A Multimodal Framework for Aphasia Severity Classification in Russian

Anastasia Kolmogorova[1], **Ekaterina Yavshits**[1], **Anastasia Margolina**[1], **Anna Sugian**[1]
[1]HSE University, Saint Petersburg, Russia
Correspondence: akolmogorova@hse.ru

Abstract

Automatic classification of aphasia severity presents persistent challenges, particularly for languages with limited clinical speech resources such as Russian. This paper explores a multimodal approach to severity estimation that combines acoustic and semantic representations of pathological speech. Acoustic features are extracted using pretrained Wav2Vec 2.0 models, while semantic information is obtained from the encoder of the Whisper model. The two representations are integrated via early feature fusion and evaluated using gradient boosting classifiers in a speaker-independent cross-validation setting. Experiments are conducted on a newly collected dataset of Russian speech recordings from patients with aphasia and neurotypical speakers (RuAphasiaBank). The results suggest that the combined use of acoustic and semantic embeddings can provide more stable severity estimates than unimodal baselines. This study contributes empirical evidence on the applicability of multimodal representation learning for aphasia severity classification under data-scarce conditions.

1 Introduction

Aphasia is a systemic speech disorder characterized by complete or partial loss of speech, resulting from localized damage to one or more areas of speech in the brain (Wiesel, 2002). This condition manifests itself in various forms. Within the framework of Russian neurolinguistics and neuropsychology, the aphasia typology proposed by A.R. Luria is predominantly used, distinguishing between three types of motor aphasia (afferent, efferent, dynamic) and four types of sensory aphasia (acoustic, acoustic-mnestic, optic-mnestic, semantic) (Luria, 1973). The distinctive speech production characteristics of the five aphasia types represented in our dataset are as follows. Afferent motor aphasia is marked by articulatory difficulties, especially with complex sounds (articulatory apraxia) and literal paraphasias where sounds are substituted with phonetically similar ones. Efferent motor aphasia is characterized by perseverations, syllable transpositions within words, and a general disruption of speech fluency. Acoustic (sensory) aphasia presents with fluent but largely incomprehensible output, often described as "verbal salad," and frequent verbal paraphasias. Acoustic-mnestic aphasia primarily involves a word-finding deficit (anomia), where patients can describe objects but fail to recall their names. Finally, semantic aphasia entails difficulties in understanding complex logical-grammatical and semantic relationships (e.g., agent-patient relations) (Tsvetkova, 1988; Shokhor-Trotskaya, 2001).

Automatic classification methods for aphasia types based on acoustic fragments of patients' expressive speech constitute one of the most pressing challenges in atypical speech processing. The interest in this problem is driven by two key factors: the growing number of patients and the corresponding increase in the workload of speech-language pathologists, as well as the necessity to develop applications and interactive environments for speech therapy in home settings.

However, several significant obstacles impede progress in solving this problem. Primarily, this is the absence or insufficiency of speech data for fine-tuning models. The largest corpus of aphasic speech is the AphasiaBank collection (Forbes et al., 2012); however, its data are also insufficient for comprehensive training of neural network models: for 7 languages, there are recordings of a total of 180 individuals with aphasia and 140 neurotypical informants. The situation is even more challenging for the Russian language. Until now, there existed only one corpus of narratives from patients with post-stroke aphasia of 4 types, without distinction by severity of impairment (Khudyakova et al., 2016), obtained from 40 patients retelling the "Pear Film" by W. Chafe. We have compiled

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 257–265

a unique dataset comprising 142 recordings from 70 patients and 20 recordings from 20 neurotypical informants.

The main advantage of this dataset, despite its relatively modest size, is the representation of different aphasia types (6 of of the 7 types distinguished by A.R. Luria) and varying degrees of speech impairment severity (from mild to severe), a diversity of patient age characteristics, and the presence of recordings of different speech genres (reading, dialogue, monologue, retelling).

Nevertheless, the problem of limited data for fine-tuning neural network models in the task of classifying Russian-language audio recordings of patients with aphasia by severity of impairment remains unresolved within our dataset as well. To address this, we decided to apply a dual-embedding, multimodal framework as a potential method to overcome the data scarcity issue. The core idea is to combine complementary representations of speech — encoding motor-acoustic and cognitive-linguistic dimensions, — via early feature fusion. To assess the contribution of each modality, we evaluated three model variants:

1. An acoustic-only baseline (Wav2Vec 2.0).
2. A semantic-only baseline (Whisper).
3. Our proposed multimodal fusion model, which concatenates both embedding vectors for a joint representation.

