_i|} \sum_{j \in S_i} \phi(s_{ij}),$$

where S_i denotes the set of speech samples for patient i, s_{ij} is the speech-derived input for sample j, and $\phi(\cdot)$ concatenates final S2L head predictions, base-learner stack outputs, residual uncertainty scales, multiclass lesion-burden outputs, and prediction-interval widths. Because the S2L outputs are out-of-fold, u_i is treated as a frozen cross-fitted representation rather than an in-sample model output. Downstream predictors are trained only on u_i or comparison feature sets.

4. Prediction Tasks and Transfer Evaluation

The modelling objective was to test whether u_i transfers to downstream targets absent from the original lesion-

supervised training. Each target was evaluated as a patient-level regression problem under nested cross-validation. Candidate feature sets included: (i) raw speech representations, comprising interpretable speech descriptors and Whisper-derived embeddings; (ii) u_i constructed as described above; (iii) demographic covariates; (iv) clinical task variables; and (v) allowed multimodal concatenations of these sets.

For each target, we first selected a target-specific non-S2L baseline from a completed feature-set benchmark. This benchmark evaluated all allowed non-S2L feature sets, including raw speech representations, demographic covariates, clinical task variables, and their non-leaking multimodal combinations. For each target–set pair, candidate regressors were compared under nested cross-validation, including ridge and elastic-net regression, partial least squares, extremely randomised trees, random forests, radial-basis kernel ridge regression, histogram gradient boosting, CatBoost, and small multilayer perceptrons, with regularisation strengths, component numbers, tree regularisation, and MLP hidden size selected inside the inner loop. The selected baseline set B_τ for each target was taken from the benchmark summary and evaluated inside the residual adaptation procedure without using held-out patients for residual-model training.

Residual adaptation was evaluated only for targets where the strongest non-S2L configuration outperformed the strongest S2L-containing configuration in the direct benchmark. In these cases, the selected non-S2L model served as the first-stage predictor, and the S2L representation was tested as a residual correction. All preprocessing steps, including imputation, scaling, variance filtering, collinearity filtering, and feature selection, were learned within training folds. The residual stage used a restricted model grid to limit overfitting when modelling fold-local residuals. Candidate residual models included ridge regression with $\lambda \in \{1, 10, 100, 1000\}$, elastic net with $\lambda \in \{0.1, 1.0\}$ and l_1-ratio 0.2, partial least squares with 2, 3, or 5 components, and constrained extremely randomised trees with 250 trees, square-root feature sampling, and minimum leaf size 8.

The residual test can be written as a two-stage correction model. Let i index patients, τ_i denote the target value, u_i the S2L representation defined above, and $b_{i,\tau}$ the patient-specific feature vector from the selected non-S2L baseline set for target τ. We first estimate a baseline model m_ω,

$$\hat{\tau}_i^{\text{base}} = m_\omega(b_{i,\tau}), \tag{1}$$

and define the residual

$$\delta_i = \tau_i - \hat{\tau}_i^{\text{base}}. \tag{2}$$

A second model q_η is then trained to predict this residual from the S2L representation,

$$\hat{\delta}_i = q_\eta(u_i), \qquad \hat{\tau}_i = \hat{\tau}_i^{\text{base}} + \hat{\delta}_i. \tag{3}$$

Within each outer training fold, the baseline predictions used to define δ_i were generated out-of-fold from the training patients, so residual targets were not computed from in-sample baseline predictions. Held-out patients were used only for final evaluation. Performance was reported using R^2 and Pearson correlation, with 95% confidence intervals estimated using 2,000 bootstrap iterations over patient-level out-of-fold predictions.

5. Results

S2L-containing configurations achieved the strongest held-out performance for most targets, with the clearest and most consistent effects observed for focal lesion bounding-box extent (Table 1). For lesion bbox y, the S2L representation alone outperformed the best non-S2L block ($R^2 = 0.241$ [95% CI: 0.026, 0.394] vs. -0.020, Pearson $r = 0.503$ vs. 0.101). For lesion bbox z, the same representation again provided the best result ($R^2 = 0.175$ [95% CI: 0.033, 0.271] vs. -0.015, Pearson $r = 0.423$ vs. 0.063). For lesion bbox x, speech interpretable features plus S2L, also outperformed the best non-S2L baseline by point estimate ($R^2 = 0.181$ [95% CI: 0.050, 0.311] vs. 0.108, Pearson $r = 0.426$ vs. 0.336). Together, these results suggest that the transferred representation carries information about lesion extent across all three spatial axes. Bootstrap confidence intervals for the reported configurations are provided in Table 3 in the Appendix.

