

Нейросетевая диагностика произношения на английском языке. Разработка веб-сервиса

М.А. Дорохин¹, С.А. Чернышев^{1✉}

¹ Санкт-Петербургский государственный университет аэрокосмического приборостроения,
Санкт-Петербург, 190000, Россия

Ссылка для цитирования

Дорохин М.А., Чернышев С.А. Нейросетевая диагностика произношения на английском языке. Разработка веб-сервиса // Программные продукты и системы. 2025. Т. 38. № 4. С. 724–732. doi: 10.15827/0236-235X.152.724-732

Информация о статье

Группа специальностей ВАК: 2.3.1, 2.3.5

Поступила в редакцию: 16.06.2025

После доработки: 24.06.2025

Принята к публикации: 30.06.2025

Аннотация. В статье рассматривается разработка веб-сервиса для коррекции произношения английского языка при помощи нейросетевого распознавания фонем. Предметом исследования является автоматическая оценка фонематической точности речи обучающегося и предоставление ему визуальной обратной связи о найденных ошибках. Для решения поставленной задачи использованы методы глубокого обучения: предварительно обученный ASR-модуль на базе Wav2Vec 2.0 преобразует аудиосигнал в последовательность фонем, а последующий классификатор сравнивает полученную транскрипцию с эталонной (генерируемой при помощи фонематической модели языка). При обнаружении несоответствий система выделяет неверные звуки цветом и предлагает пользователю повторить произношение. В качестве метода исследования применялись экспериментальные подходы к сбору специализированного набора данных, содержащего записи с ошибками произношения на уровне отдельных фонем. Для оценки качества модели использовались стандартные метрики точности и расстояние Левенштейна. Полученные результаты показали, что предложенная система достигает более 90 % точности распознавания корректных фонем. Практическая значимость работы проявляется в возможности интеграции сервиса в онлайн-курсы и приложения по изучению иностранных языков, а также в предоставлении индивидуальной обратной связи для улучшения качества речи учащихся.

Ключевые слова: интеллектуальная система, машинное обучение, ASR, MDD, CAPT, нейронные сети, Wav2Vec 2.0, фонемная разметка, расстояние Левенштейна

Благодарности. Авторы выражают признательность М.И. Иванову и В.К. Хованову за помощь в исследовании и реализации рассматриваемой в статье системы

Введение. Освоение правильного произношения является одной из ключевых задач при изучении иностранных языков. Некорректное произношение и фонетические ошибки могут значительно затруднить понимание и снизить качество коммуникации. Для решения этой проблемы используются системы компьютерного обучения произношению (*Computer-Assisted Pronunciation Training*, CAPT), основной задачей которых является анализ и выявление отклонений в речи пользователя, а также предоставление рекомендаций по их исправлению. Эффективная CAPT-система должна уметь распознавать ошибки в условиях разнообразия акцентов и шумов, чтобы провести точный анализ произношения.

Для реализации этого подхода необходимы эталонные данные с правильным произношением, на основе которых проводится сравнение, а также размеченные данные с примерами ошибок для обучения модели. Разработчики таких систем сталкиваются с рядом сложностей, например, с недостатком размеченных дан-

ных, необходимостью точного анализа минимальных различий между звуками.

Целью данной работы является создание инструмента, который не только фиксирует ошибки в произношении, но и предоставляет пользователю обратную связь. Это требует разработки модели, способной адаптироваться к фонетическим особенностям речи, обеспечивая ее точное распознавание.

В приложениях и системах для обучения произношению на иностранных языках используют различные технологии автоматического распознавания речи и анализа произношения. Среди популярных систем можно выделить как образовательные платформы общего назначения, так и специализированные приложения для анализа акцентов и фонетических ошибок. Подробное сравнение популярных CAPT-систем приведено в таблице 1.

Одним из ранних примеров подобных решений является система Tell Me More от компании Auralog. Это образовательная платформа, которая использует адаптивные технологии

Таблица 1

Сравнительный анализ аналоговых систем CAPT

Table 1

Comparative analysis of existing CAPT systems

Система	Функции	Недостатки	Целевая аудитория
Tell Me More	Адаптивные курсы, фонетическая диагностика, аудирование и повторение фраз	Высокая стоимость, отсутствие пробного периода, ограниченная доступность для частных лиц	Корпоративные клиенты, учебные заведения
Duolingo	Задания на лексику и произношение	Поверхностный анализ речи, отсутствие диагностики фонемных ошибок, игнорирование акцентов пользователей	Широкий круг пользователей
Lingualeo	Задания на лексику, грамматика, аудирование, ограниченный анализ речи	Отсутствие диагностики фонем, детальной обратной связи, оценка произношения доступна только в платной версии	Широкий круг пользователей

для обучения произношению, уделяя особое внимание аудированию и повторению слов и фраз. Однако программа имеет достаточно высокую цену и не предоставляет пробного периода, что делает ее доступной только для корпоративных клиентов или образовательных учреждений, ограничивая возможности использования индивидуальными пользователями.

