ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук
Образовательная программа «Прикладная математика и информатика»

Отчет об исследовательском проекте на тему:

Влияние языка на системы автоматической верификации говорящего

Выполнил студент:

группы # БПМИ232, 2 курса Кузнецов Степан Андреевич

Принял руководитель проекта:

Гринберг Петр Маркович
Приглашенный преподаватель
НИУ ВШЭ

Москва 2025

Содержание

Аннотация

Современные системы верификации говорящего (Speaker Verification, SV) преимущественно обучаются на англоязычных данных, что ограничивает их применимость в многоязычных и малоресурсных условиях. В данной работе исследуется влияние предварительного обучения (pretraining) на больших мультиязычных наборах данных на способность моделей эффективно решать задачу верификации говорящего при языковом разнообразии и ограниченном объёме данных. Основной гипотезой является предположение о том, что предварительное обучение на больших многоязычных корпусах улучшает обобщающую способность моделей в условиях языкового разнообразия и ограниченных ресурсов данных. Модели тестировались на мультиязычных и малоресурсных датасетах (включая языки Тамиль, Сингальский и Зулу). Эксперименты подтвердили, что мультиязычное предварительное обучение существенно повышает устойчивость и обобщающую способность моделей, что выражается в улучшении метрики средней ошибки верификации (англ. Equal Error Rate, EER), особенно при работе с короткими аудиозаписями и малоресурсными языками. Реализация моделей и конфигурации экспериментов доступны в репозитории https://github.com/Stefan2417/CourseWork

Ключевые слова

Глубинное обучение, верификация говорящего, предварительное обучение, многоязычность, малоресурсные языки.

1 Введение

Голос человека обладает уникальными характеристиками, которые зависят от строения органов артикуляции, акцента, манеры речи [22]. Такие особенности позволяют создать автоматическую систему сопоставления голоса и личности говорящего, основанную на методах машинного обучения. Распознавание говорящего является фундаментальной задачей обработки речи и находит широкое применение в реальных задачах. Например оно используется для аунтетификации персональных устройств посредством голоса, обеспечивает безопасность банковских транзакций, применяется в криминалистике [6]. Данная работа посвящена исследованию технологии, известной как «автоматическая верификация говорящего», которая является разновидностью автоматического распознавания говорящего.

Главная цель задачи верификации говорящего — определить, соответствует ли предоставленная голосовая запись заявленной личности, сравнивая её с заранее сохранённой моделью голоса (голосовым отпечатком) данного человека. Работа системы верификации, изображенная на Рисунке 1.1, включает следующие этапы:

1 **Регистрация**. Пользователь предоставляет образцы своей речи, которые используются для извлечения векторных представлений, отражающих уникальные характеристики голоса. Векторы формируют уникальный профиль пользователя.

2 **Верификация**. Пользователь предоставляет новую голосовую запись, представление которой сравнивается с его профилем.

3 **Принятие решения**. На основе сравнения векторов [14, 29], система принимает решение о подтверждении личности, используя заданные пороговые значения.

Современные системы верификации обучаются преимущественно на англоязычных данных [32], что создает препятствия для их применения в многоязычных средах из-за уникальных фонетических и лингвистических особенностей разных языков [36]. Отсутствие больших мультиязычных корпусов данных, покрывающих малоресурсные языки, является ключевой проблемой для обучения моделей верификации. Одной из задач становится обеспечение конкурентного качества верификации в условиях малого количества данных.

Хотя система верификации говорящего состоит из нескольких этапов — от предварительной обработки аудио до принятия окончательного решения, в основе всей системы лежит модель верификации, извлекающая и сравнивающая голосовые представления. Основное внимание уделяется модели верификации, исследованию влияния предварительного

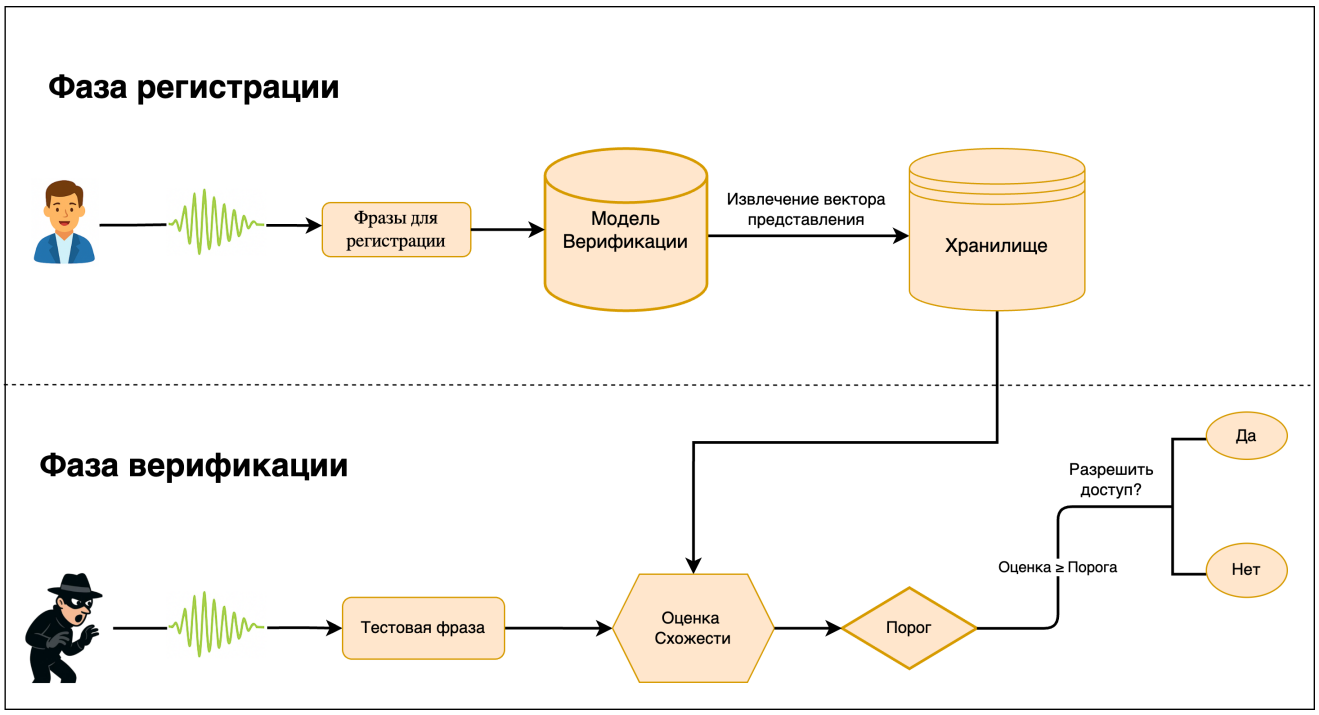

Рис. 1.1: Процесс работы системы верификации говорящего. [1, 5]

обучения на больших мультиязычных корпусах и последующего дообучения (fine-tuning) на способность модели эффективно работать в условиях языкового разнообразия, коротких записей и ограниченных данных. Результаты, полученные на двух современных моделей, обученных по принципу обучения без учителя (self-supervised learning, SSL), показывают, что такой подход позволяет значительно улучшить точность в условиях языкового разнообразия и дефицита данных.