Cognitive outcomes showed a different pattern. MoCA was better predicted by raw speech embeddings than by the best S2L-containing configuration ($R^2 = 0.486$ vs. 0.455; Pearson $r = 0.702$ vs. 0.677). Adding an S2L residual model yielded a small further improvement, increasing R^2 to 0.501 [95% CI: 0.348, 0.622]. This suggests that MoCA prediction was primarily driven by raw speech embeddings, with modest additional benefit from S2L representations. In contrast, the multi-task cognitive composite was best predicted by combining raw speech embeddings with S2L ($R^2 = 0.234$ [95% CI: 0.055, 0.428]), outperforming raw speech embeddings alone ($R^2 = 0.205$).

Other lesion-geometry summaries were weaker. S2L representations outperformed non-S2L baselines for lesion anterior-posterior distribution and lesion component count, but the absolute R^2 values were close to zero across both S2L and non-S2L configurations (Table 3 in the Appendix), with bootstrap intervals spanning zero in all cases, indicating limited predictive signal in these targets from speech regardless of representation type. Lesion superior-inferior distribution was better predicted by demographics than by the S2L-containing configuration ($R^2 = -0.020$ vs. -0.127), and adding an S2L residual model further worsened performance ($R^2 = -0.183$), confirming that none speech carries no useful signal for this target.

Table 1. Best speech-to-lesion (S2L) vs best non-S2L configuration across the targets. Residual reports performance after adding an S2L residual model to the non-S2L baseline where tested. r denotes Pearson correlation. Clinical scores refer to individual cognitive assessment scores; the cognitive composite is their z-scored mean. Bold indicates the highest R^2; negative R^2 indicates worse performance than the mean-prediction baseline. Bootstrap 95% confidence intervals are provided in Table 3 in the Appendix.

Target	Best S2L set	S2L R^2	S2L r	Best non-S2L set	Non-S2L R^2	Non-S2L r	+S2L resid. R^2
MoCA	Speech emb. + S2L	0.455	0.677	**Speech embeddings**	**0.486**	**0.702**	**0.501**
Cognitive composite	**Speech emb. + S2L**	**0.234**	**0.491**	Speech embeddings	0.205	0.466	–
Lesion bbox x	**Speech feat. + S2L**	**0.181**	**0.426**	Speech embeddings	0.108	0.336	–
Lesion bbox y	**S2L**	**0.241**	**0.503**	Speech embeddings	-0.020	0.101	–
Lesion bbox z	**S2L**	**0.175**	**0.423**	Speech embeddings	-0.015	0.063	–
Lesion ant-post	**Clinical scores + S2L**	**0.013**	**0.133**	Demographics	-0.026	-0.125	–
Lesion component	**Clinical scores + S2L**	**-0.072**	**0.104**	Demographics	-0.113	-0.299	–
Lesion sup-inf	Clinical scores + S2L	-0.127	**0.109**	**Demographics**	**-0.020**	0.069	-0.183

6. Discussion

The results indicate a constrained but interpretable form of representation reuse. The absolute effect sizes are interpreted within a clinical context where R^2 values in the range of 0.15 to 0.20 are considered meaningful for multifactorial clinical outcomes, while values near 0.10 signify more limited explained variance (Gupta et al., 2024). Under this framework, the predictions for the lesion bounding box coordinates y ($R^2 = 0.241$), x ($R^2 = 0.181$), and z ($R^2 = 0.175$) all fall within a clinically interpretable range. While the y-axis exhibits the strongest association, the x and z dimensions remain supportive and provide significant predictive value compared to several other geometry targets that remain near-baseline.