Duolingo – одно из самых известных приложений для изучения языков, включающее функции анализа произношения. Основное внимание уделяется формированию словарного запаса и практике произношения с помощью коротких заданий. Однако анализ произношения не является основным направлением Duolingo, он ограничен общим уровнем без детального разбора ошибок на уровне фонем. Система также не учитывает индивидуальные особенности речи пользователя, такие как акценты или фонетические барьеры, что делает ее менее эффективной для иностранных пользователей, желающих проверить свое произношение.

Платформа Lingualeo предоставляет пользователям интерактивные инструменты для освоения лексики, грамматики, аудирования и практики речи. На ней также есть функция анализа устной речи, однако приложение не распознает ошибки на уровне фонем, не предоставляет детальной диагностики произношения и персонализированной обратной связи. Помимо этого, полноценная оценка произношения доступна только в платной версии приложения, что тоже может ограничить ее доступность для большинства пользователей.

Таким образом, большинство из современных CAPT-систем не предоставляют полного набора функций для задач, связанных с фонетической диагностикой. Это ограничивает их применение в тех случаях, когда учащийся стремится к более глубокой работе над произношением.

Современные решения на основе нейронных сетей, такие как Wav2Vec 2.0 [1], значительно повысили точность систем распознавания речи и открыли новые возможности для анализа и обучения произношения. Эти модели основаны на глубоких нейронных сетях, обучающихся на больших объемах данных, что позволяет достигать высоких результатов даже в сложных акустических условиях, включая шумы и различные акценты. Особый интерес представляет развитие методов самообучения для распознавания фонем в аудиозаписях. Эти подходы значительно превосходят традиционные методы благодаря возможности тонкой настройки предварительно обученных моделей на небольших наборах данных, которых обычно недостаточно для классических методов обучения с учителем.

Одним из распространенных критериев для оценки качества моделей является показатель частоты ошибок в символах (*Character Error Rate, CER*), который измеряет точность распознавания фонем. Метрика CER вычисляется как отношение количества ошибок (замен, вставок, удалений) к общему числу фонем в эталонной транскрипции и принимает значения от 0 % (идеальное распознавание) до 100 % (полная ошибка). Метрика CER стала стандартом в ис-

следованиях речи после выхода работы [2], где она впервые использовалась для оценки скрытых марковских моделей.

На датасете Common Voice этот показатель для различных моделей выглядит следующим образом [3]:

- Wav2Vec 2.0 Base – 22.70;
- Hubert Large – 17.84;
- WavLM Base – 20.55.

Каждое из этих значений отражает особенности настройки, подготовки данных и методологии извлечения фонем, что позволяет оптимизировать модель под конкретные задачи.

На датасете TIMIT, который является одним из эталонов для оценки качества распознавания фонем, лучшие результаты по состоянию на конец 2020 года были достигнуты следующими моделями [4]:

- CNN + TD-filterbanks – 18.0;
- PASE – 17.2;
- Li-GRU + fMLLR – 14.9;
- wav2vec – 14.7;
- vq-wav2vec – 11.6.

Дообучение модели Wav2Vec 2.0 на датасете TIMIT позволяет достичь высокого результата, например, CER = 8 %. Демонстрация была проведена с использованием модели wav2vec2-large-xlsr, что говорит о ее эффективности [5].

Одним из ключевых преимуществ моделей на основе Wav2Vec 2.0 и методов unsupervised learning является возможность использования стандартных датасетов для автоматического распознавания речи, доразмеченных с помощью фонемайзеров. Это значительно упрощает процесс обучения и позволяет работать с большими объемами данных, что выгодно отличает такие подходы от методов, требующих специализированных и часто труднодоступных датасетов.

Существуют датасеты, аннотированные вручную с использованием продвинутых инструментов, например Praat, и содержащие различные виды ошибок. Однако доступ к этим наборам данных обычно закрыт. Например, датасет Malmberg, собранный логопедической компанией, содержит 2,5 тысячи записей детей, читающих короткие рассказы вслух. Данные также аннотированы с помощью программы Praat, но доступа к ним нет.

Датасеты, созданные для задачи автоматического распознавания речи такие, как LibriSpeech [6] или LJ Speech [7], обычно основаны на записях чистой речи, например аудиокниг. Эти наборы данных не подходят для задачи вы-

явления и исправления ошибок в произношении, поскольку не содержат их, а также не имеют фонемных транскрипций.

