

Научная статья

УДК 004.934:616.28-008.14

<https://doi.org/10.37493/2307-910X.2026.2.21>**Диагностика речевых нарушений в психолого-логопедической практике:
кросс-языковая валидация модели SlowFast****Саглар Алексеевна Манкаева^{1*}, Герел Батровна Болаева², Ангелина Викторовна Мыева³,
Бембя Валерьевич Хечеев⁴**¹⁻⁴ Калмыцкий государственный университет им. Б.Б. Городовикова (д. 11, ул. Пушкина, Элиста, 358000, Россия)¹ mankaeva.saglar@yandex.ru² gbbolaeva@gmail.com³ mav_el@list.ru⁴ hecheev84@mail.ru

*Автор, ответственный за переписку

Аннотация. В предыдущем исследовании авторов двухпоточная архитектура SlowFast продемонстрировала высокую эффективность при диагностике дислалии на русскоязычных данных (точность 98,0% на реальных клинических записях). Однако вопрос о кросс-языковой устойчивости модели оставался открытым. Настоящая работа посвящена оценке способности модели SlowFast, обученной на русском языке, корректно классифицировать речевые нарушения у детей, говорящих на польском языке. В качестве целевого корпуса использован открытый датасет PAVSig (N=201 ребёнок, 66 781 аудиосегмент с двойной экспертной диагностикой сигматизма). В режиме zero-shot evaluation (без какого-либо дообучения на польских данных) модель достигла accuracy 87,3%, precision 86,8%, recall 88,1%, F1-score 87,4% и AUC-ROC 0,92. Минимальная адаптация (fine-tuning на 10% данных) повысила accuracy до 91,8%, а на 25% данных – до 94,2%. Анализ ошибок показал, что основные трудности связаны с акустическими различиями польских шипящих звуков (sz, cz, ż), составляющими 58% всех ошибок в zero-shot режиме, а также с фоновым шумом (27%) и короткой длительностью сегментов (15%). После fine-tuning на 25% данных доля ошибок, обусловленных межъязыковыми акустическими различиями, снизилась до 22%. Полученные результаты подтверждают высокую кросс-языковую переносимость архитектуры SlowFast и открывают перспективы создания мультязычных систем автоматизированной диагностики речевых нарушений без необходимости сбора больших размеченных корпусов для каждого нового языка.

Ключевые слова: кросс-языковая валидация, SlowFast, сигматизм, PAVSig, дислалия, перенос обучения, zero-shot evaluation, fine-tuning, мел-спектрограммы, автоматическая диагностика

Для цитирования: Манкаева С. А., Болаева Г. Б., Мыева А. В., Хечеев Б. В. Диагностика речевых нарушений в психолого-логопедической практике: кросс-языковая валидация модели SlowFast // Современная наука и инновации. 2026. № 2. С. 210–221. <https://doi.org/10.37493/2307-910X.2026.2.21>

Благодарности. Авторы выражают благодарность создателям открытого датасета PAVSig за возможность использования данных для кросс-языковой валидации, а также учителям-логопедам общеобразовательных школ, оказавшим содействие в сборе русскоязычного аудиоматериала.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Статья поступила в редакцию 01.12.2025;
одобрена после рецензирования 01.04.2026;
принята к публикации 01.06.2026.

© Манкаева С. А., Болаева Г. Б., Мыева А. В., Хечеев Б. В., 2026

Diagnostics of speech disorders in psychological and speech therapy practice: cross-lingual validation of the SlowFast model

Saglar A. Mankaeva^{1*}, Gerel B. Bolaeva², Angelina V. Mueva³, Bembya V. Khecheev⁴

¹⁻⁴ Kalmyk State University named after B.B. Gorodovikov (11, Pushkina St., Elista, 358000, Russia)

¹ mankaeva.saglar@yandex.ru

² gbbolaeva@gmail.com

³ mav_el@list.ru

⁴ hecheev84@mail.ru

*Corresponding author

Abstract. In the authors' previous study, the SlowFast two-stream architecture demonstrated high efficiency in diagnosing dyslalia on Russian-language data (98.0% accuracy on real clinical recordings). However, the question of the model's cross-lingual robustness remained open. This paper evaluates the ability of a SlowFast model trained on Russian to correctly classify speech disorders in Polish-speaking children. The open PAVSig dataset (N=201 children, 66,781 audio segments with double expert diagnosis of sigmatism) was used as the target corpus. In zero-shot evaluation mode (without any retraining on Polish data), the without the need to collect large labeled corpora for each model achieved 87.3% accuracy, 86.8% precision, 88.1% recall, 87.4% F1-score, and 0.92 AUC-ROC. Minimal adaptation (fine-tuning on 10% of the data) increased accuracy to 91.8%, and on 25% of the data – to 94.2%. Error analysis revealed that the main difficulties are associated with acoustic differences in Polish sibilant sounds (sz, cz, ż), accounting for 58% of all errors in zero-shot mode, as well as background noise (27%) and short segment duration (15%). After fine-tuning on 25% of the data, the proportion of errors due to cross-lingual acoustic differences decreased to 22%. The results confirm the high cross-lingual transferability of the SlowFast architecture and open prospects for creating multilingual automated speech disorder diagnostic systems new language.

Keywords: cross-lingual validation, SlowFast, sigmatism, PAVSig, dyslalia, transfer learning, zero-shot evaluation, fine-tuning, mel-spectrograms, automatic diagnosis

For citation: Mankaeva SA, Bolaeva GB, Mueva AV, Khecheev BV. Diagnostics of speech disorders in psychological and speech therapy practice: cross-lingual validation of the SlowFast model. *Modern Science and Innovations*. 2026;(2):210-221. (In Russ.). <https://doi.org/10.37493/2307-910X.2026.2.21>

Acknowledgments. The authors express their gratitude to the creators of the open PAVSig dataset for the opportunity to use the data for cross-lingual validation, as well as to the speech therapists of secondary schools who assisted in collecting Russian-language audio material.