The obtained results demonstrate an optimistic prognosis regarding the potential of the chosen methodology.

Thus, our research question is as follows :

Can the dual-embedding, multimodal framework become an effective method for the automatic classification of aphasia severity in a data-limited scenario?

The aim of this publication is to describe the results obtained by applying this method to our dataset of recordings from Russian-speaking patients with aphasia.

Our primary contributions in this paper are: (1) our work pioneers the use of Wav2Vec2 embeddings for severity classification of Russian aphasic speech, (2) introduces a dual-embedding framework tailored for small, heterogeneous corpora, and (3) offers a comparative analysis of acoustic, semantic, and fused modeling pipelines.

The article is structured as follows. First, we present a literature review. Subsequently, we provide a detailed description of our RuAphasiaBank dataset. The following section outlines the methodology and presents the results of experiments on classifying aphasic speech recordings based on vector representations of both acoustic features and semantics. Finally, the conclusion discusses the prospects for future work.

2 Related Papers

The task of pathological speech classification (including dysarthria, speech in Parkinson's and Alzheimer's diseases, and aphasia) is being actively addressed through various approaches:

1. By extracting features from speech transcripts using classical machine learning methods (Kim et al., 2015) or by leveraging GPT models (Chi et al., 2022).

2. By employing vector representation models for acoustic and articulatory information (Ríos-Urrego et al., 2023).

Supervised machine learning methods applied to extensive labeled datasets remain the most common approach for analyzing pathological speech recordings (Vásquez-Correa et al., 2017). However, creating such datasets is inherently challenging, primarily due to the difficulties of clinical data collection. A frequently used workaround is to rely on transcripts generated by Automatic Speech Recognition (ASR) engines (Choi et al., 2024). Yet, the recognition quality for atypical speech remains critically low, even after often costly fine-tuning of these models. As a consequence of the aforementioned limitations and challenges, researchers and developers are increasingly prioritizing methods based on the vectorization of acoustic features using the family of Wav2Vec-like models (Švec et al., 2022).

The advent of SSL (Self-Supervised Learning) methods has revolutionized the classification of pathological speech data. By enabling effective model training on small datasets, SSL offers a practical solution to the long-standing problem of data scarcity in this clinical domain. Wav2Vec2 employs a three-stage pipeline to convert raw audio into meaningful representations (Baevski et al., 2020). First, a CNN-based feature extractor compresses the audio signal into a latent space. This latent sequence is then fed into a transformer encoder, which aggregates contextual information to generate high-level continuous embeddings. A final quantization step maps these embeddings onto a discrete codebook, a crucial mechanism for the model's self-supervised learning objective.

Wav2Vec2 is actively integrated into various pipelines for atypical speech classification. Researchers utilize the pre-trained Wav2Vec2 architecture for both self-supervised learning (SSL) and automatic speech recognition (ASR), and also employ it as a powerful feature extractor in automated speech assessment tools (Nguyen et al., 2024). The Wav2Vec2 model has been effectively utilized in standard classification pipelines, showing consistent and reliable results for the task of dysarthria severity classification (Javanmardi et al., 2023).

While Wav2Vec methods have advanced atypical speech classification, they face limitations: their acoustic features can lack granular detail and prove inadequate for nuanced tasks. A more promising strategy is to employ dual embeddings capturing both acoustic and linguistic properties — an approach already benefiting fields like speech emotion recognition (Atmaja et al., 2022; Pepino et al., 2020). We posit that the synergy between these two vector types will significantly boost the accuracy of classifying aphasia severity, even when data is scarce.

3 RuAphasiaBank description

3.1 Data collection

The data for the dataset were collected during collaborative work with clinicians from the Federal Scientific-Clinical Center of the Federal Medical and Biological Agency of Russia in Krasnoyarsk. Within the framework of the project, speech materials for speech therapy were developed depending on patients' sociolinguistic characteristics and language biographies (Yastrebtseva et al., 2023).