2 Обзор литературы

В последние годы глубокие нейронные сети значительно улучшили качество систем верификации говорящего. Традиционные подходы, основанные на i-vectors [37] были заменены более современными архитектурами [5].

Текущими основными направлениями для улучшения качества верификации в условиях языкового разнообразия, согласно [5, 39] являются:

1 Трансферное обучение с использованием предварительно обученных моделей для автоматического распознавания речи (Automatic Speech Recognition — ASR). Этот метод предполагает использование предварительно обученных моделей для инициализации сетей в задачах верификации говорящих.

2 Использование моделей с обучением без разметки (self-supervised learning SSL) [11]. Мо-

дели, такие как wav2vec 2.0 [4], HuBERT [19] и WavLM [8], содержат богатую фонетическую информацию. Благодаря этому можно использовать их способность к извлечению признаков.

Указанные направления отражают более широкий тренд: современные системы верификации говорящего всё чаще строятся на базе крупных предварительно обученных моделей, обученных на масштабных и мультиязычных речевых корпусах. Такие модели формируют более универсальные, независимые от языка представления, что делает их особенно устойчивыми к языковым и акустическим различиям, а также позволяет эффективно работать в условиях ограниченного количества данных [39]. Это объясняет активное использование моделей вроде Wav2Vec 2.0 [4], HuBERT [19], Whisper [30] и XEUS [10].

В статье [16], посвященной адаптации Wav2Vec 2.0 [4] на задачу верификации говорящего, авторы предобучили модель на большом корпусе данных Librispeech [28], который включает в себя только английскую речь. Далее они дообучили модель для задачи верификации говорящего, используя набор данных VoxCeleb1 [26] и получили конкурентные результаты метрики «средняя ошибка верификации» (Equal Error Rate — EER) [41]. В этой статье нет анализа мультиязычного сценария, результаты получены на датасете VoxCeleb1 [26], который содержит преимущественно английскую речь. Этот метод можно отнести ко второму направлению улучшения задачи верификации.

В качестве примера первого направления можно привести работу [43] по адаптации модели Whisper [30] для задачи верификации говорящего. Whisper [30] представляет собой многозадачную модель для обработки речи, предобученную на масштабном мультиязычном корпусе объемом 680 000 часов. Авторы разработали специальный адаптер, позволяющий эффективно использовать речевые представления Whisper [30] даже при ограниченном количестве обучающих данных. Ключевой особенностью предложенного подхода является сохранение мультиязычных возможностей исходной модели при адаптации к задаче верификации говорящего. Экспериментальные результаты показали, что такой подход демонстрирует лучшие результаты для мультиязычных датасетов. Результаты работы [43] используются для сравнения.

Для ASR-модели Whisper [30] существует еще одна статья [42]. Модель похожим образом адаптируют под задачу верификации говорящего, но используют другую архитектуру адаптера.

В статье [34] авторы предлагают подход, сочетающий в себе элементы трансферного обучения и неявного использования фонетической информации из архитектуры Conformer [18].

Основная идея заключается в использовании предварительно обученной ASR-модели Conformer [18], которую адаптируют с помощью трансферного обучения под задачу верификации говорящего. Авторы предобучают Conformer [18] на моноязычном и мультиязычном корпусе данных и получают предсказуемые результаты — модель, обученная на мультиязычном корпусе, имеет лучшие результаты в мультиязычном наборе данных для задачи верификации говорящего. Предложенный подход достигает передовых результатов на наборе VoxCeleb1 [26].

3 Методология

В данной работе уделено внимание второму направлению улучшения качества верификации в сложных языковых сценариях. В этой секции описаны используемые модели и метод адаптации к задаче верификации говорящего.

3.1 Адаптация моделей

В рамках данной работы в качестве метода адаптации предобученных речевых моделей для задачи верификации говорящего выбран подход частичной агрегации признаков на основе многомасштабного анализа (partial multi-scale feature aggregation PMFA). Этот метод предложен в работах [43, 34]. Он позволяет эффективно использовать выходы с различных уровней модели, агрегируя многомасштабную информацию и тем самым усиливая дискриминативную способность получаемых векторов представления говорящих.

В отличие от полной агрегации признаков со всех слоёв, подход PMFA предлагает выбор подмножества $S = \{S_1, S_2, \ldots, S_k\}$ слоёв предобученной модели, выходные тензоры которых затем конкатенируются и нормализуются:

$$\mathbf{H}' = \text{Concat}(\mathbf{S_1}, \mathbf{S_2}, \ldots, \mathbf{S_k}) \tag{1}$$

$$\mathbf{H} = \text{LayerNorm}(\mathbf{H}') \tag{2}$$

$\mathbf{S_i} \in \mathbb{R}^{d' \times T}$ — выход с S_i-го слоя модели, d' — размерность скрытого представления, T — длина по времени. Таким образом, итоговое представление $\mathbf{H} \in \mathbb{R}^{(k \cdot d') \times T}$ содержит информацию с различных уровней абстракции, что повышает его устойчивость к языковым и акустическим вариациям.