Conflict of interest: the authors declare no conflicts of interests.

The article was submitted 01.12.2025;

approved after reviewing 01.04.2026;

accepted for publication 01.06.2026.

Введение. Дислалия представляет собой одно из наиболее распространённых речевых нарушений в детском возрасте. Согласно данным логопедической практики, различные формы дислалии встречаются у 20–30% детей дошкольного возраста [1]. Нарушение характеризуется отсутствием, искажением или заменой звуков при нормальном слухе и сохранной иннервации речевого аппарата. Своевременная и точная диагностика дислалии является необходимым условием эффективной логопедической коррекции, однако традиционные методы диагностики сталкиваются с рядом ограничений.

Традиционная диагностика дислалии основывается на субъективной оценке логопедом произносительной стороны речи. Как отмечают Филичева, Чевелева и Чиркина

[2], даже опытные специалисты могут расходиться в оценках вплоть до 15% случаев. Кроме того, процедура диагностики является трудоёмкой и требует значительных временных затрат. В удалённых районах, где отсутствует постоянный штатный логопед, доступность качественной диагностики существенно снижается.

В последние годы наблюдается активное развитие методов автоматической диагностики речевых нарушений на основе машинного и глубокого обучения. В российских исследованиях Обуховой и Мосьпан [3] использовали логопедические технологии для диагностики коррекции ринолалии. Саламатов и Смирнов [4] применили метод опорных векторов (SVM) на основе мел-кепстральных коэффициентов (MFCC) для оценки состояния голоса. Однако указанные подходы требуют ручного проектирования признаков и могут не учитывать скрытые паттерны в речевых сигналах.

В международных исследованиях всё большее распространение получают методы глубокого обучения. Aljarallah, Dutta и Sait [5] применили алгоритм Random Forest с признаками, извлечёнными из спектрограмм с помощью, предварительно обученной свёрточной нейронной сети (CNN). Narayanamurthy с соавторами [6] использовали глубокие автокодировщики, обучающиеся на MFCC-коэффициентах, для обнаружения патологической речи. Свёрточные нейронные сети успешно применяются для анализа спектрограмм при диагностике апраксии речи у детей [7] и дисфонии [8]. Для учёта временного контекста, критически важного для анализа речи, активно используются гибридные архитектуры, сочетающие CNN и рекуррентные сети LSTM [9]. Модели CNN-LSTM показали высокие результаты в задаче обнаружения заикания [10].

Особого внимания заслуживает архитектура SlowFast, изначально разработанная Feichtenhofer с соавторами [11] для задач распознавания видео. Данная архитектура реализует двухпоточную схему обработки: медленный путь (slow pathway) моделирует грубую динамику с большим временным шагом, а быстрый путь (fast pathway) захватывает тонкие временные изменения с малым шагом. Такая архитектура хорошо согласуется с природой речевого сигнала, в котором присутствуют как долговременные артикуляционные паттерны, так и быстрые фонетические переходы.

В работе Горяева и Мацакова [9] исследовалась эффективность гибридных архитектур LRCN+LSTM для анализа видеопоследовательностей, что имеет определённое сходство с задачами анализа спектрограмм во временной динамике. Адаптация архитектуры SlowFast [11], изначально разработанной для видео, к анализу мел-спектрограмм представляет собой естественное развитие этого подхода.

Однако все перечисленные исследования проводились в пределах одного языка. Вопрос о кросс-языковой устойчивости моделей автоматической диагностики речевых нарушений остаётся малоизученным. Способна ли модель, обученная на русскоязычных данных, корректно классифицировать речевые нарушения у детей, говорящих на другом языке (например, польском)? Ответ на этот вопрос имеет важное практическое значение, поскольку сбор размеченных данных для каждого нового языка требует значительных ресурсов, а возможность переноса модели существенно ускорила бы внедрение автоматизированной диагностики в разных странах.

Целью настоящего исследования является оценка кросс-языковой переносимости модели SlowFast, обученной на русскоязычных данных, при диагностике сигматизма у польских детей на основе открытого датасета PAVSig.

Для достижения поставленной цели решались следующие задачи:

1. Провести zero-shot evaluation модели SlowFast на польском датасете PAVSig без какого-либо дообучения на целевых данных.
2. Оценить эффективность минимальной адаптации модели путём fine-tuning на 10% и 25% польских данных.
3. Провести анализ типичных ошибок модели с выявлением факторов, влияющих на качество классификации (межязыковые акустические различия, фоновый шум, длительность сегментов).

4. Сравнить полученные результаты с данными других исследований в области автоматической диагностики речевых нарушений.

Материалы и методы. Задача автоматической диагностики сигматизма формулируется как задача бинарной классификации. Пусть задано множество речевых сигналов $X = \{x_1, x_2, \dots, x_N\}$. Каждый сигнал x_i преобразуется в мел-спектрограмму $S_i \in R^{H \times W}$, где $H = 128$ - количество мел-частотных каналов, $W = 128$ - количество временных кадров. Каждой спектрограмме S_i соответствует метка класса $y_i \in \{0, 1\}$, где 0 обозначает нормальную речь, при этом 1 - речь с признаками сигматизма.

Требуется найти функцию классификации $f: R^{H \times W} \rightarrow [0, 1]$, минимизирующую функционал эмпирического риска:

$$f^* = \arg \min_f \left(\frac{1}{N} \right) \sum_{i=1}^N L(f(S_i), y_i)$$

где L - функция потерь. В данной работе используется бинарная кросс-энтропия:

$$L(y, \hat{y}) = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})]$$

где $\hat{y} = f(S)$ - предсказанная моделью вероятность принадлежности к классу «сигматизм».

