ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»
Факультет Санкт-Петербургская школа
физико-математических и компьютерных наук

Усольцев Никита Викторович

Разделение физиологических и стилистических характеристик речи для повышения качества систем верификации диктора

Бакалаврская работа

Допущена к защите.
Зав. кафедрой:
д. ф.-м. н., профессор Омельченко А. В.

Научный руководитель:
к. т. н., доцент Шуранов Е. В.

Консультант:
ООО "Техкомпания Хуавэй" Матвеева Ю. А.

Рецензент:
ПАО "Сбербанк России" Шахин З.

Санкт-Петербург
2023

Оглавление

Аннотация **3**

Abstract **4**

Введение **5**

1. Обзор литературы **9**
1.1. Метрика качества системы верификации диктора 10
1.2. Нейросетевые архитектуры кодировщиков для задачи верификации . . 11
1.2.1. X-vector 11
1.2.2. GE2E 13
1.2.3. ECAPA-TDNN 15
1.3. Неустойчивость моделей верификации на стилистических данных 16
1.4. Disentangelment 18
1.5. Выводы 19

2. Наборы данных с разметкой по эмоциям **20**
2.1. VoxCeleb 20
2.2. CREMA-D 21
2.3. MSP-Podcast 25
2.4. IEMOCAP 28
2.5. EmoV-DB 30
2.6. Выводы 32

3. Эксперименты **33**
3.1. Неустойчивость моделей верификации на стилистических данных 33
3.2. Конфигурация экспериментов 35
3.3. Дообучение 36
3.4. Состязательное обучение: слой обратных градиентов 36
3.5. Выводы 37

4. Предложения по улучшению **38**

Заключение **39**

Список литературы **40**

Аннотация

Модель верификации пользователя играет важную роль в области биометрической аутентификации и систем безопасности. Однако, существует проблема неустойчивости моделей, которая возникает при работе с экспрессивными данными или при наличии различных факторов, влияющих на процесс верификации. Для решения этой проблемы предлагается использовать методы разделения информации. Они позволяют выделить и отделить различные аспекты и характеристики пользователя, например такие как физиологические и эмоциональные данные. Это позволяет модели более точно определить личность пользователя. Одним из методов, применяемых для улучшения моделей верификации пользователя, является состязательное обучение. В этом подходе модель верификации обучается в соперничестве с дискриминатором, который старается различить эмоции. Результаты экспериментов показывают, что применение метода состязательного обучения улучшает производительность моделей верификации пользователя на экспрессивных данных, но ухудшает на менее эмоциональных данных.

Ключевые слова: модель верификации пользователя, разделение информации и состязательное обучение.

Abstract

User verification models play a crucial role in biometric authentication and security systems. However, there is a challenge of model instability when dealing with expressive data or in the presence of various factors that impact the verification process. To address this issue, the use of information separation methods is proposed. These methods aim to identify and separate different aspects and characteristics of the user, such as physiological and emotional data, enabling the model to more accurately determine the user's identity. One of the methods used to enhance user verification models is adversarial learning. In this approach, the verification model is trained in competition with a discriminator that attempts to differentiate emotions. Experimental results demonstrate that the application of adversarial learning improves the performance of user verification models on expressive data but degrades performance on less emotional data.

Keywords: user verification model, disentanglement, adversarial learning.

Введение

Актуальность работы

Биометрические системы верификации пользователей обладают неотъемлемой актуальностью и значимостью в современной информационной среде. Эти системы представляют собой эффективные механизмы, основанные на уникальных физиологических или поведенческих характеристиках индивидов, таких как отпечатки пальцев, голос, лицо или сетчатка глаза. Их применение обусловлено несколькими ключевыми факторами:

1. Высокий уровень безопасности: биометрические системы верификации пользователей обеспечивают высокий уровень безопасности. Благодаря использованию уникальных биологических атрибутов каждого индивида, эти системы в достаточной степени предотвращают возможность подделки или копирования данных, обеспечивая тем самым надежную защиту от мошенничества и несанкционированного доступа.

2. Удобство и простота использования: для осуществления процедуры аутентификации пользователь лишь требуется предоставить свою биометрическую информацию, например, путем сканирования отпечатков пальцев или произнесения определенной фразы. Такой подход является интуитивно понятным и удобным для пользователей, снижая необходимость запоминания сложных паролей или использования физических устройств.

3. Неотъемлемость от пользователя: биометрические характеристики невозможно потерять или забыть, поскольку они присутствуют у пользователя всегда в отличие от паролей или устройств аутентификации.

Системы верификации пользователя по голосу являются одной из разновидностей биометрических систем, которые используют голосовые характеристики для идентификации и аутентификации индивидуальных пользователей. Голос является уникальным биометрическим атрибутом каждого человека, поскольку он обусловлен физиологическими особенностями строением голосовых органов.

В настоящее время системы верификации пользователя по голосу широко применяются в различных областях, включая телекоммуникации, информационную безопасность, управление доступом и другие. Они обеспечивают достаточный уровень безопасности и удобства, так как не требуют использования физических устройств (таких как карты или пин-коды) для идентификации пользователя. Конечно уровень безопасности таких систем ниже, чем систем верификации пользователя по лицу или

сетчатки глаза, но при этом стоимость оборудования для верификации по голосу значительно ниже. Таким образом существует компромисс между уровням безопасности и стоимостью оборудования для верификации.

Задачи верификации пользователя по голосу можно разделить на два типа: текстонезависимая независимые от текста(text-independent) и зависимые от текста (text-dependent). Отличие между этими двумя типами задач заключается в способе проведения верификации.

Независимая от текста верификация не требует определенного текста или фразы для проведения верификационного процесса. Диктору предоставляется возможность произнести любую фразу или даже отсутствующий текст, и система должна определить, соответствует ли этот голос зарегистрированному голосу пользователя. Такая задача являются более гибкими, поскольку они не ограничивают диктора определенным набором фраз.

Зависимая от текста верификация требует от диктора произнесения определенного текста или фразы, которая предварительно известна системе. Например, это может быть фраза "Привет, Алиса". Такой тип задачи предоставляют большую степень контроля, поскольку требуют определенного текста, что облегчает обучение нейросетевой модели. В данном случае, система будет сравнивать голос диктора с его предварительно зарегистрированным голосом для указанной фразы.

Различия между независимые и зависимыми от текста задачами верификации пользователя по голосу связаны с требованиями к входным данным для проведения верификации. Каждый из этих типов задач имеет свои преимущества и ограничения, и выбор между ними зависит от конкретной ситуации и требований к системе верификации. В данной работе будет уделено внимание системам верификации независимым от текста.

Если анализировать речевой сигнал с точки зрения характеристик речи конкретного пользователя, то он представляет собой сложную комбинацию физиологических характеристик речи и других стилистических характеристик (Рис. 1):

Физиологические характеристики речи, такие как тембр голоса, определяются уникальными особенностями голосового аппарата каждого человека. Такие характеристики могут использоваться для верификации благодаря своей биометрической надежности.

К стилистическим характеристикам речи можно отнести скорость речи, высоту голоса, эмоциональность речи. Они представляют собой информацию, которая не является определяющей для верификации пользователя. В основном такие характеристики определяются текущем контекстом человека: эмоциональным состоянием или уровнем стресса. Эти аспекты речи имеют значение для понимания человека. Стилистическим характеристики делают речевой сигнал более богатым, но также создает сложности при анализе речевого сигнала с точки зрения индивидуальности пользо-

вателя.