Although S2L representation did not directly encode continuous bounding-box extent, its out-of-fold outputs appear to capture lesion extent variability, plausibly through lesion-burden and lobe-involvement supervision. This result extends classical lesion-symptom mapping, which uses lesion location or lesion load to explain behavioural impairment from structural MRI (Yourganov et al., 2015; Iorga et al., 2021). In contrast, our framework investigates whether speech can recover information about lesion geometry itself. The stronger results for bounding-box dimensions could be clinically plausible, as these capture lesion extent along anatomical axes, a coarse structural property related to language impairment (Marchina et al., 2011; Pustina et al., 2018; Na et al., 2022).

The negative or near-baseline results for anterior–posterior, superior–inferior, and component targets are also informative. Unlike bounding-box extent, these reflect coarse orientation or fragmentation rather than overall spatial spread, and their bootstrap intervals spanning zero support a cautious interpretation of limited transfer. This pattern suggests that the S2L representation is not a generic proxy for imaging-derived variables, appearing more sensitive to extent than to other geometric summaries. Cognitive outcomes followed a different pattern, with raw speech embeddings outperforming S2L representations for MoCA, and S2L adding only marginal residual signal. The resulting MoCA performance was nevertheless high for a clinical outcome prediction task ($R^2 = 0.501$). This suggests that cognitive screening is captured more by broad behavioural speech cues than by representations shaped toward focal lesion structure, though the residual contribution indicates that lesion-relevant information is not entirely orthogonal to cognitive status. This is consistent with prior work, where Wisler et al.(2020) recovered modest predictive signal from speech, language, and demographic variables alone, explaining 16.5% of MoCA variance in their study cohort.

When dense neuroimaging annotations are costly or inconsistently collected, S2L representations may provide a reusable proxy for lesion-relevant structure in downstream models. However, several limitations remain. The analysed sub-cohort is small ($N = 110$), and the wide bootstrap intervals reflect this directly; for several targets, R^2 values near zero with intervals spanning zero indicate that the null cannot be excluded at this sample size. Future work should test generalisation in external and more demographically diverse cohorts, and assess whether calibrated uncertainty estimates can identify predictions outside the model's reliable operating range.

7. Conclusion

This work shows that post-stroke speech may be able to carry structured signal about brain-injury geometry and cognitive impairment, connecting scalable behavioural measurements to targets usually derived from neuroimaging or in-person assessment. These results also speak to a broader challenge in clinical data and machine learning: imaging markers and speech variables are rarely complete across cohorts, and supervised targets are often sparsely observed across modalities. Larger external validation is needed, but these results suggest that speech-based representations may improve sample efficiency and support richer post-stroke brain-injury characterisation.

Impact Statement

This work aims to advance machine learning methods for post-stroke assessment by studying whether speech-derived representations can support prediction of brain-injury and cognitive phenotypes. If validated in larger and more diverse cohorts, such methods could help reduce reliance on costly or unevenly available clinical annotations and support more scalable post-stroke monitoring.

The work also raises important ethical considerations. Speech is sensitive personal data and may encode demographic, linguistic, cognitive, and health-related information beyond the intended prediction targets. Ethical approval and participant consent were obtained for the collection and use of the data analysed in this study; identifying details of the study sites and approvals will be disclosed after the anonymous review period. Models trained on speech may reflect cohort-specific biases, differences in language background, recording conditions, access to care, or clinical assessment practices. These systems should not be used as standalone diagnostic tools, and any clinical deployment would require external validation, uncertainty estimation and careful evaluation of fairness across demographic and linguistic groups.

References

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

Bland, M. D., Sturmoski, A., Whitson, M., Harris, H., Connor, L. T., Fucetola, R., Edmiaston, J., Huskey, T., Carter, A., Kramper, M., et al. Clinician adherence to a standardized assessment battery across settings and disciplines in a poststroke rehabilitation population. *Archives of physical medicine and rehabilitation*, 94(6):1048–1053, 2013.

Bowden, M., Beswick, E., Tam, J., Perry, D., Smith, A., Newton, J., Chandran, S., Watts, O., and Pal, S. A systematic review and narrative analysis of digital speech biomarkers in motor neuron disease. *NPJ digital medicine*, 6(1):228, 2023.