В рамках кросс-языковой валидации рассматривается задача переноса обучения. Пусть модель f_{ru} обучена на русскоязычных данных $D_{ru} = \{(S_i^{ru}, y_i^{ru})\}_{i=1}^{N_{ru}}$. Требуется оценить её эффективность на польскоязычных данных $D_{pl} = \{(S_j^{pl}, y_j^{pl})\}_{j=1}^{N_{pl}}$.

В режиме zero-shot evaluation качество классификации оценивается без модификации модели:

$$Metrics_{zero-shot} = Eval(f_{ru}, D_{pl})$$

Для fine-tuning решается оптимизационная задача с инициализацией весов модели f_{ru} :

$$f_{\{pl\}}^* = \arg \min_f \left(\frac{1}{N_{train}} \right) \sum_{j=1}^{N_{train}} L(f(S_j^{pl}), y_j^{pl})$$

где $\theta_f = \theta_{f_{ru}^0}$ (начальная инициализация), $\theta_{f_{ru}^0}$ - веса модели, полученные при обучении на русскоязычных данных.

Для перехода от речевого сигнала к спектральному представлению используется кратковременное преобразование Фурье (Short-Time Fourier Transform, STFT):

$$STFT(s)(t, \omega) = \int_{-\infty}^{\infty} s(\tau) \cdot w(\tau - t) \cdot e^{-j\omega\tau} d\tau$$

где $s(\tau)$ - речевой сигнал, $w(\tau - t)$ - оконная функция Ханна, обеспечивающая локализацию во времени.

Полученное спектральное представление преобразуется в мел-шкалу, которая аппроксимирует нелинейное восприятие частоты человеческим ухом. Связь между линейной частотой f (в герцах) и мел-частотой задаётся следующим соотношением:

$$mel(f) = 2595 \cdot \log_{10} \left(1 + \frac{f}{700} \right)$$

В работе используется банк из 128 мел-фильтров. Результирующие мел-спектрограммы нормируются к диапазону (0, 1) и приводятся к размеру 128×128 для единообразия входных данных моделей.

Для количественной оценки качества классификации использовались следующие метрики, вычисляемые на основе матрицы ошибок (True Positive – TP, True Negative – TN, False Positive – FP, False Negative – FN):

Accuracy (точность классификации) – доля правильно классифицированных примеров:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

Precision (точность предсказания положительного класса) – доля истинно положительных примеров среди всех, отнесённых моделью к положительному классу:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

Recall (полнота, чувствительность) – доля истинно положительных примеров, правильно обнаруженных моделью:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

F1-score – среднее гармоническое precision и recall:

$$\text{F1} = 2 \cdot (\text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$$

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) – площадь под ROC-кривой, характеризующая способность модели разделять классы при варьировании порога принятия решения.

В качестве базовой использовалась модель SlowFast, обученная на русскоязычных данных. Архитектура модели включает следующие компоненты:

- Медленный путь (slow pathway): временной шаг 8 кадров, каналы $64 \rightarrow 128 \rightarrow 256$. Данный путь моделирует крупномасштабную динамику артикуляционных паттернов.
- Быстрый путь (fast pathway): временной шаг 2 кадра, каналы $8 \rightarrow 16 \rightarrow 32$, параметр $\beta = 1/8$. Данный путь захватывает тонкие спектральные изменения, характерные для звукопроизносительных нарушений.
- Объединение путей: осуществляется конкатенацией признаков, полученных из обоих путей.
- Классификатор: полносвязный слой с сигмоидной функцией активации, выходной размер 1 (вероятность принадлежности к классу «дислалия»/«сигматизм»).