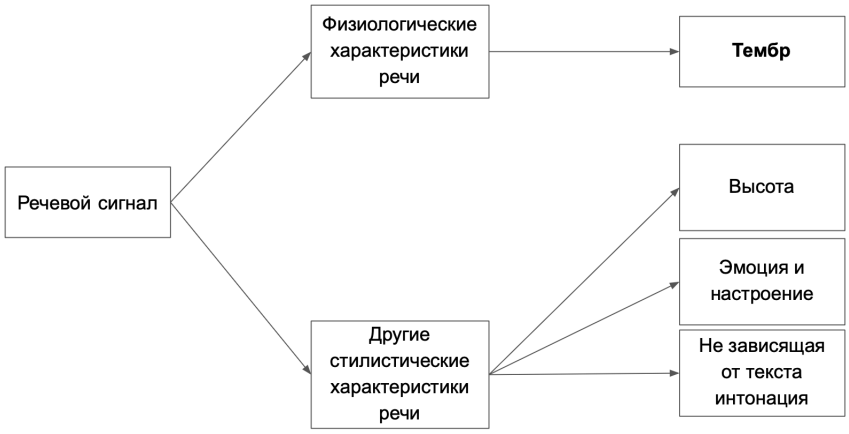

Рис. 1: Структура речевого сигнала через характеристики речи

Все эти факторы кодируются в речевом сигнале при верификации пользователя, хотя стилистическая информация является шумом для систем распознавания дикторов [5] [17] [27]. Разделение физиологических характеристик от стилистических характеристик представляют сложности, поскольку они тесно взаимодействуют друг с другом в речевом сигнале. Что создаёт проблемы в разработке эффективных систем верификации диктора.

Цель и задачи

Цель данного исследования заключается в улучшении модели верификации пользователя путем разделения физиологической и эмоциональной информации в векторном представлении.

Для достижения этой цели поставлены следующие задачи:

- Составить набор данных на основе открытых данных, содержащий информацию об эмоциях
- Подтвердить проблему неустойчивости модели при верификации на экспрессивном наборе данных
- Дообучить актуальную предобученную модель на составленном корпусе данных для верификации дикторов
- Реализовать состязательное обучение с использованием слоя обратных градиентов для улучшения модели верификации

Достигнутые результаты

- Составлен набор данных, включающий 9 эмоций и 1460 дикторов, для проведения дообучения и состязательного обучения.
- Подтверждена проблема неустойчивости классических моделей верификации на экспрессивных данных. Относительные изменения в производительности моделей составили от 6% до 670%.
- Реализовано дообучение и состязательное обучение с использованием слоя смены направления градиента. Это привело к улучшению качества модели на эмоциональных наборах данных.
- Выявлена проблема переобучения на менее экспрессивных и нейтральных данных.

Структура работы

В первой главе представлен обзор литературы, включающий основные архитектуры кодировщиков, используемые в моделях верификации, а также методы разделения (disentanglement) различных типов информации. Во второй главе подробно описывается используемый в данной работе набор данных, а также его разбиение на тренировочную, тестовую и валидационную выборки. В третьей главе описываются все проведенные эксперименты, включая верификацию проблемы, проведение дообучения и состязательного обучения.

1. Обзор литературы

Система верификации пользователя обеспечивает процесс проверки подлинности голоса пользователя. Она состоит из двух этапов: регистрация и верификация. Рассмотрим каждый этап более подробно:

1. Регистрация диктора в системе (стадия Enroll) рис. 2: В этом этапе диктор записывает несколько примеров своего голоса. Записанные аудиофрагменты проходят через кодировщик, который преобразует голосовую информацию в числовые векторы. Эти векторы представляют собой характеристики, извлеченные из голоса диктора. Затем система строит центройду путем агрегации полученных векторов, например, вычисляя среднее значение всех векторов. Центройда представляет собой усредненное представление голосовых характеристик диктора, которое будет использоваться для сравнения с векторами, полученными во время верификации.

2. Верификация диктора в системе (стадия Test) рис. 3: На этом этапе пользователь записывает свой голос для входа в систему. Запись проходит через тот же кодировщик, который возвращает текущий вектор, представляющий голос пользователя. Далее происходит оценка меры близости (например, косинусное расстояние или евклидово расстояние) между текущим вектором и ранее построенной центройдой. Более высокая мера близости указывает на большую схожесть голосовых характеристик пользователя с регистрированным диктором. Наконец, на основе оценки меры близости принимается решение о верификации. Пороговое значение может быть установлено, и если мера близости превышает этот порог, то голос пользователя считается верифицированным, и пользователю предоставляется доступ к системе. В противном случае, если мера близости не достигает порога, голос пользователя не считается достаточно близким к регистрированному диктору, и доступ может быть отклонен.

Эта система верификации пользователя по голосу предоставляет простой и эффективный способ проверки подлинности голоса для аутентификации пользователей и обеспечения безопасного доступа к системе.

Спектрограмма

Перед тем, как приступить к обзору литературы, представим несколько определений ключевых терминов, которые будут использоваться в данном исследовании:

Спектрограмма - это графическое представление спектрального содержания звукового сигнала в зависимости от времени. Она представляет собой двухмерное изображение, где по оси X откладывается время, а по оси Y - частота. Яркость или цветовая интенсивность на спектрограмме отражает амплитуду звука в каждом частотном

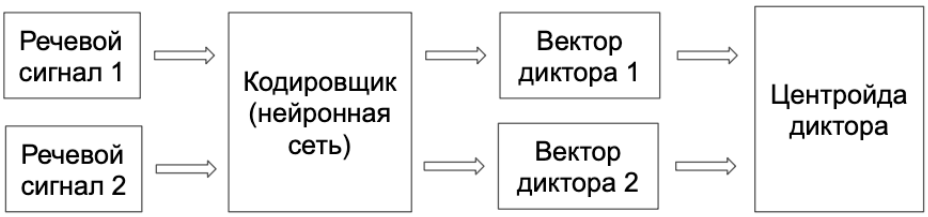

Рис. 2: Этап 1. Регистрации диктора в системе верификации

Рис. 3: Этап 2. Верификации диктора в системе

диапазоне на протяжении времени. Спектрограммы широко используются для анализа и визуализации звуковых сигналов, так как они позволяют увидеть изменения в спектральном содержании звука в течение времени. Перед подачей речевого сигнала в кодировщик, необходимо извлечь из сигнала признаки, например это может быть спектрограмма.

Срез спектрограммы представляет собой фрагмент спектрограммы, ограниченный по времени. В случае анализа звукового сигнала, срез спектрограммы представляет собой небольшой участок времени, на котором производится измерение спектральных характеристик звука. Обычно срезы спектрограммы имеют фиксированную длительность и накладываются друг на друга с определенным перекрытием, чтобы покрыть всё время записи. С помощью срезов спектрограммы можно анализировать динамику спектрального содержания звука в течение времени и извлекать характеристики для последующей обработки и распознавания.

1.1. Метрика качества системы верификации диктора

Equal Error Rate (EER) - это метрика, используемая для оценки производительности системы верификации диктора. Она является точкой, в которой вероятность ошибки первого рода (False Rejection Rate - FRR) и вероятность ошибки второго рода (False Acceptance Rate - FAR) совпадают (Рис. 4).

Ошибка первого рода представляет собой долю ложноотрицательных результатов, то есть случаи, когда система отклоняет пользователя, который зарегистрирован в нашей системе.

Ошибка второго рода представляет собой долю ложноположительных результа-

тов, то есть случаи, когда система верифицирует человека, который никогда не регистрировался в нашей системе.

EER достигается в тот момент, когда FAR и FRR равны друг другу. Это означает, что система достигает оптимального баланса между отклонением ложноположительных и ложноотрицательных результатов. Чем ниже значение EER, тем лучше производительность системы верификации диктора.