L2-ARCTIC [8] – набор данных, содержащий записи речи говорящих, для которых английский язык является вторым (L2). Он создан специально для изучения акцентов, акцентных сдвигов и для исправления произношения. Ошибки произношения в нем аннотированы, а транскрипции выполнены в алфавите ARPAbet. Однако из-за своей специфики он может оказаться недостаточным для решения задачи. Например, если CAPT-системой будет использоваться носитель языка.

TIMIT [9] представляет собой набор данных с записями речи носителей английского языка, представляющих различные диалекты. Хотя набор включает фонемные транскрипции в формате TIMIT ARPAbet, ошибки произношения в нем отсутствуют. Особенность алфавита для фонем также затрудняет работу с этим датасетом.

Common Voice [10] – крупномасштабный краудсорсинговый проект от Mozilla, созданный для поддержки технологий распознавания речи с открытым исходным кодом. Этот датасет включает множество ошибок, так как записи выполняются обычными пользователями, а фонемные транскрипции отсутствуют.

Сравнительный анализ этих наборов данных приведен в таблице 2.

Таким образом, в открытом доступе сегодня не существует набора данных, который полностью удовлетворял бы всем требованиям задачи выявления и исправления ошибок в произношении.

Это позволяет сформировать основные требования к таким датасетам:

- наличие аудиозаписей с ошибками в произношении, чтобы модель не училась исключительно на идеальной речи;
- размещение фонемных транскрипций, аннотированных вручную и фиксирующих ошибки произношения;
- соответствие транскрипций определенному фонетическому алфавиту;
- включение чистых данных, не содержащих ошибок, чтобы модель могла эффективно работать с грамотно произнесенной речью.

Кроме того, транскрипции должны быть разделены по словам, а не представлять собой непрерывный поток фонем без деления на лексические единицы.

Датасеты TIMIT и L2-ARCTIC содержат фонемные транскрипции, а также временные

метки начала и конца произнесения каждой фонемы, что позволило восстановить разбиение фонемного потока на слова. Оба датасета используют схожие алфавиты формата ARPAbet, которые были преобразованы в формат IPA. Для этого вручную были созданы два маппинга – один для преобразования TIMIT, другой для L2-ARCTIC.

Для создания маппинга использовался CMU Dictionary [11], который задает правила использования ARPAbet в английских словах. После преобразования был сформирован финальный алфавит: a, b, d, e, f, h, i, k, l, m, n, o, p, s, t, u, v, w, z, æ, ð, ɲ, ɑ, ɔ, ə, ɚ, ε, g, ɪ, ɪ, ʃ, ʊ, ʌ, ʒ, θ, ts, tj, dʒ, dz.

Набор данных Common Voice не содержит фонемных транскрипций. Для его разметки была использована модель wav2vec2-large-xlsr-53-l2-arctic-phoneme [12], обученная на L2-ARCTIC. Она выдает транскрипции в формате IPA.

Датасеты LJ Speech и LibriSpeech представляют записи идеальной речи без ошибок, а также не содержат фонемных транскрипций. Для их разметки использовалась библиотека gruut [13], которая преобразует текст в IPA-формат. Транскрипции были приведены к финальному алфавиту.

Полученные датасеты объединены таким образом, чтобы максимально обеспечить равноценный вклад данных и чтобы в дальнейшем обученная модель могла одинаково хорошо работать как с грамотной речью, так и с речью с ошибками. Состав созданного датасета приведен в таблице 3.

Общее время датасета составило 16.8 часов, из которых:

- 8.7 часов датасетов с ошибками;
- 8.14 часов датасетов без ошибок.

Общий объем аудиофайлов составил 15 112 записей различной длительности, где каждая имеет следующие атрибуты:

- audio_path: путь к wav-файлу (16 кГц);
- sentence: исходное предложение на английском языке;
- phonemes: список фонемных транскрипций для каждого слова;
- duration: длительность аудиозаписи в секундах;
- dataset: источник данных (TIMIT, CV, LS, LJ, L2).

В качестве модели для дообучения использовалась facebook/wav2vec2-large-xlsr-53 [14], предварительно обученная на датасетах CommonVoice, BABEL [15], LibriSpeech версии Wav2Vec 2.0.