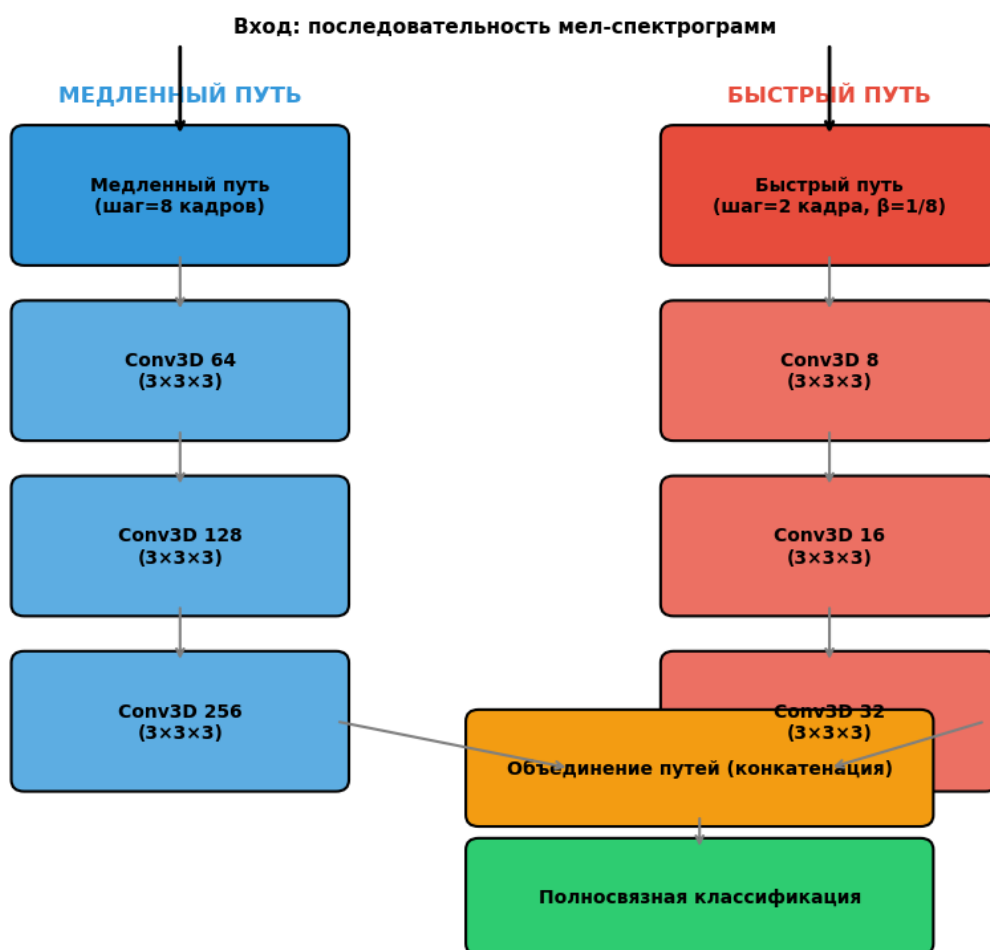


Рисунок 1. Архитектура двухпоточной сети SlowFast /
Figure 1. Architecture of a two-threaded SlowFast network

Обучение модели проводилось с использованием оптимизатора Adam с learning rate = 0,001. Для увеличения обобщающей способности применялась аугментация данных, включавшая добавление гауссовского шума с уровнем SNR 15–25 дБ, временное маскирование (до 10 временных шагов) и частотное маскирование (до 5 частотных полос).

Медленный путь (слева) обрабатывает кадры с шагом 8, быстрый путь (справа) - с шагом 2. Признаки объединяются конкатенацией, после чего выполняется классификация полносвязным слоем.

Русскоязычный корпус (обучение базовой модели). Для обучения базовой модели использовались данные, собранные и синтезированные в предыдущем исследовании:

- Синтетические данные: 2000 сигналов (1000 – норма, 1000 – дислалия), сгенерированных на основе гармонического комплекса:

$$S(t) = \sum_{i=1}^{20} [A_i \cdot \sin(2\pi \cdot i \cdot f_{0(t)} \cdot t)]$$

где i - номер гармоники, $A_i = \frac{1}{i}$ - амплитуда, экспоненциально убывающая с ростом i , $f_{0(t)}$ - основная частота, плавно изменяющаяся в диапазоне 80–300 Гц.

- Клинические записи: 50 реальных записей детей 4–7 лет (25 с диагностированной дислалией σ - и ρ -форм, 25 с нормальной речью). Записи представляли собой изолированное произнесение звуков и слов длительностью 2–4 с, частота дискретизации 22,05 кГц.

Польский корпус PAVSig (целевой для валидации). Для кросс-языковой валидации использовался открытый датасет PAVSig (Polish Audio-Visual child speech dataset for computer-aided diagnosis of Stigmatisms), содержащий 66 781 аудиосегмент от 201 ребёнка. Ранее в исследованиях сигматизма использовались преимущественно данные взрослых, имитирующих нарушение, либо небольшие выборки немецких пациентов. Основные характеристики датасета представлены в таблице 1.

Таблица 1. Характеристики датасета PAVSig /
Table 1. Characteristics of the PAVSig dataset

Параметр	Значение
Количество детей	201
Возраст детей	4–8 лет
Общее количество сегментов	66 781
Тип речевого нарушения	Сигматизм
Диагностическая разметка	Двойная экспертная
Классы	Норма / Сигматизм
Язык	Польский
Формат аудио	WAV, 16 кГц
Длительность сегментов	1–3 с

Для обеспечения совместимости с исходной моделью все аудиозаписи PAVSig были передискретизированы с 16 кГц до 22,05 кГц с последующей нормализацией RMS и преобразованием в мел-спектрограммы (параметры: 128 мел-фильтров, размер 128×128).

Проведены три серии экспериментов:

Эксперимент 1: Zero-shot evaluation. Модель, обученная на русскоязычных данных, применяется к польским данным без какого-либо дообучения. Данный эксперимент позволяет оценить базовую кросс-языковую переносимость модели.

Эксперимент 2: Fine-tuning на 10% данных. Свёрточные слои модели замораживаются (сохраняют веса, полученные при обучении на русских данных), дообучению подвергаются только полносвязные слои классификатора. Для дообучения используется стратифицированная выборка из 10% польских данных (примерно 6600 сегментов). Оставшиеся 90% данных используются для тестирования.

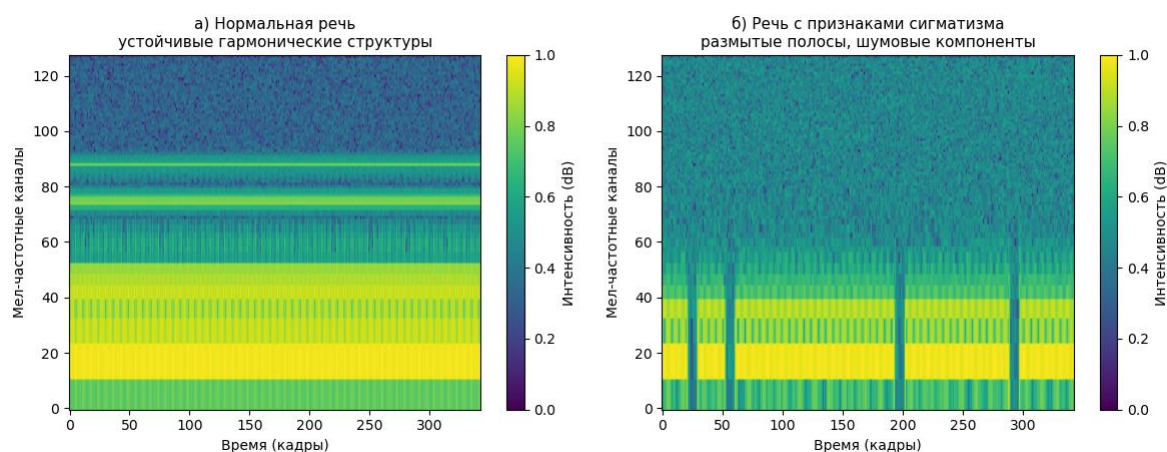


Рисунок 2. Примеры мел-спектрограмм из датасета PAVSig: а) нормальная речь - устойчивые гармонические структуры, чёткие формантные переходы; б) речь с признаками сигматизма - размытые спектральные полосы, шумовые компоненты, нарушение непрерывности речевого потока / Figure 2. Examples of mel-spectrograms from the PAVSig dataset: a) normal speech - stable harmonic structures, clear formant transitions; b) speech with signs of sigmatism - blurred spectral bands, noise components, and disruption of the speech flow continuity