Recordings were made using built-in microphones in iPhone smartphones (models 13–16) during speech therapy sessions conducted under clinical conditions. For dataset collection, informed consent was obtained from both the patients and their legal representatives, following approval from the clinic's ethics committee for the use of an iPhone in offline mode as a recording device. Prior to each session, the speech therapist created a metadata file describing the patient's status, and the patient provided specific consent for the recording and subsequent processing of anonymized data. Following the recording, the audio file was assigned a coded, anonymized name according to a coding table and transferred on a flash drive to the research laboratory. There, it was converted to WAV format, and all potentially identifying information including personal names and toponyms related to the patient's place of residence was meticulously removed from both dialogues and monologues.

The recordings were performed in the clinic's speech therapy room in a conversational setting at a table, where the speech therapist and the patient were seated side by side, with an iPhone placed on the table between them at a distance of 30–40 cm from the speaker's face. The acoustic properties of this environment provide optimal conditions for clear speech perception, ensuring comfort for both the patient and the specialist while minimizing external noise and distortions due to low reverberation, uniform sound distribution, and the absence of background noise. Speech recordings from neurotypical respondents were acquired in residential living rooms, maintaining consistent acoustic conditions and standardized interaction protocols. As a result, the collected data closely approximate natural everyday communication settings while remaining minimally affected by interference or noise artifacts.

3.2 Description of data obtained from patients with aphasia

Table 1 presents general characteristics of the patients diagnosed with aphasia (PWA), as well as of the speech therapists who worked with the patients in a dialogue mode, along with quantitative details regarding the recordings of this group.

Factor	Categorization	Overall
Records	Total duration	9,8 h
	Total number	162 rec.
People	PWA	70 pts.
	Therapist (in dialogues)	8 ther.
Gender	Male	54 pts.
	Female	16 pts.
Age	Mean male	53.4 yrs.
	Mean female	61.3 yrs.

Table 1: PWA recordings general features. Abbreviations meaning: h — hours, rec. — recordings, pts. — patients, ther. — therapists, yrs. — years.

The gender and age distribution of patients reflects the general trend observed at the partner healthcare institution of the project. According to 2018-2025 statistics, men account for 67% of patients, while women constitute only 33%. The most vulnerable age range for aphasic disorders also dif-

fers between male and female cohorts: 50-65 years for men compared to 61-73 years for women. The observed gender disparity in aphasia prevalence aligns with cross-linguistic evidence, where similar demographic patterns have been documented in patient cohorts speaking different native languages, including Russian, Spanish, English (Kolmogorova et al., 2023; Rojas, 2020; Sharma et al., 2019) and other languages.

Table 2 and Table 3 present the distribution of patients according to speech status (type of aphasia and severity of impairment). Most of the patients included in the dataset have been diagnosed with complex motor aphasia; however, recordings from patients with five other types of aphasia are also represented, based on A.R. Luria's classification (Luria, 2008). Most patients have severe or moderate impairment, although mild impairment is also present. Additionally, there are limited data available on the medications being taken by the patients, as well as the sequential number of the neurorehabilitation program each patient was undergoing.

Aphasia type	Overall	Duration
Complex motor aphasia	58 pts.	8.3 h
Sensory aphasia	5 pts.	0.77 h
Semantic aphasia	2 pts.	0.15 h
Efferent motor aphasia	2 pts.	0.31 h
Acoustic-gnostic aphasia	1 pt.	0.09 h
Total aphasia	2 pts.	0.2 h

Table 2: PWA aphasia types distribution. Abbreviations meaning: pts. – patients, pt. – patient, h – hours.

Severity	Overall	Duration
Moderate	31 pts.	5.3 h
Severe	34 pts.	3.9 h
Mild	5 pts.	0.55 h

Table 3: PWA impairment severity distribution. Abbreviations meaning: pts. – patients, h – hours.