Таблица 2

Сравнительный анализ наборов данных

Table 2

Comparative analysis of datasets

Название	Фонемный алфавит	Наличие ошибок произношения	Наличие разбиения на слова	Длительность (час)
TIMIT	TIMIT ARPAbet	–	+	5.4
L2-ARCTIC	ARPAbet	+	+	27.1
Common Voice	–	+	–	312
LJ Speech	–	–	–	24
LibriSpeech	–	–	–	1 000

Таблица 3

Итоговый состав объединенного датасета

Table 3

Final composition of the joint dataset

Название	Разметка	Наличие ошибок произношения	Процент в общем датасете	Длительность (час)
TIMIT	IPA	–	32.14	5.4
L2-ARCTIC	IPA	+	35.71	6
Common Voice	IPA, wav2vec2	+	16.07	2.7
LJ Speech	IPA, gruut	–	8.04	1.37
LibriSpeech	IPA, gruut	–	8.04	1.37

Архитектура Wav2Vec 2.0 состоит из трех базовых компонентов – энкодера признаков, трансформера, модуля квантизации (рис. 1).

Энкодер признаков принимает необработанный аудиосигнал (с частотой дискретизации 16 кГц) и преобразует его в последовательность векторов признаков. Он стандартизируется, затем проходит через семь блоков одномерных сверточных слоев с постоянным числом каналов 512, нормализацией LayerNorm и функцией активации GELU. Представим параметры свертки. Первый слой: kernel_size = 10, stride = 5; слои 2–5: kernel_size = 3, stride = 2; слои 6–7: kernel_size = 2, stride = 2. Каждый блок уменьшает временную размерность сигнала, выходные векторы имеют размерность 512.

Блок, состоящий из 24 трансформеров, обрабатывает векторы признаков, формируя контекстные представления. Каждый трансформер имеет hidden size, self-attention и feed-forward (1024 → 4096 → 1024) с GELU, Dropout и нормализацией LayerNorm. В отличие от классических трансформеров [16], в Wav2Vec 2.0 используются относительные позиционные эмбединги на основе сверточных слоев.

Модуль квантизации преобразует непрерывные выходы трансформера в дискретные представления с помощью метода Gumbel-Softmax.

Самообучение основано на двух ключевых этапах: предобучение и адаптация к новой задаче. Данный подход, известный как трансферное обучение, позволяет эффективно переис-

пользовать знания, полученные на больших наборах данных, для решения более узких, связанных задач.

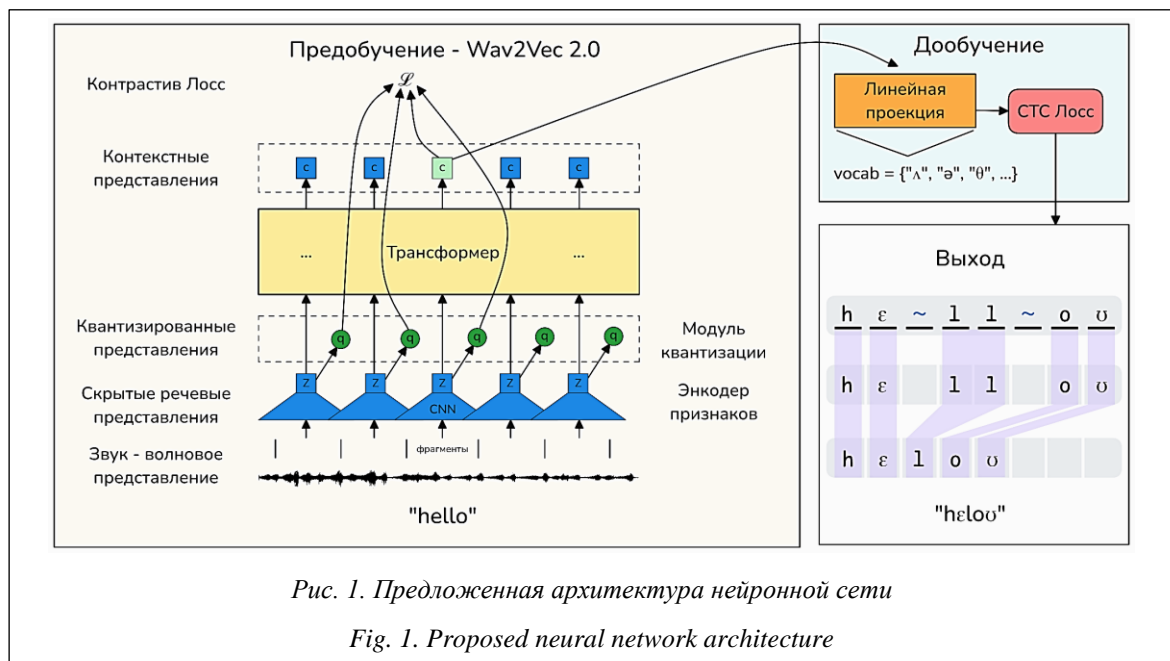
Общая цель предобучения – минимизировать комбинированную функцию потерь $L = L_m + \alpha L_d$, состоящую из контрастивной функции потерь L_m , отвечающей за производительность в задаче самообучения, и функции потерь разнообразия L_d , которая стимулирует использование всех доступных кодовых книг в модуле квантизации.