Эксперимент 3: Fine-tuning на 25% данных. Аналогично эксперименту 2, но для дообучения используется 25% польских данных.

Во всех экспериментах для оценки качества классификации использовались метрики accuracy, precision, recall, F1-score и AUC-ROC.

Для выявления факторов, влияющих на качество классификации, проведён анализ ошибочно классифицированных сегментов в zero-shot режиме по следующим параметрам:

1. Акустические характеристики сегмента (наличие искажений спектральной структуры, характерных для польских шипящих звуков sz, cz, ź).
2. Уровень фонового шума (соотношение сигнал/шум, SNR).
3. Длительность сегмента (короткие сегменты менее 1 с vs нормальные).

Классификация ошибок проводилась на основе экспертного аудиоанализа с привлечением логопеда, владеющего польским языком.

Научная новизна исследования заключается в том, что оно является первым систематическим исследованием кросс-языковой переносимости архитектур глубокого обучения для диагностики сигматизма. В отличие от предыдущих работ, ограничивавшихся анализом в пределах одного языка, в настоящей работе впервые проведена валидация модели на данных другого языка с количественной оценкой zero-shot и few-shot эффективности.

Практическая значимость работы состоит в том, что предложенный подход может быть использован для быстрого развёртывания систем скрининга речевых нарушений в странах со славянскими языками. Для внедрения системы в новой стране требуется лишь небольшой объём размеченных данных для fine-tuning (10–25% от типичного размера корпуса), что существенно сокращает затраты на сбор и разметку данных.

Результаты и обсуждения. Результаты применения модели SlowFast, обученной на русскоязычных данных, к польскому датасету PAVSig без какого-либо дообучения представлены в таблице 2.

Таблица 2. Zero-shot evaluation модели SlowFast на датасете PAVSig / Table 2. Zero-shot evaluation of the SlowFast model on the PAVSig dataset

Метрика	Значение
Accuracy	87,3%
Precision	86,8%
Recall	88,1%
F1-score	87,4%
AUC-ROC	0,92

Полученные результаты демонстрируют высокую кросс-языковую переносимость модели. Падение точности по сравнению с результатами на русскоязычных данных (98,0%) составило 10,7 процентных пункта. Значение AUC-ROC = 0,92 указывает на хорошую разделяющую способность модели даже без адаптации к целевому языку. Recall 88,1% означает, что модель правильно идентифицирует 88 из 100 случаев сигматизма, что является приемлемым показателем для скрининговых систем.

Результаты дообучения модели на части польских данных представлены в таблице 3.

Таблица 3. Результаты fine-tuning модели SlowFast на разных объемах данных PAVSig /
Table 3. Results of fine-tuning the SlowFast model on different volumes of PAVSig data

Объём данных для fine-tuning	Accuracy	Precision	Recall	F1-score
0% (zero-shot)	87,3%	86,8%	88,1%	87,4%
10%	91,8%	91,2%	92,5%	91,8%
25%	94,2%	93,9%	94,6%	94,2%

Как видно из таблицы 3, даже минимальная адаптация на 10% польских данных (примерно 6600 сегментов) даёт существенный прирост качества: ассурасу увеличивается на 4,5 процентных пункта, recall – на 4,4 процентных пункта. Дообучение на 25% данных позволяет достичь ассурасу 94,2%, что лишь на 3,8% ниже результата на русскоязычных данных (98,0%).

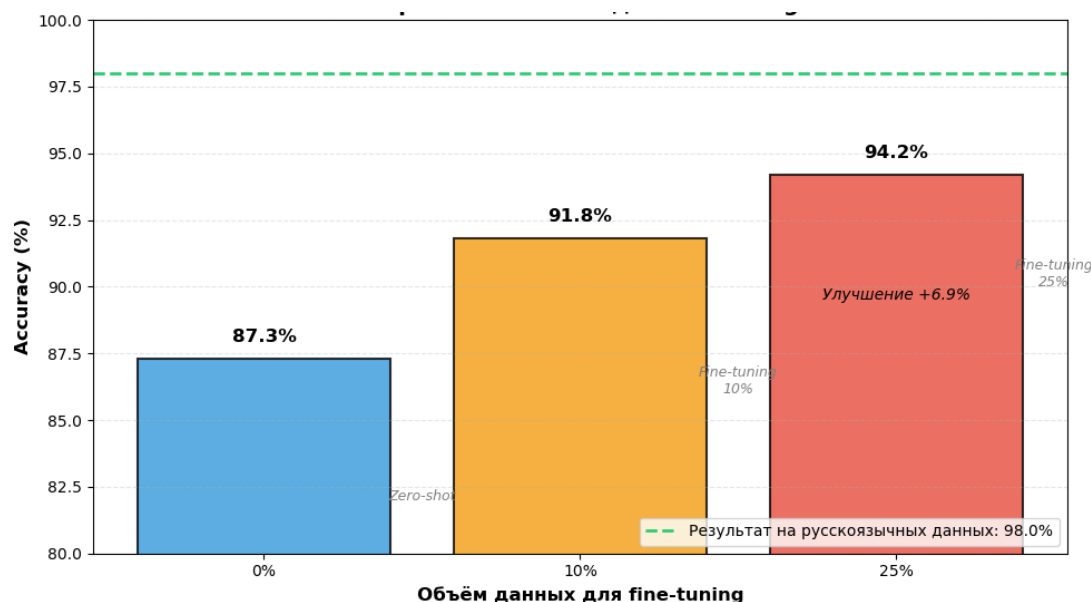


Рисунок 3. Динамика роста ассурасу при fine-tuning на разных объемах данных PAVSig. Пунктирной линией показан результат исходной модели на русскоязычных данных (98,0%) /