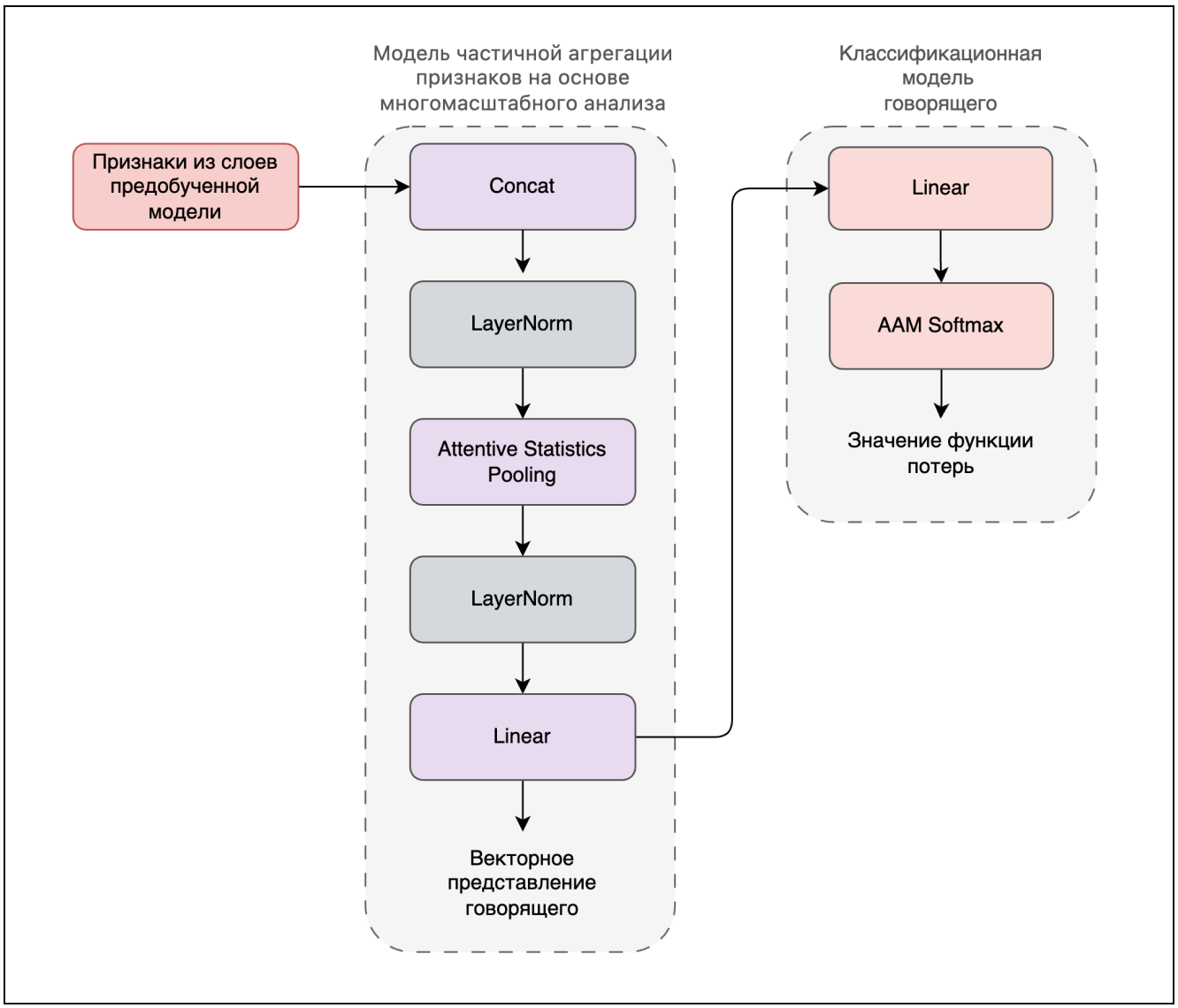

Рис. 3.1: Общая архитектура адаптера на основе частичной агрегации признаков с многомасштабным анализом. Адаптер используется для извлечения векторов представления говорящего.

Для формирования вектора представления используется статистический пуллинг с механизмом внимания (Attentive Statistics Pooling, ASP) [27]. На выходе формируется вектор $\mathbf{Z}' \in \mathbb{R}^{2 \cdot k \cdot d}$. Этот вектор нормализуется повторно с использованием `LayerNorm`, после чего подаётся в полносвязный слой проекции:

$$\mathbf{Z} = \text{Linear}(\text{LayerNorm}(\mathbf{Z}')) \in \mathbb{R}^d \tag{3}$$

где d — размерность итогового эмбеддинга, используемого для итогового сравнения аудиозаписей и подачи в классификационную голову.

Классификатор состоит из ещё одного полносвязного слоя и функции потерь Additive Angular Margin Softmax (AAM Softmax) [38], обеспечивающей повышение межклассовой раз-

личимости. Вся схема архитектуры приведена на Рисунке 3.1.

В отличие от работ [43, 34], в качестве слоя нормализации применяется `LayerNorm` вместо `BatchNorm`. Это связано с тем, что в условиях ограниченного объёма видеопамяти дообучение модели производится с использованием малого размера пакета данных (batch). Как показано в работе [3], в таких условиях `LayerNorm` демонстрирует более стабильное поведение по сравнению с `BatchNorm`, поскольку нормализация производится на уровне отдельных примеров, а не по батчу, что делает её менее чувствительной к размеру последнего.