Для адаптации предобученной модели к задаче распознавания фонем поверх выходов трансформера был добавлен классификационный слой. Этот слой состоит из одного линейного проекционного слоя, который преобразует 1024-мерные векторы признаков, выдаваемые энкодером Wav2Vec 2.0, в логиты для каждой из 39 фонем целевого алфавита.

Обучение этого слоя и адаптация всей модели производились с помощью функции потерь CTC (Connectionist Temporal Classification).

Используемая модель обучалась на большом мультиязычном наборе данных для решения задачи автоматического распознавания речи. На этапе адаптации к новой задаче модель была дообучена на подготовленном ранее наборе данных с целью получения фонемных транскрипций из аудиозаписей человеческой речи разного уровня правильности.

В процессе обучения использовалась стратегия динамического управления вычислительными ресурсами и параметрами оптимизации:



- оптимизатор AdamW с коэффициентами затухания $\beta_1 = 0.9$ и $\beta_2 = 0.999$;
- линейный планировщик скорости обучения с фазой прогрева T_{warmup} длительностью 1 000 шагов, во время которой скорость обучения плавно нарастала от 0 до $\eta_{max} = 3 \times 10^{-4}$, после фазы прогрева следовало линейное затухание скорости до нуля к концу обучения;
- динамический размер батча от 8 до 16;
- точность вычислений: 16-битная арифметика с автоматическим масштабированием градиентов для ускорения вычислений;
- группировка аудиозаписей в батчах по схожей длительности для минимизации избыточного заполнения.

Валидация и сохранение контрольных точек модели производились каждые 500 шагов. После завершения процесса обучения модель была протестирована на отложенной выборке для оценки ее итоговой производительности.

На рисунке 2 приведен график изменения функции потерь и метрик качества на валидации за цикл обучения с 14 эпохами.

После обучения модель была оценена на валидационном наборе данных. Значение CER = 0.06 (6 %), что свидетельствует о высокой точности модели в задаче распознавания фонем.

Далее была выполнена оценка работы модели на отложенной тестовой выборке (табл. 4).

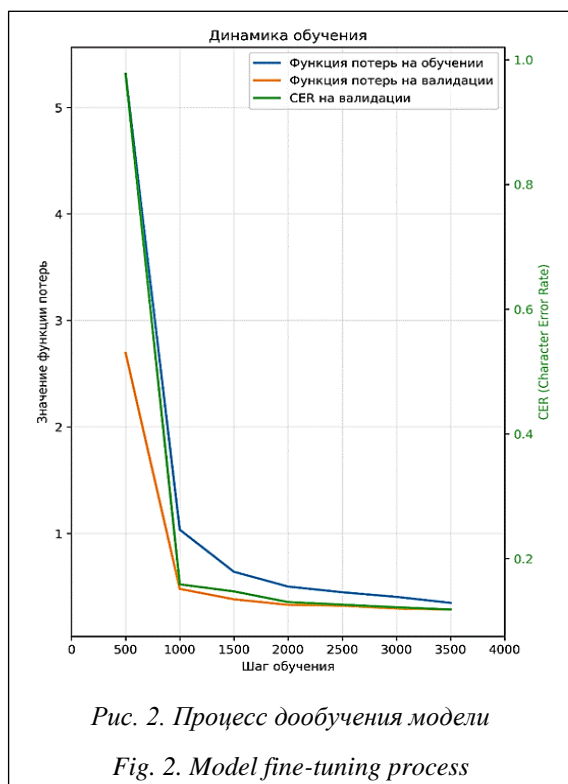


Рис. 2. Процесс дообучения модели

Fig. 2. Model fine-tuning process

Показатели тестовой выборки
Table 4
Test sample numbers

Метрика	Значение, %
CER	2.77
WER	9.29
Accuracy	90.09
Precision	91.07
Recall	90.94
F1	91.00
Замены	8.69
Удаления	0.37
Вставки	0.23
CER на гласных	4.01
CER на согласных	1.84

В контексте данной задачи accuracy (точность) определяется как 1-WER и составляет 91 %. Этот показатель интуитивно отражает долю правильно распознанной информации в общем потоке данных. WER в фонетическом пространстве – доля ошибочных слов, рассчитанная на фонемных представлениях.