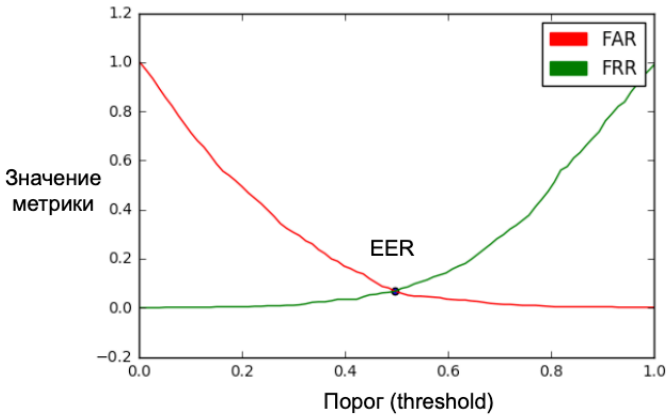

Рис. 4: Equal Error Rate

1.2. Нейросетевые архитектуры кодировщиков для задачи верификации

До появления нейросетевых архитектур, одним из наиболее эффективных решений в области верификации диктора по голосу была архитектура i-vector [11], основанная на методе гауссовых смесей (GMM). Метод гауссовых смесей (Gaussian Mixture Model, GMM) является статистическим методом моделирования вероятностных распределений данных. Он широко используется в различных областях, включая распознавание образов, анализ речи, компьютерное зрение и биометрию. С появлением глубоких нейронных сетей в области верификации диктора по голосу были представлены новые подходы, обладающие большей точностью.

1.2.1. X-vector

Архитектура X-vector [6] [21] является нейронной сетевой моделью, разработанной для задачи верификации диктора по голосу. На вход принимает T срезов спектрограммы, каждый из которых проходит через 5 слоев сети с архитектурой временной задержки (TDNN) [18].

TDNN (Time-Delay Neural Network) - это архитектура нейронных сетей, специально разработанная для обработки последовательностей, таких как речевые сигналы. Она широко применяется в области обработки речи, включая задачи распознавания и

верификации диктора. Основная идея TDNN заключается в использовании одномерной операций свертки (Рис. 5) и пулинга для захвата локальной контекстной информации в последовательности. Сверточный слой, который применяет фильтры с определенными временными задержками к последовательности входных данных. Каждый фильтр захватывает информацию из разных временных окон и выдает выходной вектор, который представляет собой сумму активаций от всех временных точек, на которые фильтр был применен. Преимущества TDNN включают его способность улавливать зависимости во временной оси и моделировать контекстную информацию в последовательностях, таких как речь. Он также обладает хорошей инвариантностью к изменениям во времени и позволяет обрабатывать последовательности разной длины. Это делает его эффективным инструментом для задач верификации диктора

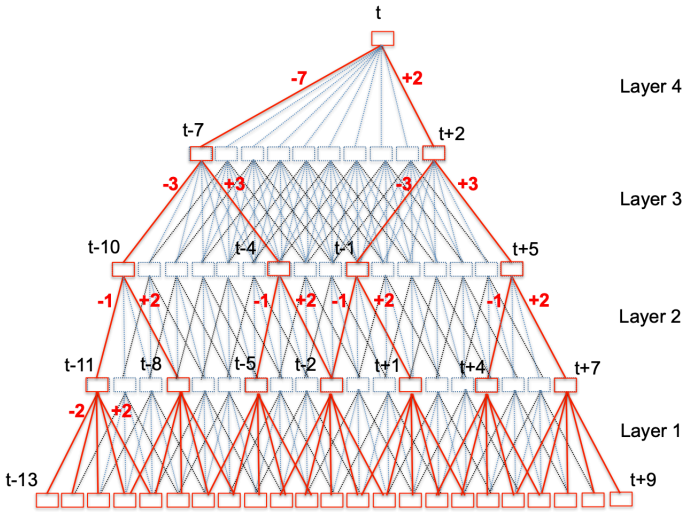

Рис. 5: Одномерная свертка из TDNN

После пропускания каждого фрейма через TDNN, получается T выходов. Далее эти выходы агрегируются с помощью Statistics Pooling, где вычисляются статистические характеристики (например, среднее и дисперсия) по временной оси, что позволяет учесть долгосрочные зависимости в голосовом сигнале.

Затем полученный результат подается на два линейных слоя с функцией активации ReLU, которые помогают извлечь более высокоуровневые признаки из агрегированного вектора. Наконец, на выходе применяется softmax-слой, который преобразует признаки в вероятности принадлежности к различным классам.

В процессе обучения X-vector используется для решения задачи классификации на N классов, где N соответствует количеству дикторов в обучающей выборке. Однако, для получения векторного представления пользователя, не присутствующего в обучающей выборке, используются два варианта (рис 6):

1. Вектор a: Берется вектор после применения статистического пулинга (Statistics

Pooling) к выходам нейронной сети.

2. Вектор b: Берется вектор после первого линейного слоя с применением нелинейности.

На этапе верификации для подсчета меры близости веторов используется предварительно обученная модель PLDA (Probabilistic Linear Discriminant Analysis). Сначала векторы центрируются, затем происходит уменьшение размерности до 25% от исходной длины с использованием LDA (Linear Discriminant Analysis). Векторы нормализуются по длине и затем сравниваются с помощью PLDA, при этом оценка PLDA также нормализуется с использованием s-нормы.

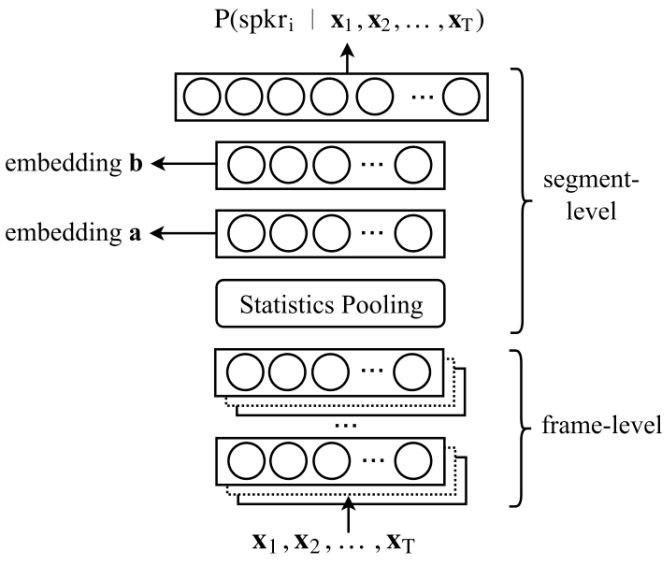

Рис. 6: Архитектура нейронной сети X-vector

1.2.2. GE2E

Архитектура GE2E (Generalized End-to-End) [12] представляет собой нейронную сеть, основанную на использовании специальной функции потерь, названной Softmax loss. Ключевая идея этой статьи Softmax loss - функция потертъ, которая приближает вектора своего диктора к центройде диктора и отдаляет его от центройд других дикторов. Идея похожа на Triplet loss. Общий процесс обучения GE2E можно описать следующим образом (Рис. 7):

Размер батча составляет $N \times M$ последовательностей, где N - количество дикторов, а M - количество аудиофайлов или последовательностей для каждого диктора. Каждый вектор обозначается как $x_{j,i}$, где j соответствует диктору, а i - номеру последовательности. Каждый вектор $x_{j,i}$ проходит через кодировщик, который состоит из

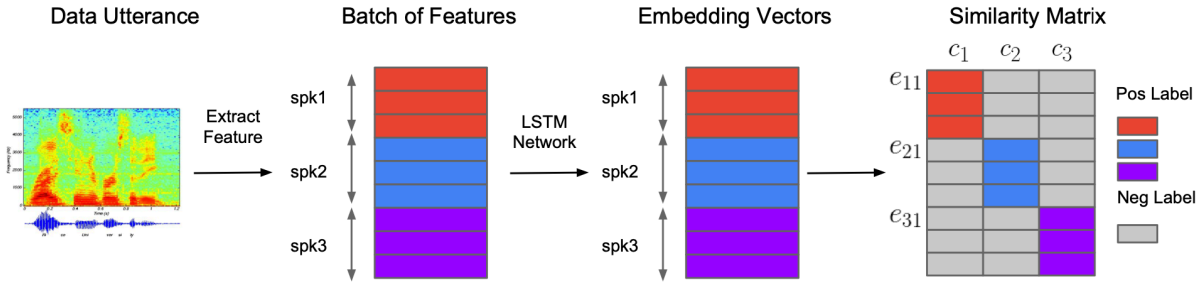

Рис. 7: Обзор системы GE2E. Разные цвета обозначают высказывания/вектора от разных дикторов.