3.2 XEUS

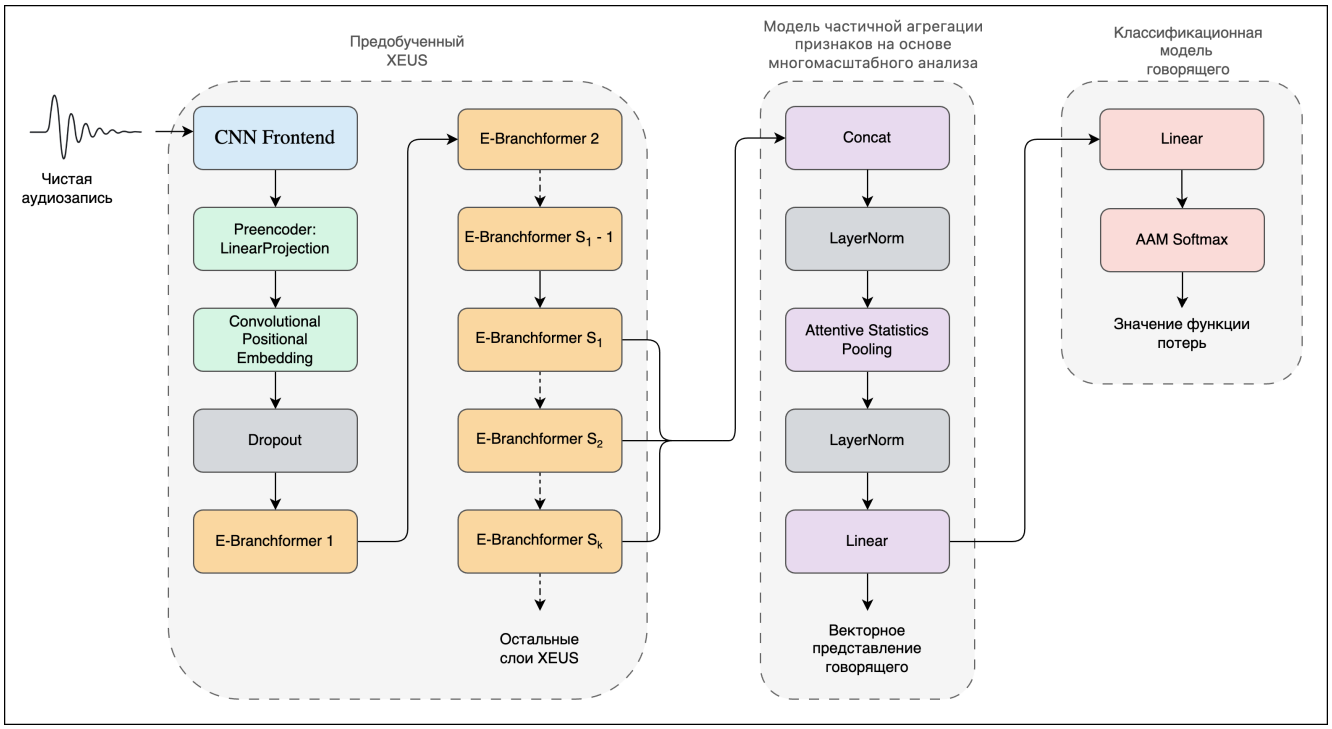

Рис. 3.2: Архитектура XEUS [10] для адаптации под задачу верификации говорящего.

В качестве первой модели, с которой проводились эксперименты, использовалась XEUS (Cross-lingual Encoder for Universal Speech) [10] — это самообучающаяся модель представления речи, ориентированная на многоязычные и акустически разнообразные условия. Архитектура модели построена на основе E-Branchformer [20] и обучена на более чем миллионе часов речевых данных, охватывающих 4057 языков.

На вход модели подаётся чистая аудиозапись с частотой дискретизации 16 кГц. Перед поступлением в основной модуль кодирования (encoder), сигнал обрабатывается с помощью CNN Frontend, состоящего из семи последовательно расположенных сверточных блоков. Этот блок понижает временное разрешение и извлекает первичные акустические признаки, формируя скрытое представление размерности 512.

Затем применяется модуль Linear Projection, преобразующий признаковое пространство в размерность 1024. Далее идут позиционное кодирование (Convolutional Positional Embedding) и регуляризация (Dropout). Полученное представление подаётся в 19 слоёв E-Branchformer.

Выходные тензоры с различных слоёв XEUS используются в рамках подхода частичной агрегации признаков PMFA 3.1. Для агрегации выбирается подмножество слоёв $S = \{S_1, S_2, \ldots, S_k\}$, что позволяет учитывать как низкоуровневые, так и высокоуровневые речевые особенности. Детали архитектуры отображены в 3.2.

3.3 Wav2Vec2-BERT 2.0

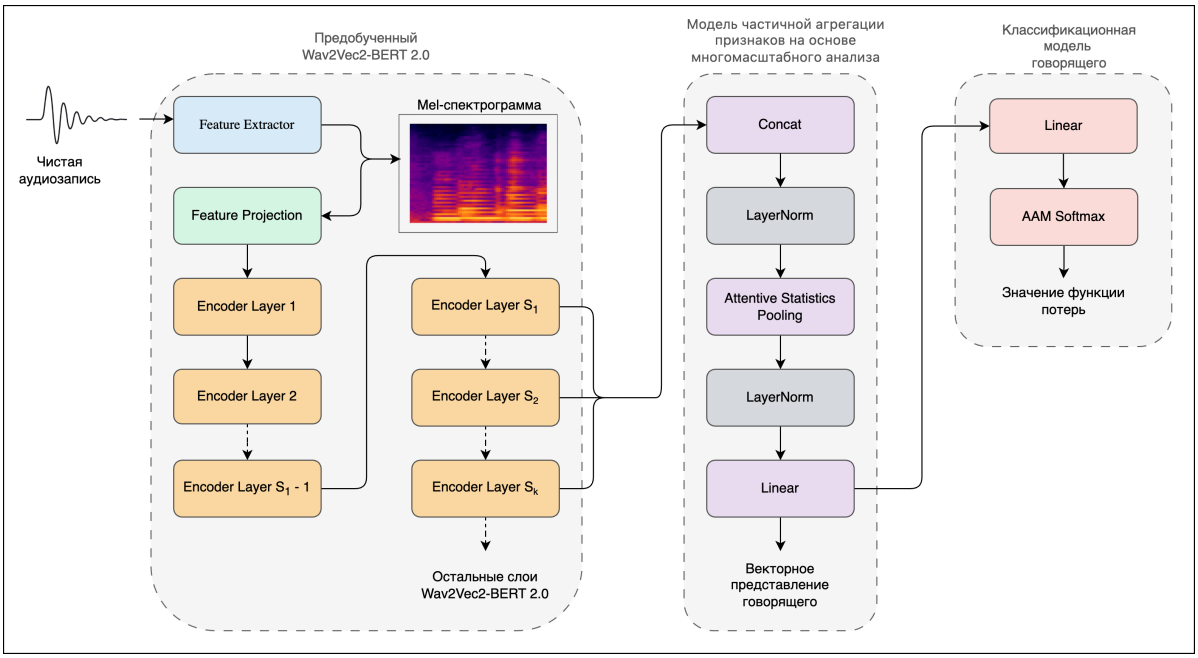

Рис. 3.3: Архитектура Wav2Vec2-BERT 2.0 [13, 12] для адаптации на задачу верификации говорящего.

В качестве второй модели использовалась Wav2Vec2-BERT 2.0 [13, 12] — продолжение Wav2Vec 2.0 [4], в котором модуль кодирования объединён с BERT-подобным трансформером. Такое объединение позволяет эффективно обрабатывать как низкоуровневую, так и высокоуровневую информацию в речевом сигнале.

Архитектура Wav2Vec2-BERT 2.0 [13, 12] сочетает извлечение акустических признаков с их последующей семантической обработкой, что делает модель особенно подходящей для задач, требующих устойчивых и универсальных речевых представлений.