Finally, Table 4 presents data on the types of speech represented in the dataset (monologic speech, dialogues, retellings, reading of texts and individual syllables), as well as the types of texts read by the patients (reading recordings). In the task of automatic speech recognition, the presence of such texts allows for immediate access to ground truth data for evaluating recognition quality. The

Factor	Categorization	Overall
Type of speech		
	Dialogue	85 rec. (52 pts.)
	Monologue	19 rec. (14 pts.)
	Reading	27 rec. (14 pts.)
	Retelling	18 rec. (11 pts.)
	Syllable reading	13 rec. (11 pts.)
Types of texts in reading		
	Didactic texts	17 rec. (10 pts.)
	Fiction texts	29 rec. (9 pts.)
	Simplified texts from our sample	21 rec. (11 pts.)

Table 4: PWA speech settings. Abbreviations meaning: rec. – recordings, pts. – patients.

dataset contains recordings of patients reading didactic texts intended for children, as well as texts of classical Russian literature, consisting either of nature descriptions or accounts of interesting events. All of these texts are included in methodological guides for speech therapy and are actively used by Russian speech-language pathologists in their practical work.

Additionally, we included recordings of patients reading texts that were simplified using our automatic simplification algorithms (the RuSentAphasia dataset (Kolmogorova and Solovyova, 2024)). The originals of these simplified texts were drawn from literary or media sources

3.3 Data from neurotypical informants

Factor	Categorization	Overall
Records	Total duration	4.3 h
	Total number	24 rec.
People	Informants	20 ppl.
	Interviewer (in dialogues)	8 ppl.
Gender	Male	16 ppl.
	Female	4 ppl.
Age	Mean male	56.1 yrs.
	Mean female	70.2 yrs.

Table 5: General features of data obtained from neurotypical informants. Abbreviations meaning: h – hours, rec. – recordings, ppl. – people, yrs. - years.

To ensure demographic comparability, the neurotypical group was constructed to parallel the patient cohort in terms of gender distribution (with a

male predominance) and age, where the majority of male participants fell within the 53–56 year age range.

4 Experiments with RuAphasiaBank

4.1 Methodology and Setup

To jointly capture both low-level motor-speech impairments and higher-level cognitive–linguistic deficits associated with aphasia, we construct two complementary representations for each speech recording, corresponding to acoustic–motor and semantic–linguistic dimensions of impairment.

The acoustic representation is obtained by extracting a 768-dimensional embedding from the final hidden layer of the pretrained wav2vec 2.0 Base model (Baevski et al., 2020). These embeddings primarily encode language-independent properties of the speech signal, such as articulation stability, prosodic control, speech rate, and temporal regularity. Such cues are particularly informative for detecting motor-speech disturbances and dysfluencies, that frequently accompany moderate and severe forms of aphasia.

In parallel, we derive a semantic representation by feeding the same recordings into the encoder of the Whisper Large model (Radford et al., 2022) and extracting a 1280-dimensional embedding. In contrast to the acoustic features, these representations capture higher-level linguistic organization, including lexical diversity, syntactic structure, and overall semantic coherence of the produced utterance, which are known to degrade even in relatively mild aphasic conditions.

Although neither model is trained specifically on Russian pathological speech, Whisper is a large-scale multilingual model and has been shown to transfer well across languages and recording conditions (Radford et al., 2022). In addition, self-supervised wav2vec 2.0 representations are widely used as general-purpose acoustic embeddings and have proven effective as feature extractors in pathological-speech assessment under limited supervision (Baevski et al., 2020; Nguyen et al., 2024; Javanmardi et al., 2023).

Based on these two modalities, we evaluate three model configurations. The acoustic-only baseline employs an XGBoost classifier (Chen and Guestrin, 2016) trained exclusively on wav2vec 2.0 embeddings, focusing on motor-speech characteristics. The semantic-only baseline uses the same classifier architecture, but operates solely on Whisper encoder embeddings, targeting cognitive–linguistic impairments. Finally, the multimodal fusion model combines both representations via early fusion by concatenating the acoustic and semantic feature vectors into a single high-dimensional input:

$$X_{combined} = [X_{acoustic}; X_{semantic}]$$

This fused representation is then used as input to an XGBoost classifier. XGBoost is chosen across all configurations due to its ability to efficiently handle high-dimensional dense features and to model non-linear interactions between heterogeneous feature subsets, which is particularly important in the fusion setting where acoustic and semantic cues may interact in clinically meaningful ways.