LSTM (Long Short-Term Memory) модуля и полно связного слоя. Кодировщик извлекает характеристики из аудиофайлов и преобразует их в вектора. Полученные вектора нормализуются по L2-норме, что позволяет стандартизировать их длину (Формула 1).

$$\mathbf{e}_{ji} = \frac{f\left(\mathbf{x}_{ji}; \mathbf{w}\right)}{\left\| f\left(\mathbf{x}_{ji}; \mathbf{w}\right) \right\|_2} \tag{1}$$

Центройды дикторов строятся на основе полученных нормализованных векторов признаков. Формула 2 используется для вычисления центройдов, где каждый центройд представляет собой усредненный вектор признаков для соответствующего диктора.

$$\mathbf{c}_k = E_m\left[\mathbf{e}_{km}\right] = \frac{1}{M}\sum_{m=1}^{M}\mathbf{e}_{km} \tag{2}$$

После построения центройдов дикторов, в архитектуре GE2E строится матрица схожести с использованием формулы 3. Здесь w и b представляют обучаемые параметры модели. Матрица схожести представляет собой матрицу размером $N \times M$, где каждый элемент матрицы отражает степень схожести между вектором вложения и центройдом диктора.

$$\mathbf{S}_{ji,k} = w \cdot \cos\left(\mathbf{e}_{ji}, \mathbf{c}_k\right) + b \tag{3}$$

Вводится функция потерь, представленная формулой 4. Эта функция потерь приближает каждый вектор вложения к соответствующему центройду диктора и удаляет его от всех остальных центройдов (Рис. 8). Это позволяет улучшить различимость векторов вложения и повысить качество верификации диктора.

$$L\left(\mathbf{e}_{ji}\right) = -\mathbf{S}_{ji,j} + \log\sum_{k=1}^{N}\exp\left(\mathbf{S}_{ji,k}\right) \tag{4}$$

Таким образом, архитектура GE2E позволяет эффективно извлекать пользовательские характеристики из аудиофайлов и использовать их для задачи верификации

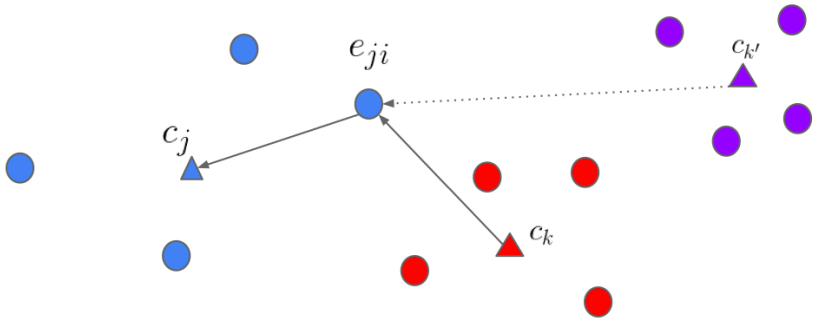

Рис. 8: Функция потерь GE2E толкает вложение к центроиду истинного динамика, а также от центроида самого похожего говорящего.

диктора по голосу с помощью специальной функции потерь и матрицы схожести. Это приводит к повышению точности и надежности системы верификации диктора.

1.2.3. ECAPA-TDNN

ECAPA-TDNN [7] является модификацией архитектуры x-vector. Она превосходит x-vector благодаря нескольким ключевым усовершенствованиям:

1. Увеличение размера контекста: ECAPA-TDNN использует блоки Squeeze-Excitation (SE) [13], чтобы расширить рецептивное поле модели и включить больший контекст. Это помогает модели учесть больше контекстуальной информации и улучшить ее способность улавливать долгосрочные зависимости.

2. Остаточные соединения: Вдохновленная архитектурой ResNet, ECAPA-TDNN включает остаточные соединения, которые позволяют более эффективно передавать градиенты по сети. Это не только облегчает обучение, но также позволяет строить более глубокие модели с большим количеством параметров, что потенциально улучшает представительную способность сети.

3. Пулинг со статистикой, зависящей от канала: ECAPA-TDNN вводит новый подход, называемый пулингом со статистикой, зависящей от канала. Он расширяет механизм временного внимания на измерение каналов, позволяя сети выбирать специфичные для канала характеристики диктора. Например, модель может различать дикторо-специфичные свойства гласных звуков и дикторо-специфичные свойства согласных звуков. Это расширение улучшает дискриминирующую способность модели, обращая внимание на специфичные для канала информацию о дикторе.

В целом, эти улучшения в ECAPA-TDNN способствуют ее превосходству по сравнению с традиционной архитектурой x-vector. За счет включения большего контекста,

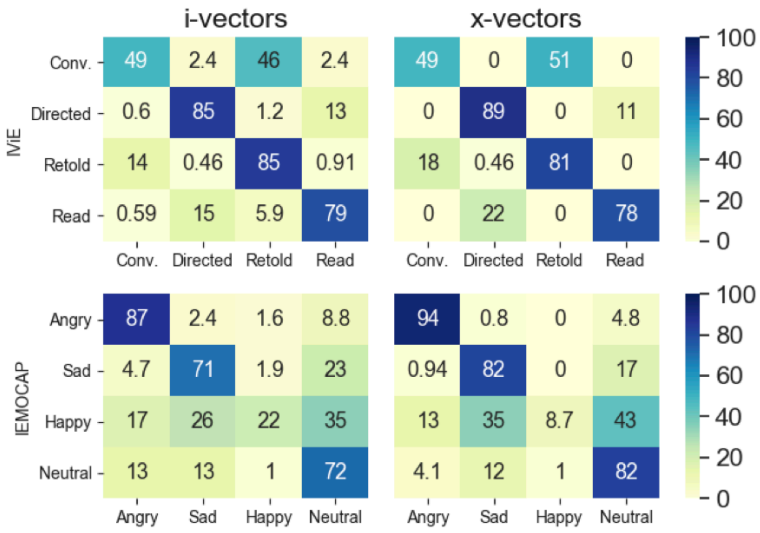

Рис. 9: Матрица неточности(confusion matrix): процент примеров, классифицированных по категориям

использования остаточных соединений и введения пулинга со статистикой, зависящей от канала, ECAPA-TDNN обладает расширенными возможностями моделирования для задач распознавания дикторов.

1.3. Неустойчивость моделей верификации на стилистических данных

Содержание информации об эмоциях в векторах из кодировщика системы верификации пользовотеля

В данной статье [20] было показано, что два типа представления на уровне векторов(I-векторы и X-векторы) содержат информацию, предсказывающую стиль и эмоциональную окраску. Аудио файлы пропустили через кодировщик(I-vector и X-vector) и полученные вектора классифицировали на эмоции. Матрица неточности (Рис. 9) показывает хорошие результаты классификации на обоих наборах данных IEMOCAP и IViE. Этот результат подтверждает существование стилевых факторов в векторах представления дикторов для верификации пользователя. В этой работе был сделал вывод, что разделение таких факторов было бы чрезвычайно полезно во многих приложениях речи, включая речевой перевод, синтез речи и верификацию диктора. Этот вывод подтверждает актуальность моей работы.