На вход модели также подаётся аудиозапись с частотой дискретизации 16 кГц. Сначала сигнал преобразуется в Mel-спектрограмму с 80 каналами с помощью блока Feature

Extractor. Далее следует Feature Projection — линейная проекция с последующей нормализацией, которая приводит признаки к размерности $d = 512$.

Полученное представление подаётся на вход модулю кодирования, состоящему из 24 слоёв Transformer-подобной архитектуры. В рамках подхода PMFA 3.1 из модели извлекаются выходы подмножества слоёв $S = \{S_1, S_2, \ldots, S_k\}$, используемые для построения вектора представления говорящего. Детали архитектуры отображены в 3.3.

4 Конфигурация экспериментов

В данном разделе описаны конфигурации, с использованием которых проводились эксперименты.

4.1 Наборы данных

Для оценки эффективности моделей были использованы следующие наборы данных:

- VoxCeleb2 [26] — масштабный аудиовизуальный набор данных для задач распознавания говорящего, содержащий более миллиона высказываний от более чем 6 тысяч знаменитостей. Использовался для обучения всех моделей.
- VoxCeleb1 [26] — набор данных, содержащий свыше 150,000 высказываний от 1,251 говорящего. Применялся для тестирования моделей. В экспериментах использовались следующие списки пар:
  - VoxCeleb1-O — оригинальный список пар;
  - VoxCeleb1-Clean, VoxCeleb1-H, VoxCeleb1-E — списки с различными уровнями сложности и степенью очистки;
  - VoxSRC2021 val — валидационная часть официального теста VoxSRC 2021.
- SL-Celeb [2] — семейство аудиодатасетов, предназначенное для оценки моделей на специфических южноазиатских языках. В данной работе использовались два набора:
  - SL-Celeb Tamil — записи на языке тамиль;
  - SL-Celeb Sinhala — записи на сингальском языке.

  Для тестирования применялись официальные списки пар.

- NCHLT isiZulu Speech Corpus [35] — открытый речевой корпус языка Зулу, представителя щёлкающих языков, распространённых в Южной Африке. На его основе был сформирован тестовый набор пар, подготовленный в рамках данной работы для задачи верификации говорящего. При формировании учитывался баланс положительных и отрицательных примеров, а также гендерный баланс (соотношение 1:1).

4.2 Конфигурация моделей

В экспериментах использовались три модели: XEUS [10], Wav2Vec2-BERT 2.0 [13, 12] и ECAPA-TDNN [31, 15]. Все модели были адаптированы под задачу верификации говорящего и сравнивались в идентичных условиях.

- XEUS: использовалась публичная реализация модели из ESPnet [40]. Для адаптации применялся предложенный в работе адаптер на основе частичной агрегации признаков (PMFA), описанный в разделе 3.1. Для итогового вектора представления выбрана размерность 256. В качестве слоёв для частичной агрегации признаков были выбраны слои $\{6, 7, 8, 9, 10, 11\}$. Такой выбор мотивирован результатами недавних исследований [43, 24], где показано, что представления промежуточных слоёв SSL моделей содержат наибольшее количество дискриминативной информации, релевантной задаче верификации говорящего. Более высокие уровни зачастую фокусируются на семантическом содержании, тогда как нижние слои отражают в основном акустические особенности. Использование же признаков из средних слоёв позволяет достичь наилучшего баланса между обобщением и сохранением персонализированных признаков голоса.

- Wav2Vec2-BERT 2.0 [13, 12]: для экспериментов была использована точка сохранения `facebook/w2v-bert-2.0`, доступная на платформе Hugging Face[1]. Данная точка сохранения обучена на 4,5 миллионах часов немаркированных аудиоданных, охватывающих более 143 языков, и демонстрирует высокую эффективность при переносе знаний на новые языки и задачи. Для итогового вектора представления выбрана размерность 512. В качестве слоёв для агрегации были выбраны $\{12, 13, 14, 15, 16, 17, 18\}$ — область, также соответствующая промежуточным уровням модели, как рекомендовано в ряде исследований [43, 24].

- ECAPA-TDNN [31, 15]: использовалась в качестве сильного базового метода. Модель была обучена на датасете VoxCeleb2 [26] и применялась без дополнительной адаптации

[1] https://huggingface.co/facebook/w2v-bert-2.0

для малоресурсных языков. Конфигурация соответствует стандартной реализации из SpeechBrain [31]. Итоговый вектор представления имеет размерность 192.

4.3 Детали реализации

Для повышения устойчивости моделей к акустическим и шумовым искажениям использовались следующие методы аугментации:

- добавление фонового шума из датасета MUSAN [33];
- добавление реверберации на основе импульсных характеристик помещений из набора RIRs [17].

Таблица 4.1: Параметры обучения для разных моделей

Параметр	**XEUS [10]**	**Wav2Vec2-BERT 2.0 [13, 12]**
Этап 1: обучение адаптера		
Длина аудиофрагментов	5 секунд (фикс.)	3 секунды (фикс.)
Learning rate	10^{-3}	10^{-3}
Scheduler	StepLR, $\gamma = 0.7$	StepLR, $\gamma = 0.7$
Оптимизатор	Adam [21], weight decay $2 \cdot 10^{-5}$	Adam [21], weight decay $2 \cdot 10^{-5}$
Этап 2: полное дообучение		
Длина аудиофрагментов	5 секунд (фикс.)	3 секунды (фикс.)
Learning rate	10^{-4}/адаптер, 10^{-5}/модель	10^{-4}/адаптер, 10^{-5}/модель
Scheduler	StepLR, $\gamma = 0.32$	StepLR, $\gamma = 0.5$
Weight decay	10^{-6}	10^{-6}
Функция потерь	AAM-Softmax [38], параметры: $margin = 0.2$, $scale = 30.0$	
Этап 3: дообучение с увеличенным отступом		
Применение	✗	✓
Длина аудиофрагментов	—	5 секунд (фикс.)
Learning rate	—	$4 \cdot 10^{-6}$
$margin$ в AAM-Softmax	—	0.5

4.4 Методы оценки

Для оценки качества систем использовалась метрика Equal Error Rate (EER), основанная на косинусной близости между векторами представления аудиозаписей.

Перед вычислением сходства каждая аудиозапись приводилась к фиксированной длине: если длительность записи была менее 5 секунд, она дополнялась повтором самого сигнала до нужной длины; если аудиофайл превышал 40 секунд, он усекался до первых 40 секунд.