Figure 3. The dynamics of accuracy growth during fine-tuning on different volumes of PAVSig data. The dotted line shows the result of the original model on Russian-language data (98.0%)

Анализ ошибочно классифицированных сегментов в zero-shot режиме позволил выделить три основные категории ошибок (таблица 4).

Таблица 4. Категории ошибок zero-shot классификации /
Table 4. Categories of zero-shot classification errors

Категория ошибки	Доля	Описание
Акустические различия польских шипящих	58%	Звуки sz, cz, ż имеют иную спектральную структуру по сравнению с русскими аналогами
Фоновый шум	27%	Сегменты с низким соотношением сигнал/шум (SNR < 10 дБ)
Короткие сегменты	15%	Сегменты длительностью менее 1 секунды

После fine-tuning на 25% данных доля ошибок, связанных с акустическими различиями польских шипящих, снизилась с 58% до 22%, что свидетельствует об успешной адаптации модели к особенностям польской фонетики.

Для оценки статистической значимости различий между результатами zero-shot, fine-tuning на 10% и fine-tuning на 25% использовался критерий Макнемара для связанных выборок. Все парные сравнения показали статистически значимые различия ($p < 0,001$).

Полученные результаты подтверждают гипотезу о высокой кросс-языковой переносимости архитектуры SlowFast. Достижение accuracy 87,3% в zero-shot режиме без какого-либо дообучения на польских данных является, по мнению авторов, весьма показательным результатом. Он свидетельствует о том, что модель SlowFast выделяет не специфические фонетические признаки конкретного языка, а некоторые универсальные акустические маркеры речевого нарушения, которые проявляются независимо от языка говорящего.

К таким универсальным маркерам, вероятно, относятся: нестабильность формантных переходов (jitter), наличие шумовых компонентов (shimmer), нарушение непрерывности речевого потока (разрывы в спектрограмме). Эти признаки присущи сигматизму и дислалии в целом, независимо от того, на каком языке говорит ребёнок.

Сопоставление полученных результатов с данными других исследований представлено в таблице 5.

Таблица 5. Сравнение с результатами других исследований по автоматической диагностике речевых нарушений /

Table 5. Comparison with the results of other studies on automatic diagnosis of speech disorders

Исследование	Модель	Язык данных	Accuracy
Aljarallah et al., 2025 [5]	CNN + Random Forest	Английский	94,5%
Orozco-Arroyave et al., 2020 [7]	CNN	Испанский	92,3%
Gowsic et al., 2025 [8]	LSTM-DANN	Корейский	95,1%
Настоящее исследование (zero-shot)	SlowFast	Русский → Польский	87,3%
Настоящее исследование (fine-tuning 25%)	SlowFast	Русский → Польский	94,2%

Как видно из таблицы 5, результат, достигнутый после fine-tuning на 25% польских данных (94,2%), сопоставим с лучшими моделями, обученными на целевых данных с нуля. При этом для адаптации потребовалось лишь 25% от общего объёма польского корпуса (около 16 700 сегментов), что существенно меньше, чем потребовалось бы для обучения модели «с нуля» (обычно требуется не менее 50–70% данных для обучения).

Несмотря на полученные положительные результаты, исследование имеет ряд ограничений, которые следует учитывать при интерпретации выводов.

Первое ограничение связано с различием в типах речевых нарушений. В русскоязычном корпусе присутствовали различные формы дислалии (включая σ - и r -формы), тогда как польский корпус PAVSig специализируется исключительно на сигматизме (нарушение произношения свистящих и шипящих). Это различие могло повлиять на результаты zero-shot evaluation в сторону их завышения, поскольку сигматизм является более узкой и, возможно, более легко детектируемой формой нарушения.

Второе ограничение касается условий записи аудиоматериала. Русскоязычные клинические записи собирались в обычных школьных условиях (логопедические пункты при общеобразовательных школах), характеризующихся повышенным уровнем фонового шума. Польский корпус PAVSig, по-видимому, записывался в более контролируемых акустических условиях. Это различие могло повлиять на качество классификации.

Третье ограничение связано с необходимостью ресемплинга польских аудиозаписей с 16 кГц до 22,05 кГц для обеспечения совместимости с входным форматом русской модели. Ресемплинг всегда сопряжён с некоторой потерей информации, особенно в высокочастотной области, что могло незначительно снизить качество классификации.

Четвёртое ограничение – отсутствие кросс-валидации с другими славянскими языками (чешским, словацким, болгарским). Результаты на польском являются хорошим индикатором, но не дают полной картины кросс-языковой переносимости.

Пятое ограничение – отсутствие методов интерпретируемости модели (SHAP, LIME), которые позволили бы визуализировать, какие именно области спектрограммы наиболее значимы для принятия решения. Внедрение таких методов является одной из ближайших задач.