In preliminary experiments, we also evaluated a Random Forest classifier (Breiman, 2001) as a non-boosting baseline, using the same acoustic, semantic, and fused feature representations. While Random Forests are commonly used for tabular clinical data, we observed consistently inferior performance compared to gradient boosting methods, particularly in terms of MAE and stability across cross-validation folds. This degradation is likely due to the very high-dimensional input space (up to 2048 concatenated features in the fusion setting) and the relatively small sample size, which make Random Forests prone to noisy splits and suboptimal feature utilization. Consequently, we focus our analysis on XGBoost, which demonstrated superior robustness and predictive accuracy in all configurations.

All models are trained and evaluated on the same corpus of 187 recordings from 90 unique speakers. To prevent patient-level data leakage and to approximate a realistic clinical deployment scenario, we adopt Group K-Fold cross-validation with k=5, using the speaker identity as a grouping variable. Consequently, all recordings produced by a given patient are assigned either to the training split or to the test split in each fold, but never to both. Aphasia severity is encoded as an ordinal variable with four levels (*Neurotypical*, *Mild*, *Moderate*, *Severe*). Because the labels form an ordered scale rather than a set of independent categories, we use Mean Absolute Error (MAE) between true and predicted severity levels as the primary evaluation metric, in addition to standard classification metrics.

The class distribution in the corpus is imbalanced, with *Moderate* and *Severe* cases dominating and *Mild* cases being relatively rare. To mitigate this, we apply class-weighted learning in all three models by assigning sample weights inversely pro-

portional to the empirical class frequencies. This encourages the classifier to allocate sufficient capacity to under-represented classes and reduces the tendency to bias predictions toward the majority categories.

4.2 Results

We first compare model configurations in terms of overall severity prediction accuracy (Figure 1). Across all feature sets, XGBoost consistently outperforms Random Forest baselines (evaluated in preliminary experiments) and achieves lower MAE with reduced variance across folds. Among the unimodal settings, semantic features derived from Whisper outperform purely acoustic representations, suggesting that lexical and structural information is particularly informative for severity estimation. However, the best performance is achieved by the multimodal fusion model with XGBoost (MAE = 0.431), improving over both unimodal XGBoost baselines (0.452 for Whisper and 0.569 for Wav2Vec2). For Random Forest, in contrast, fusion does not yield an improvement: the lowest MAE is obtained with Whisper features (0.457), while the combined feature set reaches 0.484.

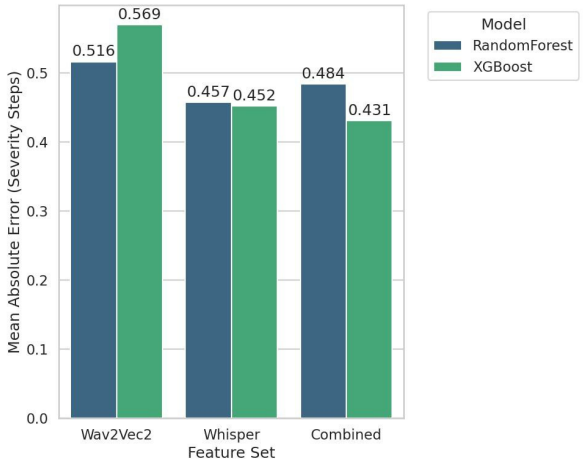

Figure 1: Classification performance (MAE) across model configurations.

To better understand model behavior across clinical subtypes, we further analyze prediction errors stratified by aphasia type (Figure 2). Error rates vary substantially between diagnostic categories, with the lowest MAE observed for neurotypical speakers and total aphasia, and higher errors for sensory and semantic aphasia. These patterns are consistent with clinical expectations: conditions characterized by diffuse or subtle impairments present a greater challenge for automated

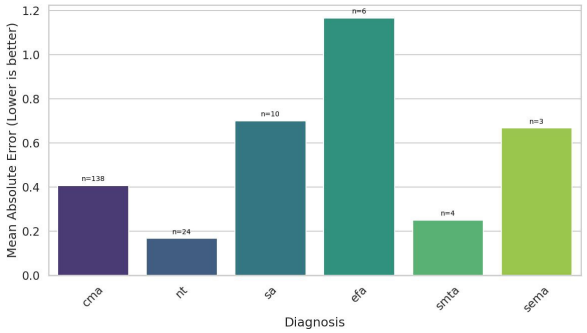

Figure 2: Prediction Error (MAE) by Aphasia Type (n $\geq$ 3). Abbreviations meaning: cma – complex motor aphasia, nt – neurotypical, sa – sensory aphasia, efa – efferent motor aphasia, smta – total aphasia, sema – semantic aphasia.