5 Результаты и анализ

Таблица 5.1: Сравнение моделей на подмножествах VoxCeleb1 (EER, %)

.

Модель	VoxCeleb1-O	VoxCeleb1-Clean	VoxCeleb1-H	VoxCeleb1-E	Fine-tuning	Large-margin
ECAPA-TDNN	1.42	1.28	2.74	1.50	✗	✗
XEUS	1.41	1.28	2.53	1.40	✗	✗
Wav2Vec2-BERT 2.0	1.18	1.03	2.18	1.27	✗	✗
XEUS	1.29	1.15	2.24	1.16	✓	✗
Wav2Vec2-BERT 2.0	0.60	0.46	1.10	0.56	✓	✗
Wav2Vec2-BERT 2.0	0.46	0.33	0.99	0.51	✓	✓
WavLM Large	0.431	—	1.154	0.538	✓	✗
UniSpeech-SAT Large	0.564	—	1.230	0.561	✓	✗
Wav2Vec2.0 (XLSR)	0.564	—	1.230	0.605	✓	✗
HuBERT Large	0.585	—	1.342	0.654	✓	✗
Whisper-PMFA	1.42	—	—	—	✗	✗
Whisper-SV	1.71	—	—	—	✗	✗
MFA-Conformer	0.606	—	1.918	0.903	✓	✓

Таблица 5.2: Сравнение моделей на мультиязычных и малоресурсных наборах данных (EER, %).

Модель	VoxSRC21-Val	SL-Celeb-Tamil	SL-Celeb-Sinhala	Zulu	Fine-tuning	Large-margin
ECAPA-TDNN	5.05	3.27	4.72	3.30	✗	✗
XEUS	4.39	6.42	3.47	3.00	✗	✗
Wav2Vec2-BERT 2.0	4.38	5.99	5.10	1.97	✗	✗
XEUS	4.06	6.01	3.79	2.50	✓	✗
Wav2Vec2-BERT 2.0	2.00	2.73	4.37	1.64	✓	✗
Wav2Vec2-BERT 2.0	1.82	1.39	4.28	1.9	✓	✓
MFA-Conformer	3.77	—	—	—	✓	✓
snowstar (Track 2)	**1.85**	—	—	—	—	—
JTBD (Track 2)	2.05	—	—	—	—	—

В данной секции представлены результаты экспериментов и анализируются причины наблюдаемых различий в эффективности моделей в задаче верификации говорящего с акцентом на многоязычные и малоресурсные условия. В таблицах 5.1 5.2 указано наличие дообучения (Fine-tunning) и дообучения с увеличинным отступом (Large-margin).

Результаты сравнения моделей на подмножествах датасета VoxCeleb1 представлены в Таблице 5.1. Модель Wav2Vec2-BERT 2.0 [13, 12] показывает наиболее низкие значения EER на всех подмножествах, существенно превосходя модель XEUS [10] и базовую архитектуру ECAPA-TDNN [31, 15]. Высокое качество модели Wav2Vec2-BERT 2.0 [13, 12] связано с успешным применением подхода адаптации через частичную агрегацию признаков (PMFA) и эффективной схемой двухэтапного обучения.

Дополнительно были рассмотрены результаты других авторов [25, 11] , представленные в нижней части Таблицы 5.1. Из них видно, что модель Wav2Vec2-BERT 2.0 [13, 12] демонстрирует сопоставимые и в некоторых случаях даже лучшие результаты по сравнению

с такими известными архитектурами, как WavLM Large [8], UniSpeech-SAT Large [9], Whisper-SV [42], Whisper-PFMA [43], MFA-Conformer [18] и HuBERT Large [19], которые считаются одними из наиболее эффективных в задаче верификации говорящего на английском языке. Это подчёркивает конкурентоспособность выбранного подхода адаптации и применяемой архитектуры.

Анализ результатов на малоресурсных наборах данных (SL-Celeb и Zulu), представленных в Таблице 5.2, выявил важные особенности работы моделей:

На наборах SL-Celeb модель XEUS [10] показала менее конкурентные результаты, что можно объяснить доменным несоответствием между условиями обучения и тестирования. Средняя продолжительность аудиозаписей SL-Celeb составляет примерно 2.5 секунды, что существенно меньше продолжительности фрагментов, использованных при обучении XEUS [10] (5 секунд). Таким образом, короткие сегменты речи оказывают негативное влияние на качество верификации говорящего при использовании XEUS [10].

Ещё одним фактором, отрицательно сказавшимся на эффективности XEUS [10], является её склонность к переобучению. В процессе экспериментов было замечено, что значение функции потерь XEUS [10] приближалось к нулю уже через пару эпох обучения, что свидетельствует о слишком быстром запоминании обучающего набора и недостаточном обобщении. Модель имеет потенциал, возможно необходима другая стратегия выбора слоев и гиперпараметров.

Модель Wav2Vec2-BERT 2.0 [13, 12] показала высокую эффективность даже при коротких аудиозаписях (около 2.5 секунд), несмотря на то, что при обучении использовались записи длительностью около 3 секунд. Это свидетельствует о том, что модель способна эффективно обобщать речевые представления на временные интервалы, отличающиеся от использованных при обучении.

На наборе данных языка Зулу модель Wav2Vec2-BERT 2.0 [13, 12] также продемонстрировала лучшие результаты по сравнению с другими моделями. Это указывает на её способность успешно адаптироваться и сохранять эффективность при значительных фонетических и акустических различиях от языка обучения. Наиболее заметны ее конкурентные результаты в VoxSRC21, для которого были взяты 2 лучших участника, согласно результатам [7]. Дообучение с увеличинным отступом позволило добиться более впечетляющих результатов на всех тестовых наборах, кроме Зулу. Это может быть связано с усилением межклассового разделения, которое не всегда эффективно в условиях ограниченного количества данных и высокой акустической вариативности языка Зулу. Увеличенный отступ в функции потерь AAM Softmax требует от модели более чётких границ между классами, что

при низком покрытии фонетических особенностей языка в обучающих данных может приводить к переадаптации под шум или нерелевантные признаки. В результате — снижение обобщающей способности и ухудшение результатов на тесте.

Дополнительно стоит отметить, что двухэтапная схема обучения, включающая предварительное обучение адаптера с замороженными весами модуля кодирования, а затем дообучение всей модели, оказалась эффективной и с точки зрения качества, и с точки зрения вычислительных ресурсов. Такой подход позволяет избежать негативного влияния случайной инициализации адаптера, которая может вносить шум в ранее выученные представления, и тем самым сохраняет преимущества предварительного обучения.