Bowren Jr, M., Bruss, J., Manzel, K., Edwards, D., Liu, C., Corbetta, M., Tranel, D., and Boes, A. D. Post-stroke outcomes predicted from multivariate lesion-behaviour and lesion network mapping. *Brain*, 145(4):1338–1353, 2022.

Bredin, H. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*, 2023.

Butler, R. A., Lambon Ralph, M. A., and Woollams, A. M. Capturing multidimensionality in stroke aphasia: mapping principal behavioural components to neural structures. *Brain*, 137(12): 3248–3266, 2014.

Chen, S., Li, L., Han, S., Luo, W., Wang, W., Yang, Y., Wang, X., Zhang, W., Chen, M., and Wang, Z. Review of voice biomarkers in the screening of neurodegenerative diseases. *Interdisciplinary Nursing Research*, 3(3):190–198, 2024.

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

den Ouden, D.-B., Malyutina, S., Basilakos, A., Bonilha, L., Gleichgerrcht, E., Yourganov, G., Hillis, A. E., Hickok, G., Rorden, C., and Fridriksson, J. Cortical and structural-connectivity damage correlated with impaired syntactic processing in aphasia. *Human brain mapping*, 40(7):2153–2173, 2019.

Ernst, M., Boers, A., Forkert, N., Berkhemer, O., Roos, Y., Dippel, D., van der Lugt, A., Van Oostenbrugge, R., Van Zwam, W., Vettorazzi, E., et al. Impact of ischemic lesion location on the mrs score in patients with ischemic stroke: a voxel-based approach. *American Journal of Neuroradiology*, 39(11):1989–1994, 2018.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., 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.

Flowers, H. L., Silver, F. L., Fang, J., Rochon, E., and Martino, R. The incidence, co-occurrence, and predictors of dysphagia, dysarthria, and aphasia after first-ever acute ischemic stroke. *Journal of communication disorders*, 46(3):238–248, 2013.

Forkel, S. J., Thiebaut de Schotten, M., Dell'Acqua, F., Kalra, L., Murphy, D. G., Williams, S. C., and Catani, M. Anatomical predictors of aphasia recovery: a tractography study of bilateral perisylvian language networks. *Brain*, 137(7):2027–2039, 2014.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

Geranmayeh, F., Leech, R., and Wise, R. J. Network dysfunction predicts speech production after left hemisphere stroke. *Neurology*, 86(14):1296–1305, 2016.

Gupta, A., Stead, T. S., and Ganti, L. Determining a meaningful r-squared value in clinical medicine. *Academic Medicine & Surgery*, 2024.

Honnibal, M. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *(No Title)*, 2017.

Iorga, M., Higgins, J., Caplan, D., Zinbarg, R., Kiran, S., Thompson, C. K., Rapp, B., and Parrish, T. B. Predicting language recovery in post-stroke aphasia using behavior and functional mri. *Scientific reports*, 11(1):8419, 2021.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. Fsl. *Neuroimage*, 62(2):782–790, 2012.

Kertesz, A. Western aphasia battery–revised. 2007.

Lazar, R. M. and Boehme, A. K. Aphasia as a predictor of stroke outcome. *Current neurology and neuroscience reports*, 17(11): 83, 2017.

Li, X., Song, H., Guo, N., Kang, C., Gong, X., Ji, X., and Jie, Z. Machine learning models in post-stroke aphasia: a scoping review. *Frontiers in Neurology*, 17:1806856, 2026.

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

Marchina, S., Zhu, L. L., Norton, A., Zipse, L., Wan, C. Y., and Schlaug, G. Impairment of speech production predicted by lesion load of the left arcuate fasciculus. *Stroke*, 42(8):2251–2256, 2011.

Mazziotta, J., Toga, A., Evans, A., Fox, P., Lancaster, J., Zilles, K., Woods, R., Paus, T., Simpson, G., Pike, B., et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 356(1412):1293–1322, 2001.

Mirheidari, B., Bell, S. M., Harkness, K., Blackburn, D., and Christensen, H. Spoken language-based automatic cognitive assessment of stroke survivors. *Language and Health*, 2(1): 32–38, 2024.