Class	Precision	Recall	F1-Score	Support
Neurotypical	0.85	0.92	0.88	24
Mild	0.50	0.33	0.40	12
Moderate	0.61	0.66	0.63	90
Severe	0.55	0.52	0.53	60
Weighted Avg	0.62	0.62	0.62	186

Table 6: Performance evaluation of the fusion classification model.

severity estimation, particularly under limited data availability.

Table 6 reports per-class precision, recall, and F1-scores for the fusion model. Performance is strongest for the *Neurotypical* and *Moderate* classes, which are both well represented in the corpus. In contrast, the *Mild* category remains the primary bottleneck (F1 = 0.40, n = 12), with most errors corresponding to confusions with the *Moderate* class. This indicates that the model reliably detects the presence of impairment but struggles to distinguish early or subtle deficits from more pronounced cases when labeled examples are scarce.

The confusion matrix of the fusion model (Figure 3) reveals two clinically important patterns. First, the system behaves as a reliable screening tool: no *Severe* patients are misclassified as healthy controls, and only two healthy speakers are erroneously flagged as pathological. This asymmetry suggests that the model is conservative with respect to false negatives at the severe end of the spectrum, which is preferable in a screening context.

Second, the primary performance bottleneck is the *Mild* category, which achieves an F1-score of 0.40 with only 12 supporting instances. Most errors for this class correspond to misclassifications

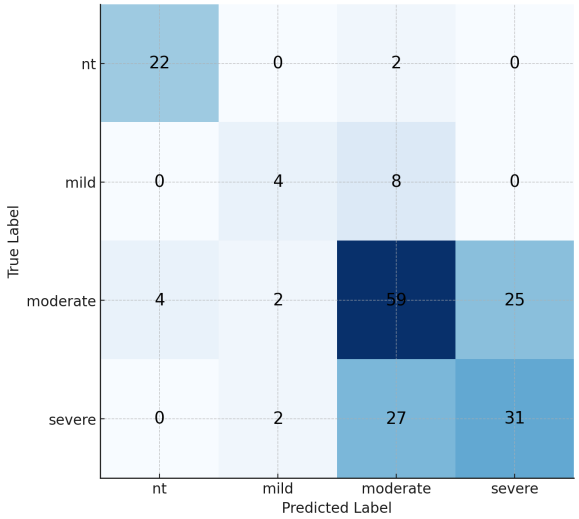

Figure 3: Confusion matrix of the fusion model.

of *Mild* as *Moderate*, indicating that the model reliably detects the presence of impairment but struggles to separate subtle cases from more pronounced deficits when labeled data are scarce.

Comparing the fusion model to the unimodal baselines, we observe that incorporating both acoustic and semantic representations improves not only MAE but also the weighted F1-score. Intuitively, the acoustic channel captures prosodic and articulatory deviations that are particularly informative for distinguishing healthy controls from pathological speech, while the semantic channel encodes lexical and structural abnormalities that refine severity estimation.

A direct comparison of the obtained metrics (F1-score 40–88%) with the results of previous studies, which reported F1-scores of 78–97% (Wagner et al., 2023; Nivedha et al., 2023; Qin et al., 2020), must account for key methodological differences. While the cited research addressed narrower tasks—such as classification into three aphasia types or binary severity classification (severe vs. mild)—using corpora in other languages, our study is the first to tackle the complex problem of multiclass classification (five aphasia types and three severity levels) on a Russian-language speech corpus. This inherently increases the task complexity and explains the somewhat lower absolute performance values.

5 Conclusion and Future Work

This study describes experiments with a dual-embedding multimodal framework for the automatic classification of aphasia severity in Russian, under conditions of limited and heterogeneous data. Using the newly developed dataset RuAphasiaBank, which encompasses multiple aphasia types, levels of impairment, and speech genres, we conducted a comparative evaluation of acoustic-only, semantic-only, and fused modeling pipelines.