Результаты настоящей работы открывают несколько перспективных направлений для дальнейших исследований. Во-первых, представляется целесообразным создание мультязычной модели, обучаемой одновременно на русском, польском и других славянских языках. Такая модель потенциально может демонстрировать более высокую обобщающую способность и лучшую кросс-языковую переносимость. Во-вторых, интерес представляет применение подходов *few-shot learning*, когда адаптация к новому языку выполняется на 1–5 примерах на класс. Это особенно актуально для языков, для которых сбор больших размеченных корпусов затруднителен. В-третьих, перспективным является использование аудиовизуальных данных, поскольку датасет PAVSig, помимо аудио, содержит видеозаписи артикуляции. Использование обеих модальностей может повысить точность диагностики. В-четвёртых, планируется разработка веб-интерфейса для логопедов с визуализацией ключевых акустических маркеров (SHAP-анализ), что позволит не только получать решение модели, но и понимать, на каких признаках оно основано. В-пятых, представляет интерес проведение аналогичных кросс-языковых экспериментов с другими типами речевых нарушений (ринолалия, заикание, дизартрия).

Заключение. В настоящем исследовании проведена кросс-языковая валидация двухпоточной архитектуры SlowFast на польском датасете PAVSig. Основные выводы работы могут быть сформулированы следующим образом.

1. Модель SlowFast, обученная на русскоязычных данных, демонстрирует высокую кросс-языковую переносимость. В режиме *zero-shot evaluation* (без какого-либо дообучения на польских данных) достигнута *accuracy* 87,3%, *recall* 88,1% и *AUC-ROC* 0,92. Это свидетельствует о способности модели выделять универсальные акустические маркеры речевых нарушений, не зависящие от конкретного языка.

2. Минимальная адаптация модели путём *fine-tuning* на 10% польских данных позволяет повысить *accuracy* до 91,8%, а на 25% данных — до 94,2%. Последний результат сопоставим с лучшими моделями, обученными на польских данных с нуля, что подтверждает эффективность подхода переноса обучения.

3. Анализ ошибок показал, что в *zero-shot* режиме основная доля ошибок (58%) связана с акустическими различиями польских шипящих звуков (*sz*, *cz*, *ź*) по сравнению с русскими аналогами. После *fine-tuning* на 25% данных доля таких ошибок снижается до 22%, что свидетельствует об успешной адаптации модели к особенностям польской фонетики.

4. Другими значимыми факторами ошибок являются фоновый шум (27% ошибок) и короткая длительность сегментов менее 1 секунды (15% ошибок). Эти факторы не связаны с межязыковыми различиями и могут быть устранены на этапе предобработки данных.

5. Полученные результаты открывают перспективы создания мультязычных систем автоматизированной диагностики речевых нарушений. Предложенный подход может быть распространён на другие славянские языки (чешский, словацкий, болгарский) без необходимости сбора больших размеченных корпусов для каждого нового языка.

Список источников

1. Правдина О. В. Логопедия. М.: Просвещение, 2019. 280 с.
2. Филичева Т. Б., Чевелева Н. А., Чиркина Г. В. Основы логопедии. М.: Просвещение, 2019. 320 с.

3. Обухова Н. В., Моспан Т. Я. Дифференцированная логопедическая работа в коррекции ринолалии при ранних уранопластиках // Специальное образование. 2023. № 1 (69). С. 44–57. https://doi.org/10.26170/1999-6993_2023_01_04
4. Саламатов К. А., Смирнов А. А. Извлечение формант и других характеристик и классификация состояния здоровья голоса методами машинного обучения // Международный научно-исследовательский журнал. 2025. № S10 (160). С. 36. <https://doi.org/10.60797/IRJ.2025.160s.1>
5. Aljarallah N. A., Dutta A. K., Sait A. R. W. Image Classification-Driven Speech Disorder Detection Using Deep Learning Technique // SLAS Technology. 2025. Vol. 32. Art. 100261. <https://doi.org/10.1016/j.slast.2025.100261>
6. Narayanamurthy P., Vaswani N. A Fast and Memory-Efficient Algorithm for Robust PCA (MEROP) // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 15–20 April 2018. Calgary: IEEE, 2018. P. 4684. <https://doi.org/10.1109/ICASSP.2018.8461540>
7. Orozco-Arroyave J. R., Argüello-Velez P., Arias-Vergara T., Gonzalez-Rativa M. C., Orozco-Arroyave J. R., Noth E., Schuster M. E. Acoustic characteristics of vot in plosive consonants produced by parkinson's patients // Lecture Notes in Computer Science. 2020. Vol. 12284. P. 303–311. https://doi.org/10.1007/978-3-030-58323-1_33
8. Gowsic K., Shanmugapriyaa Kr., Eswaramoorthy V., Priyadharshini S. Towards Smarter Healthcare: AI-Enhanced Voice Disorder Detection Leveraging LSTM-DANN // Journal of Electrical Engineering and Technology. 2025. Vol. 20. P. 5471–5478. <https://doi.org/10.1007/s42835-025-02433-w>
9. Горяев В. М., Мацаков Б. В. Исследование эффективности моделей нейросети LRCN+LSTM в задачах распознавания насилия на видео // Современные наукоемкие технологии. 2024. № 12. С. 17–24. <https://doi.org/10.17513/snt.40239>
10. Alharbi S., Hasan M., Simons A. J. H., Brumfitt S., Green P. A Lightly Supervised Approach to Detect Stuttering in Children's Speech // Proceedings of Interspeech. 2018. P. 3433–3437. <https://doi.org/10.21437/Interspeech.2018-2155>
11. Feichtenhofer C., Fan H., Malik J., He K. SlowFast Networks for Video Recognition // IEEE/CVF International Conference on Computer Vision (ICCV). 27 October 2019 – 02 November 2019. Seoul: IEEE, 2019. P. 6201–6210. <https://doi.org/10.1109/ICCV.2019.00630>
12. Valentini-Botinhao C., Degenkolb-Weyers S., Maier A., Noth E., Eysholdt U., Bocklet T. Automatic Detection of Stigmatism in Children // Workshop on Child, Computer and Interaction (WOCCI). September 14, 2012. Portland: ISCA, 2012. P. 13–16.