Неустойчивость моделей верификации на стилистических данных

В исследовании [27] авторы изучают связь между эмоциями и производительностью системы распознавания диктора с использованием архитектуры X-vector в каче-

стве кодировщика диктора. В работе исследуется влияние эмоциональной изменчивости на распознавание диктора. Исследование показало, что существует значительная связь между эмоциями и производительностью распознавания диктора: эмоциональная изменчивость оказывает отрицательное влияние на метрику верификации диктора EER. Предобученная модель x-vector регистрировала(стадия enroll) нейтральные записи, а верифицировалась(стадия test) на 4 множествах: нейтральном, злом, счастливом и грустном. Эксперименты проводились на 3 корпусах: CREMA-D, MSP-Podcast и IEMOCAP. Результаты представлены в таблицах 1, 2, 3.

Анализ влияния эмоций на модель верификации диктора показал, что модель чрезвычайно чувствительны к изменениям эмоционального состояния дикторов на этапе верификации. В работе было обнаружено, что конфигурация с нейтральной регистрацией и нейтральной верификацией показывают более лучшие результаты по сравнению с другими. Кроме того, было обнаружено, что злые эмоции на этапе верификации больше остальных ухудшают результат. Это исследование подчеркивает важность решения проблемы кодирования стилевой информации в эмбеддингах верификации диктора.

Test / Enroll	Злость	Счастье	Грусть	Нейтраль
Нейтраль	32.36	31.08	28.43	26.92

Таблица 1: CREMA-D: EER для системы верификации диктора

Test / Enroll	Злость	Счастье	Грусть	Нейтраль
Нейтраль	44.35	43.2	43.27	39.4

Таблица 2: IEMOCAP: EER для системы верификации диктора

Test / Enroll	Злость	Счастье	Грусть	Нейтраль
Нейтраль	12.98	11.63	11.89	8.95

Таблица 3: MSP-Podcast: EER для системы верификации диктора

В исследовании, описанном в работе [17], было исследовано влияние шепота на метрики верификации диктора. Результаты показали, что во всех экспериментальных условиях Equal Error Rate для нейтральной тестовой выборки значительно ниже, чем для речи с шепотом. Это свидетельствует о том, что вектора шепота из кодировщика, значительно отличаются от нейтральных векторов в терминах косинусного сходства.

1.4. Disentangelment

В работе [9] исследовалось обучение двух векторов: вектора стиля и вектора пользователя, с целью улучшения результатов распознавания дикторов. Обучение проводилось на основе набора данных VoxCeleb1, который был размечен на эмоции с применением методом semi-supervised обучения [10]. Главной целью работы было превзойти базовые модели по метрике EER на наборе данных VoxCeleb eval. Авторам удалось достичь значительного улучшения по сравнению с базовыми моделями по метрике EER на наборе данных VoxCeleb eval. Результаты показали, что обучение двум векторам - стиля и пользователя - позволило улучшить эффективность системы. Предложенный подход дал возможность выделить и отделить атрибуты стиля и пользователя, что позволило более точно моделировать и представлять голосовые данные.

В работе [1] исследователи применили метод обратного градиента (Gradient Reversal Layer, GRL) для улучшения производительности системы верификации диктора на наборе данных Hey Cortana, содержащем шумы различных типов.

Основная идея заключалась в использовании дискриминатора, обученного на распознавание различных типов шумов, чтобы помочь модели адаптироваться и более эффективно работать с такими шумовыми условиями. Такой способ обучение приводил к тому, что модель обучалась игнорировать шум и фокусироваться на более устойчивых и дискриминативных признаках диктора. Эксперименты показали, что включение GRL привело к улучшению показателя EER. Это свидетельствует о том, что модель, обученная с использованием GRL, стала более устойчивой к шумам и способна лучше разделять голосовые характеристики диктора от внешних шумовых воздействий.

Слой обратных градиентов (Gradient Reversal Layer, GRL) - это техника, которая часто используется в задачах доменной адаптации и безнадзорной доменной адаптации. Он разработан для снижения влияния доменного сдвига, путем поощрения построения инвариантных к домену представлений признаков. GRL работает путем изменения потока градиента во время обратного распространения. Он вводит специальный слой, который инвертирует градиенты и предотвращает обновление параметров в сети. Путем включения GRL в архитектуру сети модель поощряется изучать признаки, которые одновременно являются дискриминативными для целевой задачи и устойчивыми к вариациям домена (напимер шумы или эмоции). Это позволяет модели хорошо обобщаться на невидимые данные из разных доменов, улучшая производительность и устойчивость модели в сценариях доменной адаптации.

В работе [8] исследователи предложили метод разделения физиологической информации о дикторе и информации о звукозаписывающем устройстве путем минимизации взаимной информации между ними. Для проведения экспериментов они использовали набор данных FFSVC2022, который был размечен на четыре класса: iPhone,

iPad, Android и обычный микрофон. Эксперименты показали, что предложенный метод успешно разделил физиологическую информацию о дикторе от информации о звукозаписывающем устройстве. Таким образом, полученные представления дикторов были более чистыми и свободными от влияния устройства записи, что может быть полезным в различных приложениях, таких как системы идентификации дикторов и распознавания речи.

1.5. Выводы

В ходе обзора литературы были рассмотрены различные исследования, связанные с областью распознавания дикторов и анализа речи. Одним из ключевых результатов обзора было выявлено, что стилистическая изменчивость оказывает значительное влияние на метрики распознавания дикторов. Также было отмечено, что использование дополнительных данных, таких как информация о стиле, о шуме или о звукозаписывающем устройстве, может привести к улучшению производительности моделей распознавания дикторов.

2. Наборы данных с разметкой по эмоциям

В данной главе представлен обзор и анализ набора данных для экспериментов в этой работе.

В соответствии с поставленной научной целью, требуется использовать размеченные данные, содержащие информацию о эмоциональных выражениях и идентификации дикторов. В таблице 4 представленная краткая информация о наборах данных используемых в моей работе в качестве тренировочной, валидационной и тестовой выборки.

Название	Количество дикторов	Количество записей	Количество эмоций
MSP-Podcast	1355	62140	8
CREMA-D	91	7442	7
IEMOCAP	10	7527	5
EmoV-DB	4	6653	3

Таблица 4: Краткая информация о наборах данных используемых при обучение, валидации и тестирование

2.1. VoxCeleb

Набор данных VoxCeleb является одним из наиболее широко используемых и популярных наборов данных для задач верификации диктора по голосу и распознавания диктора. Под VoxCeleb я понимаю объединение VoxCeleb1[16] и VoxCeleb2[3] . VoxCeleb представляет собой масштабный набор данных по распознаванию дикторов, полученный автоматически из открытых медиа-ресурсов. Среди них присутствуют интервью с красных ковров, выступления перед большой аудиторией, записи на открытых стадионах и в тихих закрытых студиях, отрывки из профессионально снятых мультимедийных материалов, а также даже примитивные видеозаписи, снятые на ручные устройства. Набор данных VoxCeleb2 состоит из более чем миллиона фраз от более чем 7 тысяч дикторов. Поскольку набор данных собран ”в естественных условиях”, речевые сегменты коррелируют с реальными звуками окружающей среды, включая смех, перекрестные разговоры, эффекты канала передачи, музыку и другие звуки. Набор данных также является многоязычным, охватывая широкий спектр акцентов, возрастов, этнических групп и языков. VoxCeleb является сбалансированным по гендеру, потому что в VoxCeleb2 61% дикторов являются мужчинами. Набор данных является аудиовизуальным, поэтому также полезен для ряда других приложений. VoxCeleb предоставляет возможность исследователям и разработчикам различных методов и алгоритмов проводить эксперименты, увеличивать производительность своих моделей и оценивать свои модели на этом наборе данных. Благодаря

своей широкой покрытию и разнообразию дикторов, VoxCeleb способствует развитию и улучшению систем верификации и распознавания диктора по голосу.