6 Заключение

Таким образом, полученные результаты подтверждают роль предварительного обучения на больших многоязычных корпусах для улучшения качества и обобщающей способности моделей верификации говорящего, особенно в условиях ограниченного объёма данных и коротких аудиозаписей.

Модель Wav2Vec2-BERT 2.0, уже обладающая богатым набором мультиязычных признаков, показала значительно более высокую устойчивость к доменным и языковым сдвигам, чем решения, прошедшие менее обширное или одноязычное предобучение. Это проявилось как при сравнении на крупных сетах VoxCeleb1, так и на малоресурсных языках (Тамиль, Сингальский и Зулу).

XEUS продемонстрировала результаты, которые могут быть ограничены не столько архитектурными особенностями, сколько методикой её адаптации к новым данным. Поведение модели указывает на то, что при работе с короткими записями и в условиях доменного несоответствия особую роль играет корректная настройка и использование предобученных мультиязычных представлений.

В целом, эксперименты подтвердили, что многоязычное предварительное обучение даёт моделям заметное преимущество при решении задачи верификации говорящего в сложных условиях, и при последующем дообучении может способствовать дополнительному росту качества.

Благодарности

Исследование выполнено с использованием суперкомпьютерного комплекса НИУ ВШЭ [23].

Список литературы

[1] K.N.R.K. Raju Alluri и Anil Kumar Vuppala. “Chapter 7 - A study on the emotional state of a speaker in voice bio-metrics”. В: *Advances in Ubiquitous Computing*. Под ред. Amy Neustein. Advances in ubiquitous sensing applications for healthcare. Academic Press, 2020, с. 223—236. DOI: https://doi.org/10.1016/B978-0-12-816801-1.00007-4. URL: https://www.sciencedirect.com/science/article/pii/B9780128168011000074.

[2] Dimuthu Anuraj, Jarashanth Selvarajah, Kanagasundaram Ahilan, Ragupathyraj Valluvan, Thiruvaran Tharmarajah и Anantharajah Kaneswaran. *SLCeleb for Speaker Verification*. 2023. DOI: 10.21227/smmf-e298. URL: https://dx.doi.org/10.21227/smmf-e298.

[3] Jimmy Lei Ba, Jamie Ryan Kiros и Geoffrey E. Hinton. *Layer Normalization*. 2016. arXiv: 1607.06450 [stat.ML]. URL: https://arxiv.org/abs/1607.06450.

[4] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed и Michael Auli. *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: 2006.11477 [cs.CL]. URL: https://arxiv.org/abs/2006.11477.

[5] Zhongxin Bai и Xiao-Lei Zhang. *Speaker Recognition Based on Deep Learning: An Overview*. 2021. arXiv: 2012.00931 [eess.AS]. URL: https://arxiv.org/abs/2012.00931.

[6] Homayoon Beigi. “Speaker Recognition: Advancements and Challenges”. В: нояб. 2012, с. 3—29. ISBN: 978-953-51-0859-7. DOI: 10.5772/52023.

[7] Andrew Brown, Jaesung Huh, Joon Son Chung, Arsha Nagrani, Daniel Garcia-Romero и Andrew Zisserman. *VoxSRC 2021: The Third VoxCeleb Speaker Recognition Challenge*. 2022. arXiv: 2201.04583 [cs.SD]. URL: https://arxiv.org/abs/2201.04583.

[8] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu и Furu Wei. “WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing”. В: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (окт. 2022), с. 1505—1518. ISSN: 1941-0484. DOI: 10.1109/jstsp.2022.3188113. URL: http://dx.doi.org/10.1109/JSTSP.2022.3188113.

[9] Sanyuan Chen, Yu Wu, Chengyi Wang, Zhengyang Chen, Zhuo Chen, Shujie Liu, Jian Wu, Yao Qian, Furu Wei, Jinyu Li и Xiangzhan Yu. *UniSpeech-SAT: Universal Speech Representation Learning with Speaker Aware Pre-Training*. 2021. arXiv: 2110.05752 [cs.CL]. URL: https://arxiv.org/abs/2110.05752.

[10] William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu и Shinji Watanabe. *Towards Robust Speech Representation Learning for Thousands of Languages*. 2024. arXiv: 2407.00837 [cs.CL]. URL: https://arxiv.org/abs/2407.00837.

[11] Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian и Michael Zeng. *Large-scale Self-Supervised Speech Representation Learning for Automatic Speaker Verification*. 2022. arXiv: 2110.05777 [cs.SD]. URL: https://arxiv.org/abs/2110.05777.

[12] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang и Yonghui Wu. *W2v-BERT: Combining Contrastive Learning and Masked Language Modeling for Self-Supervised Speech Pre-Training*. 2021. arXiv: 2108.06209 [cs.LG]. URL: https://arxiv.org/abs/2108.06209.

[13] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang и Mary Williamson. *Seamless: Multilingual Expressive and Streaming Speech Translation*. 2023. arXiv: 2312.05187 [cs.CL]. URL: https://arxiv.org/abs/2312.05187.

[14] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel и Pierre Ouellet. "Front-End Factor Analysis for Speaker Verification". В: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), с. 788—798. DOI: 10.1109/TASL.2010.2064307.

[15] Brecht Desplanques, Jenthe Thienpondt и Kris Demuynck. "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification". В: *Interspeech 2020*. interspeech$_2$020. ISCA, окт. 2020. DOI: 10.21437/interspeech.2020-2650. URL: http://dx.doi.org/10.21437/Interspeech.2020-2650.

[16] Zhiyun Fan, Meng Li, Shiyu Zhou и Bo Xu. *Exploring wav2vec 2.0 on speaker verification and language identification*. 2021. arXiv: 2012.06185 [cs.SD]. URL: https://arxiv.org/abs/2012.06185.

[17] María Pilar Fernández-Gallego и Doroteo T. Toledano. "A Study of Data Augmentation for ASR Robustness in Low Bit Rate Contact Center Recordings Including Packet Losses". В: *Applied Sciences* 12.3 (2022). ISSN: 2076-3417. DOI: 10.3390/app12031580. URL: https://www.mdpi.com/2076-3417/12/3/1580.

[18] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu и Ruoming Pang. *Conformer: Convolution-augmented Transformer for Speech Recognition*. 2020. arXiv: 2005.08100 [eess.AS]. URL: https://arxiv.org/abs/2005.08100.