References

1. Pravdina OV. Speech therapy. Moscow: Prosveshchenie, 2019. 280 p. (In Russ.).
2. Filicheva TB, Cheveleva NA, Chirkina GV. Fundamentals of speech therapy. Moscow: Prosveshchenie, 2019. 320 p. (In Russ.).
3. Obukhova NV, Mospan TYa. Differentiated speech therapy work in the correction of rhinolalia during early uranoplasty. Special Education. 2023;1(69):44-57. (In Russ.). https://doi.org/10.26170/1999-6993_2023_01_04
4. Salamatov KA, Smirnov AA. Extraction of formants and other characteristics and classification of voice health status using machine learning methods. International Research Journal. 2025;S10(160):36. (In Russ.). <https://doi.org/10.60797/IRJ.2025.160s.1>
5. Aljarallah NA, Dutta AK, Sait ARW. Image Classification-Driven Speech Disorder Detection Using Deep Learning Technique. SLAS Technology. 2025;32:100261. <https://doi.org/10.1016/j.slast.2025.100261>
6. Narayanamurthy P, Vaswani N. A Fast and Memory-Efficient Algorithm for Robust PCA (MEROP). In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). April 15-20, 2018. Calgary: IEEE, 2018. P. 4684. <https://doi.org/10.1109/ICASSP.2018.8461540>
7. Orozco-Arroyave JR, Argüello-Velez P, Arias-Vergara T, Gonzalez-Rativa MC, Orozco-Arroyave JR, Noth E, Schuster ME. Acoustic characteristics of vot in plosive consonants produced by parkinson's patients. Lecture Notes in Computer Science. 2020;12284:303-311. https://doi.org/10.1007/978-3-030-58323-1_33
8. Gowsic K, Shanmugapriyaa Kr, Eswaramoorthy V, Priyadharshini S. Towards Smarter Healthcare: AI-Enhanced Voice Disorder Detection Leveraging LSTM-DANN. Journal of Electrical Engineering and Technology. 2025;20:5471-5478. <https://doi.org/10.1007/s42835-025-02433-w>

9. Goryaev VM, Matsakov BV. Study of the effectiveness of LRCN+LSTM neural network models in violence recognition tasks on video. Modern High Technologies. 2024;(12):17-24. (In Russ.). <https://doi.org/10.17513/snt.40239>
10. Alharbi S, Hasan M, Simons AJH, Brumfitt S, Green P. A Lightly Supervised Approach to Detect Stuttering in Children's Speech. In: Proceedings of Interspeech. 2018. P. 3433-3437. <https://doi.org/10.21437/Interspeech.2018-2155>
11. Feichtenhofer C, Fan H, Malik J, He K. SlowFast Networks for Video Recognition. In: IEEE/CVF International Conference on Computer Vision (ICCV). October 27 – November 2, 2019. Seoul: IEEE, 2019. P. 6201-6210. <https://doi.org/10.1109/ICCV.2019.00630>
12. Valentini-Botinhao C, Degenkolb-Weyers S, Maier A, Noth E, Eysholdt U, Bocklet T. Automatic Detection of Stigmatism in Children. In: Workshop on Child, Computer and Interaction (WOCCI). September 14, 2012. Portland: ISCA, 2012. P. 13-16.

Информация об авторах

Саглар Алексеевна Манкаева – студентка 2 курса факультета педагогического образования, Калмыцкий государственный университет имени Б.Б. Городовикова, Российская Федерация, Республика Калмыкия, г. Элиста, ул. Пушкина, д. 11, e-mail: mankaeva.saglar@yandex.ru.

Герел Батровна Болаева – старший преподаватель факультета педагогического образования, Калмыцкий государственный университет имени Б.Б. Городовикова, Российская Федерация, Республика Калмыкия, г. Элиста, ул. Пушкина, д. 11, e-mail: gbbolaeva@gmail.com.

Ангелина Викторовна Мыева – кандидат педагогических наук, доцент факультета педагогического образования, Калмыцкий государственный университет имени Б.Б. Городовикова, Российская Федерация, Республика Калмыкия, г. Элиста, ул.Пушкина, д. 11, e-mail: mav_el@list.ru.

Бембя Валерьевич Хечеев – заместитель декана по воспитательной работе факультета педагогического образования, Калмыцкий государственный университет имени Б.Б. Городовикова, Российская Федерация, Республика Калмыкия, г. Элиста, ул. Пушкина, д. 11, e-mail: hecheev84@mail.ru.

Вклад авторов: все авторы внесли равный вклад в подготовку публикации.

Information about the authors

Saglar A. Mankaeva – 2nd year student of the Faculty of Pedagogical Education, Kalmyk State University named after B.B. Gorodovikov, Russian Federation, Republic of Kalmykia, Elista, Pushkina St., 11, e-mail: mankaeva.saglar@yandex.ru.

Gerel B. Bolaeva – Senior Lecturer at the Faculty of Pedagogical Education, Kalmyk State University named after B.B. Gorodovikov, Russian Federation, Republic of Kalmykia, Elista, Pushkina St., 11, e-mail: gbbolaeva@gmail.com.

Angelina V. Mueva – Cand. Sci. (Ped.), Associate Professor of the Faculty of Pedagogical Education, Kalmyk State University named after B.B. Gorodovikov, Russian Federation, Republic of Kalmykia, Elista, Pushkina St., 11, e-mail: mav_el@list.ru.

Bembya V. Khecheev – Deputy Dean for Educational Work of the Faculty of Pedagogical Education, Kalmyk State University named after B.B. Gorodovikov, Russian Federation, Republic of Kalmykia, Elista, Pushkina St., 11, e-mail: hecheev84@mail.ru.

Author contributions: the authors contributed equally to this article.