Набор данных VoxCeleb представляет собой значительно эмоционально окрашенный набор, содержащий множество разговоров с известными личностями. Однако отмечается отсутствие разметки данных по эмоциональному состоянию. В рамках исследования будет использоваться тестовый набор VoxCeleb1 Test, состоящий из 40 дикторов, с целью оценки качества системы распознавания пользователей.

2.2. CREMA-D

Набор данных Crema-D представляет собой мультимодальный набор данных (аудио и видео), включающий 91 профессионального актера, воплощающих целевую эмоцию для заранее заготовленного списка из 12 предложений. Набор данных был записан в студии. В набор данных входят 48 мужских и 43 женских актера с разнообразным этническим и возрастным составом: актеры в возрасте от 20 до 74 лет, представленные различными расами и этническими группами (афроамериканцы, азиаты, кавказцы, испанцы и неуказанные).

Перед рассказам про разбиение на тренировочную, валидационную и тестовую выборку важно рассказать про устройство этапа подтверждения(validation) при обучение систем распознавания пользователя, потому что этот этап накладывает ограничения на размер валидационной выборки.

Как было отмечено в литературном обзоре, при обучении модели верификации мы используем задачу классификации, где каждый диктор из обучающего набора данных представляет отдельный класс. Одной из задач валидационной выборки является определение момента, когда модель начинает переобучаться на обучающей выборке. Однако в валидационной выборке присутствуют другие дикторы (классы), и невозможно измерить классификационные метрики на этой выборке в стандартной постановке задачи классификации.

При обучении систем верификации на валидационной выборке используются метрики системы верификации, такие как EER (Equal Error Rate). Поэтому обучение следует остановить, когда EER начинает увеличиваться или изменяться незначительно в течение нескольких десятков эпох.

В данном исследовании проводилась валидация следующим образом. Имеются две валидационные выборки на основе CREMA-D и MSP-Podcast. В каждом из этих наборов данных мы создали наборы всех возможных пар без учета порядка на основе аудиозаписей. Из комбинаторики ясно, что количество таких пар равно $\frac{n \cdot (n-1)}{2}$, где n - число аудизаписей в наборе данных. Первый элемент каждой пары соответствует вектору регистрации, а второй элемент - вектору верификации. Таким образом, мы можем рассчитать EER на валидационной выборке. Необходимо отметить, что

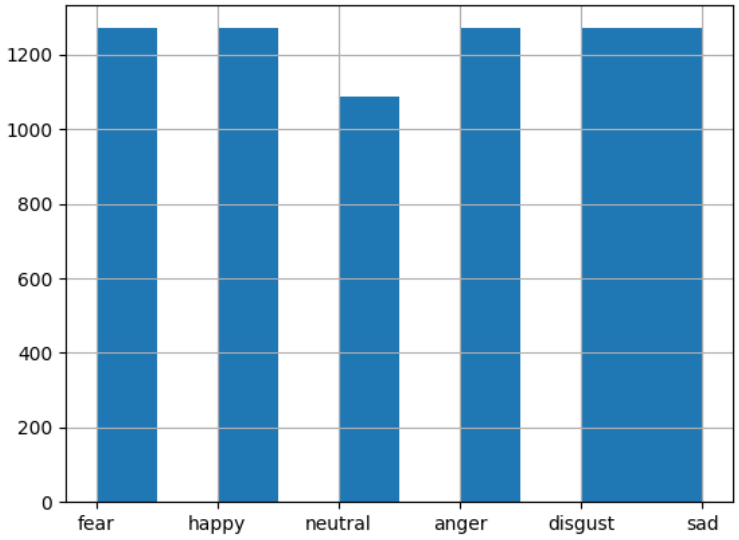

Рис. 10: CREMA-D - Столбчатая диаграмма: количества аудио записей по эмоциям

подсчет такой метрики является вычислительно затратным и требует значительного времени, что ограничивает размер валидационной выборки. Для ускорения процесса валидации были использованы матричные операции, а оценка косинусного сходства производилась в рамках каждого батча с помощью класса torch.utils.data.Dataloader. Изначально был реализован подход, при котором каждый раз пересчитывался один и тот же вектор путем прохода через кодировщик, что приводило к почти квадратичному увеличению вычислений. Однако здесь очевидно, что мы можем предварительно вычислить выход кодировщика для каждой аудиозаписи из валидационной выборки и затем загружать их из файловой системы. В результате простого процесса валидации, который изначально занимал примерно 14 минут 15 секунд, удалось сократить время до 3 минут 5 секунд.

CREMA-D размечен на 6 эмоций: нейтральность, злость, грусть, радость, отвращение, страх. На гистограммах 10 и 11 видно, что датасет хорошо сбалансирован: на каждую эмоцию приходится примерно одинаковое число аудиозаписей. Внутри каждого пользователя дататасет так же сбалансирован, что показывает гистограмма 11.

Размер тренировочной выборки составил 75 пользователей. Размер валидационной выборки составил 6 пользователей. Размер тестовой выборки составил 10 пользователей. Отдельно хочется отметить, что внутри каждой выборки у каждого диктора присутствуют всех эмоции, это видно на графика 12, 13, 14.