Дальнейший анализ показал, что уровень ошибок при распознавании гласных фонем (CER ≈ 4.01 %) выше, чем для согласных (CER ≈ 1.84 %). Это ожидаемый результат, связанный с большей вариативностью артикуляции гласных в зависимости от акцента говорящего. Анализ наиболее частых замен подтвердил эту гипотезу: модель в основном путает фонетически близкие пары, такие как нейтральные гласные /ə/ и /ʌ/.

Преобладание ошибок замены является типичным для систем, использующих технологии автоматического распознавания речи, и указывает на то, что модель корректно сегментирует речь, но иногда путает акустически схожие звуки.

Тестирование производительности проводилось без графического ускорителя. Обработка идет в 6.62 раз быстрее реального времени (60 386.73 с аудио обработано за 9 126.16 с), то есть 1 час аудио ≈ 9.1 мин. Этого достаточно для интерактивных веб-сервисов.

В совокупности полученные результаты демонстрируют высокое качество разработанной модели. Она успешно решает поставленную задачу распознавания фонем в речи с различными акцентами, что открывает широкие возможности для ее практического применения в обучающих системах.

В рамках решения задачи был разработан веб-сервис, который предоставляет пользова-

телям возможность оценки произношения через выполнение упражнений и позволяет сравнить фонемную транскрипцию своей речи с эталонной. Для удобства восприятия пользователю предоставляется информация (выделяется цветом) о том, в каком звуке была допущена ошибка. Эталонная транскрипция формируется с помощью библиотеки `gruut`, которая автоматически преобразует текст в последовательность фонем в формате IPA.

Система выдает как визуальную, так и численную оценку качества произношения. Она рассчитывается на основе сопоставления последовательностей фонем и выражается в виде процента правильности, где 100 % соответствует идеальному совпадению (<http://www.swsys.ru/uploaded/image/2025-4/16.jpg>).

Для расчета различий между пользовательской и эталонной транскрипциями применяется алгоритм Вагнера – Фишера для нахождения расстояния Левенштейна [17]. Численная оценка качества произношения, представленная пользователю, является прямым производным от этой метрики и выражается в процентах.

Помимо анализа ошибок, пользователю предоставляется возможность прослушать пример идеального произношения предложения. Для этого используется функционал Web Speech API, поддерживаемый большинством современных веб-браузеров и позволяющий воспроизводить эталонное произношение, что помогает пользователю понять, как именно должны звучать слова и отдельные фонемы.

В ходе данной работы была успешно дообучена модель Wav2Vec 2.0 для задачи распознавания фонем в речи. Формирование собственного набора данных и расчет метрик, отражающих разные аспекты работы модели, позволили добиться высокой точности распознавания речи и эффективности анализа фонетической транскрипции.

Для обеспечения воспроизводимости результатов и участия научного сообщества дообученная модель была опубликована на платформе Hugging Face [18]. Следует отметить, что пользователи уже скачивают приложение, что подтверждает интерес к решениям на базе Wav2Vec 2.0 для задач фонетического анализа.

Внедрение созданной модели в веб-сервис позволило сделать процесс обучения произношению более эффективным и понятным для пользователя. Разработанная система демонстрирует высокую точность и способна предоставлять пользователям детализированную обратную связь.

Проведенное исследование показало, что современные технологии автоматического распознавания речи обладают большим потенциалом не только для перевода речи в текст, но и для анализа фонетических признаков на уровне фонем.

Перспективы дальнейшего развития проекта включают расширение функционала системы, интеграцию с другими инструментами обучения и улучшение алгоритмов обратной связи для более детального анализа ошибок произношения.