[19] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov и Abdelrahman Mohamed. *HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units*. 2021. arXiv: 2106.07447 [cs.CL]. URL: https://arxiv.org/abs/2106.07447.

[20] Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han и Shinji Watanabe. *E-Branchformer: Branchformer with Enhanced merging for speech recognition*. 2022. arXiv: 2210.00077 [eess.AS]. URL: https://arxiv.org/abs/2210.00077.

[21] Diederik P. Kingma и Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[22] Tomi Kinnunen и Haizhou Li. "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors". В: *Speech Communication* 52 (янв. 2010), с. 12—40. DOI: 10.1016/j.specom.2009.08.009.

[23] P. S. Kostenetskiy, R. A. Chulkevich и V. I. Kozyrev. "HPC Resources of the Higher School of Economics". В: *Journal of Physics: Conference Series* 1740.1 (2021), с. 012050. DOI: 10.1088/1742-6596/1740/1/012050.

[24] Weiwei Lin, Chenhang He, Man-Wai Mak и Youzhi Tu. *Self-supervised Neural Factor Analysis for Disentangling Utterance-level Speech Representations*. 2023. arXiv: 2305.08099 [cs.SD]. URL: https://arxiv.org/abs/2305.08099.

[25] Microsoft. *UniSpeech: Speaker Verification (GitHub Repository)*. https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification. Accessed: 31.03.2025. 2021.

[26] Arsha Nagrani, Joon Son Chung, Weidi Xie и Andrew Zisserman. “Voxceleb: Large-scale speaker verification in the wild”. В: *Computer Science and Language* (2019).

[27] Koji Okabe, Takafumi Koshinaka и Koichi Shinoda. “Attentive Statistics Pooling for Deep Speaker Embedding”. В: *Interspeech 2018*. interspeech$_2$018. ISCA, сент. 2018. DOI: 10.21437/interspeech.2018-993. URL: http://dx.doi.org/10.21437/Interspeech.2018-993.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey и Sanjeev Khudanpur. “Librispeech: An ASR corpus based on public domain audio books”. В: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, с. 5206—5210. DOI: 10.1109/ICASSP.2015.7178964.

[29] Simon J.D. Prince и James H. Elder. “Probabilistic Linear Discriminant Analysis for Inferences About Identity”. В: *2007 IEEE 11th International Conference on Computer Vision*. 2007, с. 1—8. DOI: 10.1109/ICCV.2007.4409052.

[30] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey и Ilya Sutskever. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS]. URL: https://arxiv.org/abs/2212.04356.

[31] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori и Yoshua Bengio. *SpeechBrain: A General-Purpose Speech Toolkit*. 2021. arXiv: 2106.04624 [eess.AS]. URL: https://arxiv.org/abs/2106.04624.

[32] R. Sharma, D. Govind, J. Mishra, A. K. Dubey, K. T. Deepak и S. R. M. Prasanna. “Milestones in speaker recognition”. В: *Artificial Intelligence Review* 57.3 (2024), с. 58. ISSN: 1573-7462. DOI: 10.1007/s10462-023-10688-w. URL: https://doi.org/10.1007/s10462-023-10688-w.

[33] David Snyder, Guoguo Chen и Daniel Povey. *MUSAN: A Music, Speech, and Noise Corpus*. arXiv:1510.08484v1. 2015. eprint: 1510.08484.

[34] Zhida Song, Liang He, Penghao Wang, Ying Hu и Hao Huang. “Introducing Multilingual Phonetic Information to Speaker Embedding for Speaker Verification”. В: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, с. 10091—10095. DOI: 10.1109/ICASSP48485.2024.10446546.

[35] South African Centre for Digital Language Resources (SADiLaR). *Zulu NCHLT Speech Corpus*. https://repo.sadilar.org/handle/20.500.12185/275. Accessed: 2025-03-31. 2022.

[36] Dávid Sztahó и Attila Fejes. "Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings". В: *Journal of Forensic Sciences* 68.3 (март 2023), с. 871—883. ISSN: 1556-4029. DOI: 10.1111/1556-4029.15250. URL: http://dx.doi.org/10.1111/1556-4029.15250.

[37] Ville Vestman, Kong Aik Lee и Tomi H. Kinnunen. *Neural i-vectors*. 2020. arXiv: 2004.01559 [eess.AS]. URL: https://arxiv.org/abs/2004.01559.

[38] Feng Wang, Jian Cheng, Weiyang Liu и Haijun Liu. "Additive Margin Softmax for Face Verification". В: *IEEE Signal Processing Letters* 25.7 (июль 2018), с. 926—930. ISSN: 1558-2361. DOI: 10.1109/lsp.2018.2822810. URL: http://dx.doi.org/10.1109/LSP.2018.2822810.

[39] Shuai Wang, Zhengyang Chen, Kong Aik Lee, Yanmin Qian и Haizhou Li. *Overview of Speaker Modeling and Its Applications: From the Lens of Deep Speaker Representation Learning*. 2024. arXiv: 2407.15188 [eess.AS]. URL: https://arxiv.org/abs/2407.15188.

[40] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala и Tsubasa Ochiai. *ESPnet: End-to-End Speech Processing Toolkit*. 2018. arXiv: 1804.00015 [cs.CL].

[41] Ahmad Zairi Zaidi, Chun Yong Chong, Zhe Jin, Rajendran Parthiban и Ali Safaa Sadiq. "Touch-based continuous mobile device authentication: State-of-the-art, challenges and opportuniti B: *Journal of Network and Computer Applications* 191 (2021), с. 103162. ISSN: 1084-8045. DOI: https://doi.org/10.1016/j.jnca.2021.103162. URL: https://www.sciencedirect.com/science/article/pii/S1084804521001740.

[42] Li Zhang, Ning Jiang, Qing Wang, Yue Li, Quan Lu и Lei Xie. *Whisper-SV: Adapting Whisper for Low-data-resource Speaker Verification*. 2024. arXiv: 2407.10048 [cs.SD]. URL: https://arxiv.org/abs/2407.10048.

[43] Yiyang Zhao, Shuai Wang, Guangzhi Sun, Zehua Chen, Chao Zhang, Mingxing Xu и Thomas Fang Zheng. *Whisper-PMFA: Partial Multi-Scale Feature Aggregation for Speaker Verification*

using Whisper Models. 2024. arXiv: 2408.15585 [cs.SD]. URL: https://arxiv.org/abs/2408.15585.