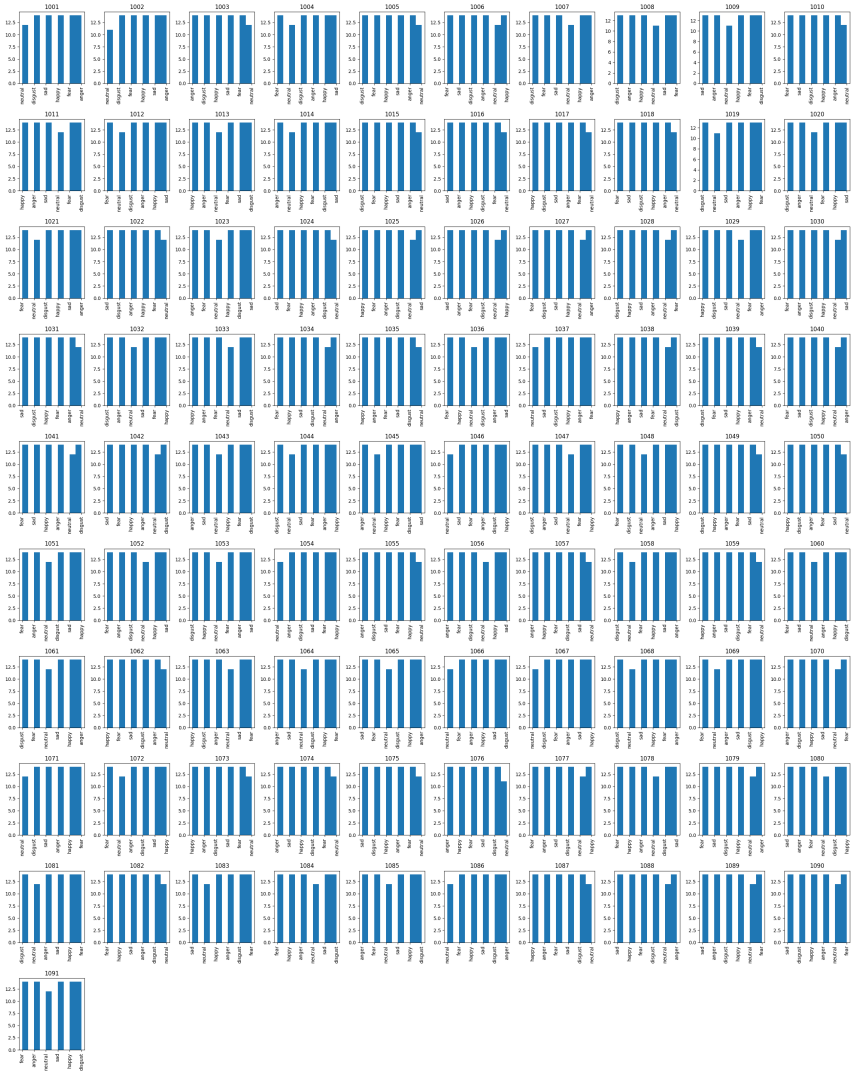

Рис. 11: CREMA-D - Столбчатая диаграмма: количества аудио записей по эмоциям для каждого диктора

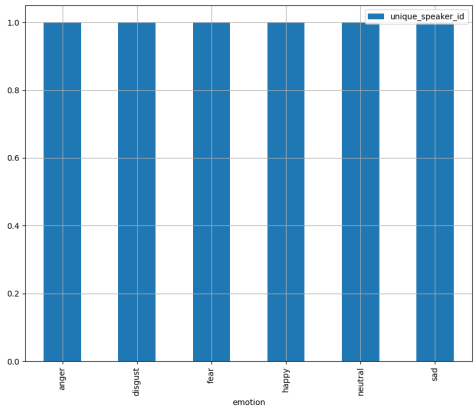

Рис. 12: CREMA-D - Столбчатая диаграмма: количества уникальных дикторов по эмоциям в тренировочной выборке в абсолютных величинах

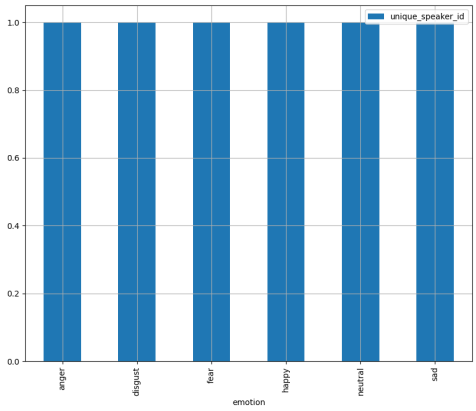

Рис. 13: CREMA-D - Столбчатая диаграмма: количества уникальных дикторов по эмоциям в валидационной выборке в абсолютных величинах

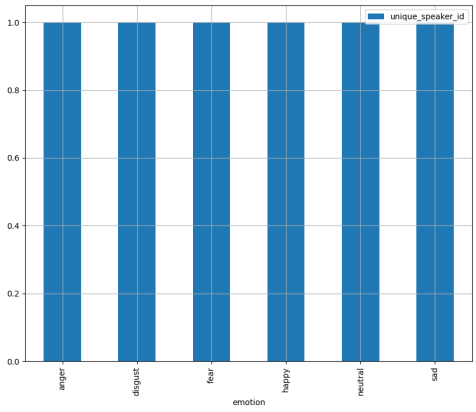

Рис. 14: CREMA-D - Столбчатая диаграмма: количества уникальных дикторов по эмоциям в тестовой выборке в абсолютных величинах

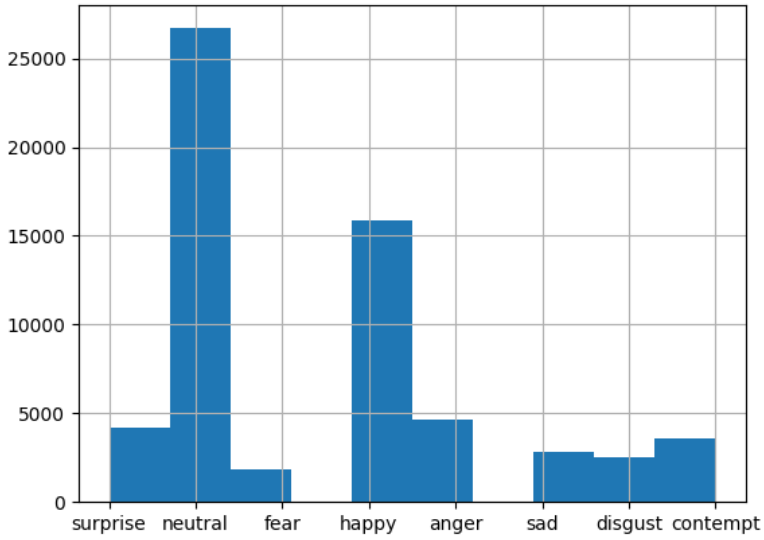

Рис. 15: MSP-Podcast - Столбчатая диаграмма: количества аудио записей по эмоциям

2.3. MSP-Podcast

Корпус MSP-Podcast [15] содержит речевые сегменты из записей студийных подкастов, которые были аннотированы с использованием краудсорсинга. Он является одним из крупнейших натуралистических речевых наборов данных размеченных на эмоции. Корпус содержит 62 140 речевых записей (100 часов). Корпус аннотирован эмоциональными метками, используя атрибутивные дескрипторы (активация, доминирование и валентность) и категориальные метки (злость, грусть, радость, отвращение, презрение, удивление, страх и нейтральность). Большинство сегментов в обычных подкастах являются нейтральными. В наборе данных представлено 1355 уникальных диктора.

MSP-Podcast размечен на 8 эмоций. На гистограмме 15 видно, что корпус плохо сбалансирован: на каждую эмоцию приходится разное число аудиозаписей. Нейтральных и счастливых аудизаписей сильно больше, чем остальных. Внутри каждой эмоции распределение дикторов по количеству аудио записей так же плохо сбалансировано, что показывает гистограмма 16.

Размер тренировочной выборки составил 1207 пользователей. Размер валидационной выборки составил 24 пользователя. Размер тестовой выборки составил 97 пользователей. Отдельно хочется отметить, что внутри каждой выборки распределение уникальных дикторов по эмоциям похожи, что важно при обучение и тестирование модели 17, 18, 19.