Список литературы

1. Baeviski A., Zhou H., Mohamed A., Auli M. wav2vec 2.0: A framework for self-supervised learning of speech representations. ArXiv, 2020, art. 2006.11477. URL: <https://arxiv.org/abs/2006.11477> (дата обращения: 18.06.2024).
2. Lee K.-F. Automatic Speech Recognition: The Development of the Sphinx System. Kluwer Academic Publ., 1989, 309 p. doi: 10.1007/978-1-4615-3650-5.
3. Apavou C., Belkada Y., Leo T., Arthur Z. Self-supervised pretraining for phoneme recognition, and generalization on foreign languages. MVA, 2022, pp. 1–9. URL: https://github.com/ASR-project/Multilingual-PR/blob/main/ASR_report_project_Apavou_Belkada_Tronchon_Zucker.pdf (дата обращения: 5.02.2024).
4. Ravanelli M., Brakel P., Omologo M., Bengio Y. Light gated recurrent units for speech recognition. IEEE Transactions on Emerging Topics in Computational Intelligence, 2018, vol. 2, no. 2, pp. 92–102. doi: 10.1109/TETCI.2017.2762739.
5. Vitouphy. Hugging Face URL: <https://huggingface.co/vitouphy/wav2vec2-xls-r-300m-timit-phoneme> (дата обращения: 18.06.2024).
6. Panayotov V., Chen G., Povey D., Khudanpur S. Librispeech: An ASR corpus based on public domain audio books. Proc. ICASSP, 2015, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.
7. The LJ Speech Dataset. URL: <https://keithito.com/LJ-Speech-Dataset/> (дата обращения: 18.06.2024).
8. Zhao G., Sonsaat S., Silpachai A. et al. L2-ARCTIC: A non-native english speech corpus. Interspeech, 2018, pp. 2783–2787. doi: 10.21437/Interspeech.2018-1110.
9. Garofolo J.S., Lamel L.F., Fisher W.M. et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Gaithersburg, 1992. 95.
10. Rosana A., Branson M., Davis K. et al. Common voice: A massively-multilingual speech corpus. Proc. Twelfth Language Resources and Evaluation Conf., 2020, pp. 4218–4222.
11. The CMU Pronouncing Dictionary. URL: <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> (дата обращения: 18.06.2024).

12. Mrrubino. Hugging Face. URL: <https://huggingface.co/mrrubino/wav2vec2-large-xlsr-53-l2-arctic-phoneme> (дата обращения: 18.06.2024).
13. Gruut Documentation. URL: <https://rhasspy.github.io/gruut/> (дата обращения: 18.06.2024).
14. Conneau A., Baevski A., Collobert R. et al. Unsupervised cross-lingual representation learning for speech recognition. *ArXiv*, 2020, art. 2006.13979v2. URL: <https://arxiv.org/abs/2006.13979> (дата обращения: 18.06.2024).
15. LDC Catalog by Year. URL: <https://catalog.ldc.upenn.edu/byyear> (дата обращения: 18.06.2024).
16. Vaswani A., Shazeer N., Parmar N. et al. Attention is all you need. *ArXiv*, 2017, art. 1706.03762v5. URL: <https://arxiv.org/pdf/1706.03762v5> (дата обращения: 18.06.2024).
17. Kobzdeja P., Waligóraa D., Wielebińska K., Paprzycki M. Parallel application of levenshtein distance to establish similarity between strings. *Frontiers in Artificial Intelligence and Applications*, 2003, vol. 12, no. 4.
18. Мохеееее. Hugging Face. URL: <https://huggingface.co/moxeeeee/wav2vec2-finetuned-pronunciation-correction> (дата обращения: 18.06.2024).

Software & Systems

doi: 10.15827/0236-235X.152.724-732

2025, 38(4), pp. 724–732

Neural network-based pronunciation diagnosis for english language. Web service developmentMikhail A. Dorokhin¹, Stanislav A. Chernyshev^{1✉}¹ Saint Petersburg State University of Aerospace Instrumentation, St. Petersburg, 190000, Russian Federation**For citation**Dorokhin, M.A., Chernyshev, S.A. (2025) 'Neural network-based pronunciation diagnosis for english language. Web service development', *Software & Systems*, 38(4), pp. 724–732 (in Russ.). doi: 10.15827/0236-235X.152.724-732**Article info**

Received: 16.06.2025

After revision: 24.06.2025

Accepted: 30.06.2025

Abstract. This paper presents the development of a web service designed to improve English pronunciation through neural network-based phoneme recognition. The subject of the research is the automatic assessment of a learner's phonemic accuracy and provision of visual feedback about detected pronunciation errors. The solution employs deep learning methods: a pretrained Wav2Vec 2.0-based ASR module converts audio signals into phoneme sequences, while a subsequent classifier compares the resulting transcription with a reference (generated using a language phonemic model). The system identifies pronunciation errors through color-coded highlighting and initiates user repetition exercises for correction. As a research method, the authors employ experimental approaches to collect a specialized dataset containing recordings with pronunciation errors at the individual phoneme level. To evaluate model quality, the authors use standard accuracy metrics and Levenshtein distance. The obtained results demonstrate that the proposed system achieves over 90% accuracy in recognizing correct phonemes. The practical significance of this work lies in the potential integration of the service into online language-learning platforms and mobile applications, as well as its ability to provide personalized feedback to help learners improve their speech quality.