The results indicate that the early fusion of Wav2Vec 2.0 acoustic embeddings with Whisper-based semantic representations enables a more comprehensive modeling of pathological speech by jointly capturing its motor-acoustic and cognitive-linguistic dimensions. The proposed multimodal approach demonstrated robust performance and exceeded the chance-level classification baseline, particularly in distinguishing moderate aphasia from other severity levels.

The favourable outcomes for the multimodal model may be partially attributable to the variety of data employed, which included multiple aphasia subtypes and speech production tasks. Nevertheless, this interpretation remains speculative and warrants empirical validation, for instance, by utilizing more controlled, homogeneous datasets.

These findings support the hypothesis that complementary multimodal representations can partially compensate for data scarcity, a critical constraint in atypical speech research. Notably, this work shows that a relatively small yet carefully curated and diverse dataset can serve as a meaningful testbed for modern self-supervised and multimodal methods in clinical linguistics. The proposed framework is not intended to replace expert clinical assessment but rather to complement it by providing scalable and reproducible severity estimates.

Current expert annotation by speech therapists reveals that severity grading (mild/moderate/severe) is more clearly defined for some aphasia types than others (e.g., semantic aphasia poses greater difficulty). We intend to balance our sub-corpora across severity levels for each aphasia type and subsequently analyze how this typological factor influences classification accuracy for severity.

Our future work will focus on integrating Russian-specific and domain-adapted SSL models, expanding the dataset with longitudinal recordings, and investigating the relative contributions of different speech genres and modalities. Overall, this study highlights the promise of embedding-based multimodal approaches for developing practical tools for aphasia assessment in low-resource language settings.

Limitations

The conducted experiments possess certain acknowledged limitations. Firstly, a multilingual, rather than Russian-specific, wav2vec model was employed. Secondly, the dataset exhibits variability in recording quality due to the use of different iPhone models. On the other hand, this very heterogeneity of the data can be viewed as an advantage: the tested method enables the differentiation of moderate aphasia from other conditions with precision and recall rates exceeding those of a random classifier, even given such acoustically diverse input. Furthermore, future experiments appear promising, particularly those involving Russian-language wav2vec models, as well as multilingual models with varied parameter configurations.

Acknowledgments

The article and the dataset were prepared within the framework of the Basic Research Program at HSE University in 2026. The authors are registered rights holders of the database described in this study. The database "RuAphasiaBank: database of phonograms of speech of patients with aphasia" is registered with the *Russian Federal Service for Intellectual Property* (No. 6.0021-2025). The authors have no competing interests to declare that are relevant to the content of this article.

Compliance with ethical standarts statement

This study was approved by the *Commission on Internal University Surveys* and *Ethical Evaluation of Empirical Research Projects of the Higher School of Economics*.

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Bagus Tris Atmaja, Akira Sasou, and Masato Akagi. 2022. Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion. *Speech Communication*, 140:11–28.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Nathan A. Chi, Peter Washington, Aaron Kline, Arman Husic, Cathy Hou, Chloe He, Kaitlyn Dunlap, and Dennis P. Wall. 2022. Classifying autism from crowdsourced semi-structured speech recordings: A machine learning approach. *arXiv*. ArXiv:2201.00927.

Yerin Choi, Jeehyun Lee, and Myoung-Wan Koo. 2024. Speech recognition-based feature extraction for enhanced automatic severity classification in dysarthric speech. *arXiv*. ArXiv:2412.03784.

Margaret M. Forbes, Davida Fromm, and Brian MacWhinney. 2012. Aphasiabank: A resource for clinicians. *Seminars in Speech and Language*, 33(3):217–222.

Farhad Javanmardi, Saska Tirronen, Manila Kodali, Sudarsana Reddy Kadiri, and Paavo Alku. 2023. Wav2Vec-based detection and severity level classification of dysarthria from speech. In *Proceedings of ICASSP 2023*, pages 1–5.

Maria V. Khudyakova, Maria B. Bergelson, Yulia S. Akinina, Elena V. Iskra, Svetlana Toldova, and Olga V. Dragoy. 2016. Russian CLIPS: A corpus of narratives by brain-damaged individuals. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC)*, pages 22–26.