Na, Y., Jung, J., Tench, C. R., Auer, D. P., and Pyun, S.-B. Language systems from lesion-symptom mapping in aphasia: A meta-analysis of voxel-based lesion mapping studies. *NeuroImage: Clinical*, 35:103038, 2022.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., and Chertkow, H. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

Pustina, D., Avants, B., Faseyitan, O. K., Medaglia, J. D., and Coslett, H. B. Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations. *Neuropsychologia*, 115:154–166, 2018.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. pp. 28492–28518, 2023.

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., and Brecher, A. The philadelphia naming test: scoring and rationale. *Clinical aphasiology*, 24:121–133, 1996.

Roehrich-Gascon, D., Small, S. L., and Tremblay, P. Structural correlates of spoken language abilities: A surface-based region-of interest morphometry study. *Brain and language*, 149:46–54, 2015.

Salvalaggio, A., De Filippo De Grazia, M., Zorzi, M., Thiebaut de Schotten, M., and Corbetta, M. Post-stroke deficit prediction from lesion and indirect structural and functional disconnection. *Brain*, 143(7):2173–2188, 2020.

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

Swinburn, K., Porter, G., and Howard, D. Comprehensive aphasia test. *APA PsycTests*, 2004.

Thomas, M. R., Gudnason, J., and Naylor, P. A. 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.

Wardlaw, J., Brazzelli, M., Miranda, H., Chappell, F., McNamee, P., Scotland, G., Quayyum, Z., Martin, D., Shuler, K., Sandercock, P., et al. An assessment of the cost effectiveness of magnetic resonance including diffusion-weighted imaging in patients with transient ischaemic attack and minor stroke. a systematic review, meta-analysis and economic evaluation. *Health technology assessment*, 18(27):1–368, 2014.

Weigel, K., Gaser, C., Brodoehl, S., Wagner, F., Jochmann, E., Güllmar, D., Mayer, T. E., and Klingner, C. M. Acute stroke severity assessment: The impact of lesion size and functional connectivity. *Brain Sciences*, 15(7):735, 2025.

Wisler, A. A., Fletcher, A. R., and McAuliffe, M. J. Predicting montreal cognitive assessment scores from measures of speech and language. *Journal of Speech, Language, and Hearing Research*, 63(6):1752–1761, 2020.

Wolf, T. and Debut, L. e. a. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Yang, Y., Zhao, X., Zhao, P., Ying, D., Wang, J., Jiang, Y., and Wan, Q. Ai-driven speech biomarkers for disease diagnosis and monitoring: a systematic review and meta-analysis. *BMJ Evidence-Based Medicine*, 31(1):46–56, 2026.

Yourganov, G., Smith, K. G., Fridriksson, J., and Rorden, C. Predicting aphasia type from brain damage measured with structural mri. *Cortex*, 73:203–215, 2015.

Appendix

Table 2. Description of downstream prediction targets used in the paper.

Target	Variable family	Description
MoCA	Cognitive	Montreal Cognitive Assessment score at session 1; global cognitive screening measure.
Cognitive composite	Cognitive	Composite computed from the cognitive/task assessments collected at the time of the speech task. These include orientation, attention, memory, calculation, comprehension, semantic judgement, visuospatial processing, task recall, and activities of daily living.
Lesion bbox x	Focal lesion geometry	Left-right spatial extent of the focal lesion mask, computed as the bounding-box width along the x axis in MNI space.
Lesion bbox y	Focal lesion geometry	Anterior-posterior spatial extent of the focal lesion mask, computed as the bounding-box width along the y axis in MNI space.
Lesion bbox z	Focal lesion geometry	Inferior-superior spatial extent of the focal lesion mask, computed as the bounding-box width along the z axis in MNI space.
Lesion component count	Focal lesion geometry	Number of spatially connected components in the focal lesion mask after binarization.
Lesion sup-inf index	Focal lesion geometry	Relative superior-inferior distribution of the focal lesion mask; summarizes whether lesion burden is more superior or inferior.
Lesion ant-post index	Focal lesion geometry	Relative anterior-posterior distribution of the focal lesion mask; summarizes whether lesion burden is more anterior or posterior.