Keywords: intelligent system, machine learning, ASR, MDD, CAPT, neural networks, Wav2Vec 2.0, phonemic annotation, Levenshtein distance

Acknowledgements. The authors are grateful to M.I. Ivanov and V.K. Khovanov for their help in researching and implementing the system discussed in the paper

References

1. Baevski, A., Zhou, H., Mohamed, A., Auli, M. (2020) 'wav2vec 2.0: A framework for self-supervised learning of speech representations', *ArXiv*, art. 2006.11477, available at: <https://arxiv.org/abs/2006.11477> (accessed June 18, 2024).
2. Lee, K.-F. (1989) *Automatic Speech Recognition: The Development of the Sphinx System*. Kluwer Academic Publ., 309 p. doi: 10.1007/978-1-4615-3650-5.
3. Apavou, C., Belkada, Y., Leo, T., Arthur, Z. (2022) 'Self-supervised pretraining for phoneme recognition, and generalization on foreign languages', *MVA*, pp. 1–9, available at: https://github.com/ASR-project/Multilingual-PR/blob/main/ASR_report_project_Apavou_Belkada_Tronchon_Zucker.pdf (accessed February 5, 2024).
4. Ravanelli, M., Brakel, P., Omologo, M., Bengio, Y. (2018) 'Light gated recurrent units for speech recognition', *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2), pp. 92–102. doi: 10.1109/TETCI.2017.2762739.
5. Vitouphy. *Hugging Face*, available at: <https://huggingface.co/vitouphy/wav2vec2-xls-r-300m-timit-phoneme> (accessed June 18, 2024).

6. Panayotov, V., Chen, G., Povey, D., Khudanpur, S. (2015) 'Librispeech: An ASR corpus based on public domain audio books', *Proc. ICASSP*, pp. 5206–5210. doi: 10.1109/ICASSP.2015.7178964.
7. *The LJ Speech Dataset*, available at: <https://keithito.com/LJ-Speech-Dataset/> (accessed June 18, 2024).
8. Zhao, G., Sonsaat, S., Silpachai, A. et al. (2018) 'L2-ARCTIC: A non-native english speech corpus', *Interspeech*, pp. 2783–2787. doi: 10.21437/Interspeech.2018-1110.
9. Garofolo, J.S., Lamel, L.F., Fisher, W.M. et al. (1992) *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Gaithersburg, 95 p.
10. Rosana, A., Branson, M., Davis, K. et al. (2020) 'Common Voice: A massively-multilingual speech corpus', *Proc. Twelfth Language Resources and Evaluation Conf.*, pp. 4218–4222.
11. *The CMU Pronouncing Dictionary*, available at: <http://www.speech.cs.cmu.edu/cgi-bin/cmudict> (accessed June 18, 2024).
12. *Mrrubino. Hugging Face*, available at: <https://huggingface.co/mrrubino/wav2vec2-large-xlsr-53-12-arctic-phoneme> (accessed June, 18 2024).
13. *Gruut Documentation*, available at: <https://rhaspy.github.io/gruut/> (accessed June 18, 2024).
14. Conneau, A., Baeviski, A., Collobert, R. et al. (2020) 'Unsupervised cross-lingual representation learning for speech recognition', *ArXiv*, art. 2006.13979v2, available at: <https://arxiv.org/abs/2006.13979> (accessed June 18, 2024).
15. *LDC Catalog by Year*, available at: <https://catalog.ldc.upenn.edu/byyear> (accessed June 18, 2024).
16. Vaswani, A., Shazeer, N., Parmar, N. et al. (2017) 'Attention is all you need', *ArXiv*, art. 1706.03762v5, available at: <https://arxiv.org/pdf/1706.03762v5> (accessed June 18, 2024).
17. Kobzdej, P., Waligóra, D., Wielebińska, K., Paprzycki, M. (2003) 'Parallel application of levenshtein distance to establish similarity between strings', *Frontiers in Artificial Intelligence and Applications*, 12(4).
18. *Moxeeem. Hugging Face*, available at: <https://huggingface.co/moxeeem/wav2vec2-finetuned-pronunciation-correction> (accessed June 18, 2024).

Авторы

Дорохин Михаил Андреевич¹, бакалавр,
michaelsagittarius08@gmail.com
Чернышев Станислав Андреевич¹,
к.т.н., доцент,
chernyshev.s.a@bk.ru

Authors

Mikhail A. Dorokhin¹, Bachelor of Sci.,
michaelsagittarius08@gmail.com
Stanislav A. Chernyshev¹,
Cand. of Sci. (Engineering), Associate Professor
chernyshev.s.a@bk.ru

¹ Санкт-Петербургский государственный университет
аэрокосмического приборостроения,
Санкт-Петербург, 190000, Россия

¹ Saint Petersburg State University
of Aerospace Instrumentation,
St. Petersburg, 190000, Russian Federation