Jinyu Kim, Nikhil Kumar, Andreas Tsiartas, Ming Li, and Shrikanth S. Narayanan. 2015. Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech and Language*, 29(1):132–144.

Anastasia Kolmogorova, Svetlana Lyamzina, and Olga N. Nikolskaya. 2023. Linguistic biography of an individual as a tool for modeling mental vocabulary. *Tomsk State University Journal of Philology*, 81:30–50. In Russian.

Anastasia V. Kolmogorova and Marina V. Solovyova. 2024. Text simplification methodology based on cognitive difficulty criteria for patients with aphasia. In *Proceedings of the Tenth International Conference on Cognitive Science*, pages 178–180. In Russian.

Alexander R. Luria. 1973. *Foundations of Neuropsychology*. Moscow State University Press, Moscow. In Russian.

Alexander R. Luria. 2008. *Higher human cortical functions*. Piter. In Russian.

Tuan Nguyen, Corinne Fredouille, Alain Ghio, Mathieu Balaguer, and Virginie Woisard. 2024. Exploring pathological speech quality assessment with ASR-powered wav2vec 2.0 in data-scarce context. *arXiv*. ArXiv:2403.20184.

E. Nivedha, A. Chandrasekar, and S. Jothi. 2023. An optimal hybrid AI-ResNet for accurate severity detection and classification of patients with aphasia disorder. *Signal, Image and Video Processing*, 17:3913–3922.

Leonardo Pepino, Pablo Riera, Luciana Ferrer, and Agustín Gravano. 2020. Fusion approaches for emotion recognition from speech using acoustic and text-based features. In *Proceedings of ICASSP 2020*, pages 6484–6488.

Yu Qin, Yuxuan Wu, Tan Lee, and Anthony P. H. Kong. 2020. An end-to-end approach to automatic speech assessment for cantonese-speaking people with aphasia. *Journal of Signal Processing Systems*, 92(8):819–830.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv*. ArXiv:2212.04356.

Cristian David Ríos-Urrego, Jan Rusz, Elmar Nöth, and Juan Rafael Orozco-Arroyave. 2023. Automatic classification of hypokinetic and hyperkinetic dysarthria based on GMM supervectors. In *Proceedings of Interspeech 2023*, pages 2368–2372.

Macarena Rojas. 2020. Post-stroke aphasia in spanish language: The effect of demographic variables. *Aphasiology*.

Sonal Sharma, Patrick M. Briley, Heather H. Wright, Joseph L. Perry, Xiaofeng Fang, and Charles Ellis. 2019. Gender differences in aphasia outcomes: Evidence from the AphasiaBank. *International Journal of Language and Communication Disorders*, 54(5):806–813.

Maria K. Shokhor-Trotskaya. 2001. *Speech and aphasia*. Akademicheskii Proekt, Moscow. In Russian.

Jan Švec, Filip Polák, Aleš Bartoš, Michaela Zapletalová, and Martin Víta. 2022. Evaluation of Wav2Vec speech recognition for speakers with cognitive disorders. In *Text, Speech, and Dialogue*, pages 501–512.

Larisa S. Tsvetkova. 1988. *Aphasia and rehabilitative learning*. Prosveshchenie, Moscow. In Russian.

Juan Camilo Vásquez-Correa, Juan Rafael Orozco-Arroyave, and Elmar Nöth. 2017. Convolutional neural network to model articulation impairments in patients with parkinson's disease. In *Proceedings of Interspeech 2017*, pages 314–318.

Lukas Wagner, Maximilian Zusag, and Tobias Bloder. 2023. Careful Whisper: Leveraging advances in automatic speech recognition for robust and interpretable aphasia subtype classification. *arXiv*. ArXiv:2308.01327.

Tatiana G. Wiesel. 2002. *Neurolinguistic Analysis of Atypical Forms of Aphasia (A Systemic Integrated Approach)*. Ph.D. thesis, Moscow. In Russian.

Irina P. Yastrebtseva, Evgeny A. Biryukov, Viktoria V. Belova, and Lyudmila Y. Deryabkina. 2023. Results of targeted training in the rehabilitation of patients with a combination of motor and speech disorders. *Bulletin of Rehabilitation Medicine*, 22(3):49–58. In Russian.