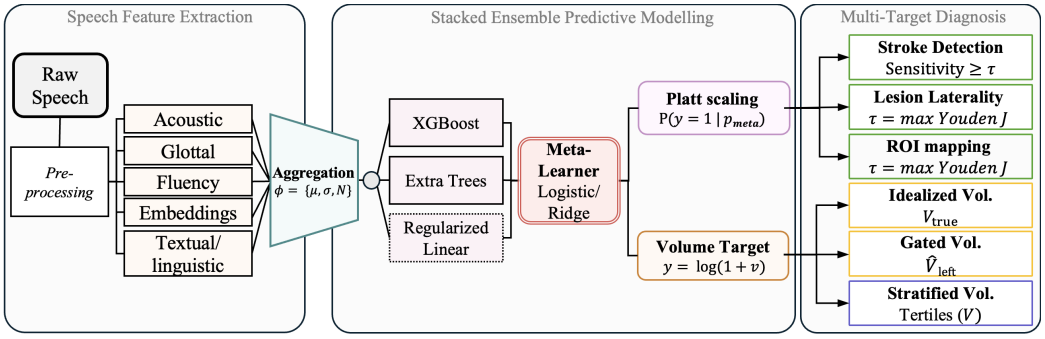

Figure 1. **Speech-to-lesion (S2L) Pipeline.** The framework is a hierarchical multi-head architecture organized according to clinically interpretable levels of lesion characterization. Each head utilizes three base learners: (a) XGBoost ($n_{\text{estim}} = 800$, learning rate $= 0.05$, max depth $= 4$, subsample $= 0.85$, colsample $= 0.85$, $L_2 = 1.0$, bagging over seeds $\{0, 1, 2\}$, early stopping patience $= 40$), (b) Extremely Randomized Trees ($n_{\text{estim}} = 600$, max_features=`sqrt`, min_samples_leaf $= 2$), and (c) an L_2-regularized linear model. Base predictions are fused using a logistic regression meta-learner for classification and a ridge regression meta-learner for volumetry, with ridge α selected from $\{1, 10, 10^2, 10^3, 10^4\}$ via inner-validation. Stroke detection operates under a sensitivity-constrained threshold, reflecting clinically prioritised true-positives, while the other binary thresholds were selected via Youden's criterion to balance sensitivity and specificity. For classification, meta-probabilities p_{meta} are calibrated via Platt scaling. Lesion volume targets include (i) idealized regression (V_{true}) trained on confirmed left-hemisphere manual clinical labels, (ii) gated regression ($\hat{V}_{\text{left}}$) activated upon predicted laterality, and (iii) three-class lesion size stratification into tertiles. Volume is modeled as $y = \log(1 + v)$ to reduce skewness. Regional and volumetric heads focus on left-hemisphere damage, retaining frontal ($N = 298$), parietal ($N = 221$), temporal ($N = 210$), subcortical ($N = 210$), and occipital ($N = 62$) ROIs. Heads were trained only when $n_{\text{train}} \geq 75$, excluding the left cerebellum ($n = 11$).

Table 3. Bootstrap 95% confidence intervals for the configurations reported in Table 1. Intervals were computed from patient-level out-of-fold predictions. Clinical scores refer to individual cognitive assessment scores.

Target	Configuration	R^2 [95% CI]	Pearson r [95% CI]
MoCA	Speech emb. + S2L residual	0.501 [0.350, 0.620]	0.713 [0.603, 0.801]
Cognitive composite	Speech emb. + S2L	0.234 [0.055, 0.428]	0.491 [0.330, 0.665]
Lesion bbox x	Speech feat. + S2L	0.181 [0.050, 0.311]	0.426 [0.194, 0.612]
Lesion bbox y	S2L	0.241 [0.026, 0.394]	0.503 [0.343, 0.647]
Lesion bbox z	S2L	0.175 [0.033, 0.271]	0.423 [0.253, 0.565]
Lesion ant-post	Clinical scores + S2L	0.013 [-0.102, 0.099]	0.133 [-0.082, 0.364]
Lesion component	Clinical scores + S2L	-0.072 [-0.302, 0.076]	0.104 [-0.090, 0.303]
Lesion sup-inf	Demographics	-0.020 [-0.143, 0.063]	0.069 [-0.141, 0.269]