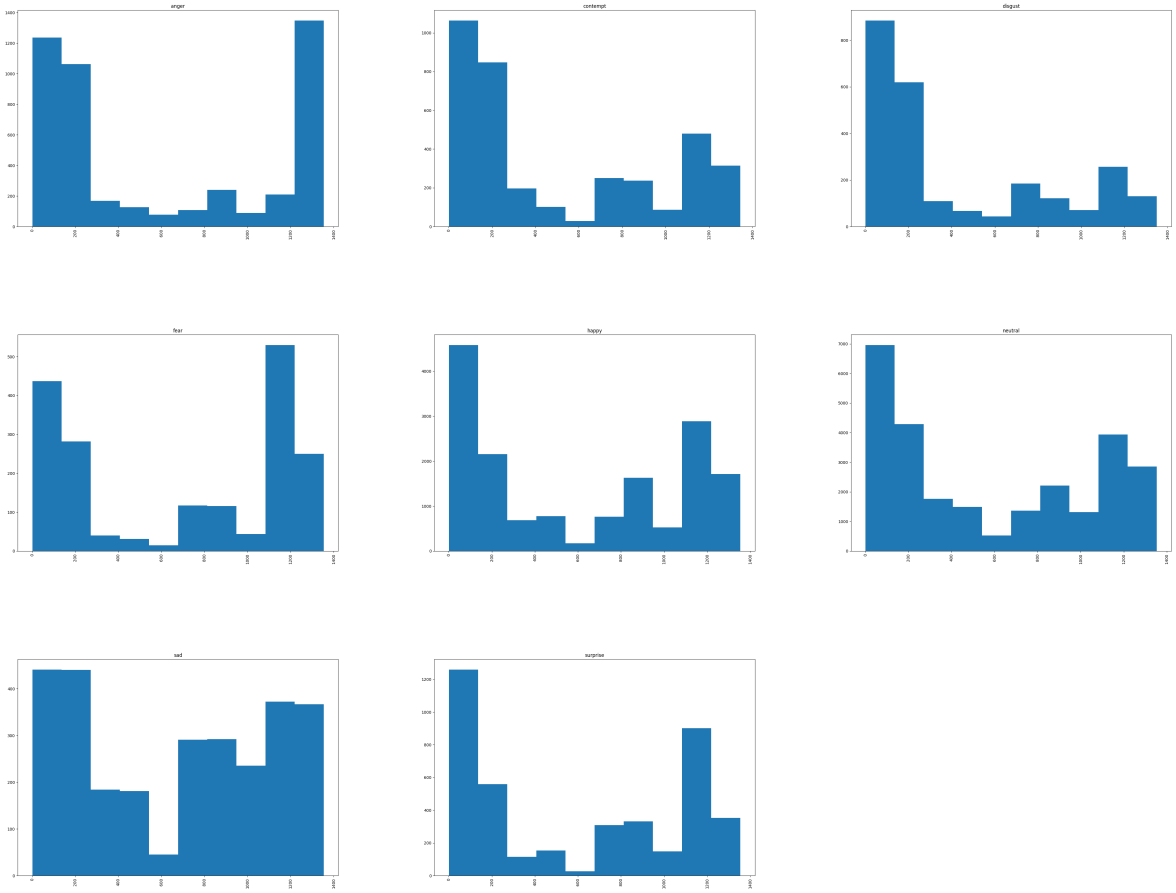

Рис. 16: MSP-Podcast - Столбчатая диаграмма: количества аудио записей по дикторам для каждой эмоции

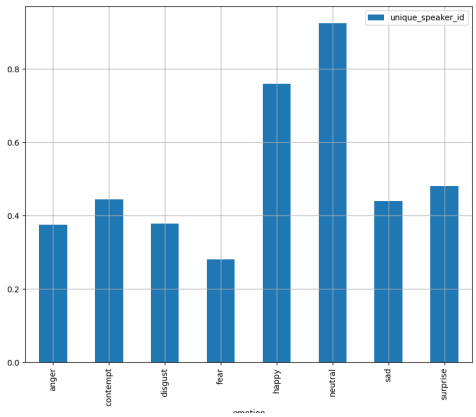

Рис. 17: MSP-Podcast - Столбчатая диаграмма: количества уникальных дикторов по эмоциям в тренировочной выборке в абсолютных величинах

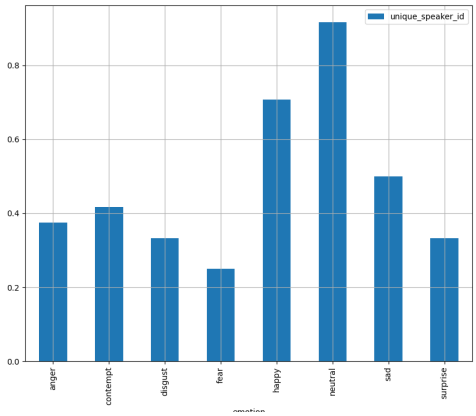

Рис. 18: MSP-Podcast - Столбчатая диаграмма: количества уникальных дикторов по эмоциям в валидационной выборке в абсолютных величинах

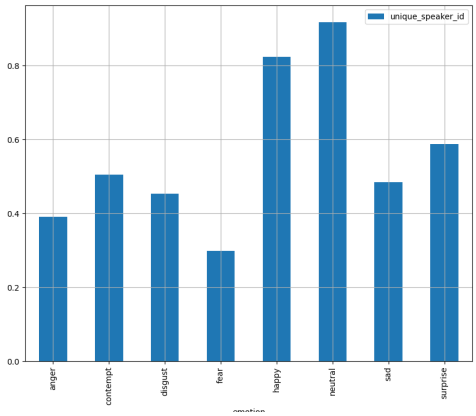

Рис. 19: MSP-Podcast - Столбчатая диаграмма: количества уникальных дикторов по эмоциям в тестовой выборке в абсолютных величинах

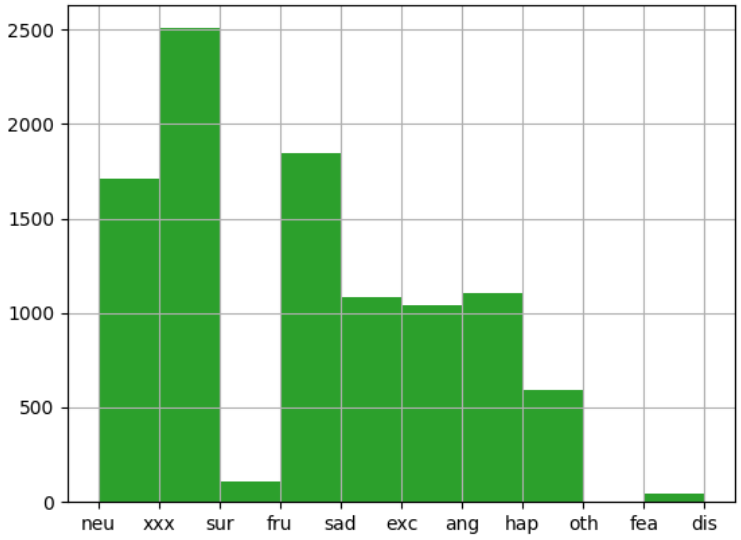

Рис. 20: IEMOCAP - Столбчатая диаграмма: количества аудио записей по эмоциям

2.4. IEMOCAP

Набор данных IEMOCAP[14] представляет собой мультимодальный диадический набор данных разговоров, записанный с участием 5 женщин и 5 мужчин, все дикторы профессиональные актеры. Он содержит беседы из 5 сессий, в каждой из которых один мужчина и одна женщина общаются на заданную тему. Каждая сессия разделена на фразы вручную, и каждая фраза аннотируется как минимум 3 аннотаторами для классификации по одному из 9 классов эмоций (злость, радость, взволнованность, нейтральность, грусть, удивление, отвращение, страх, разочарование) и два класса других меток('oth' - другое и 'xxx' - аннотаторы не смогли прийти к единому ммнению). Беседы имеют сценарный и импровизационный характер. Но у него есть небольшой недостаток - есть перекрытые голосов в некоторых аудиозаписях из-за того, что ведется диалог, поэтому метрики систем верификации диктора на таких данных выше чем на остальных.

IEMOCAP содержит разметку на 9 эмоций. Из гистограммы, представленной на рисунке 20, видно, что корпус имеет неравномерное распределение: количество аудиозаписей для каждой эмоции сильно различается. Кроме того, внутри каждого диктора наблюдается неравномерное распределение эмоций по количеству аудиозаписей, что иллюстрирует гистограмма 21. Было принято решение избавиться от данных с пометкой 'oth' и 'xxx', потому что они не информативны. Так же были выкинуты записи с отметкой 'dis', потому что там всего 2 аудио записи. После прослушивания 'exc'(взволнованность) и 'hap'(радость) было принято решение объединить эти эмоции под меткой 'радость'.

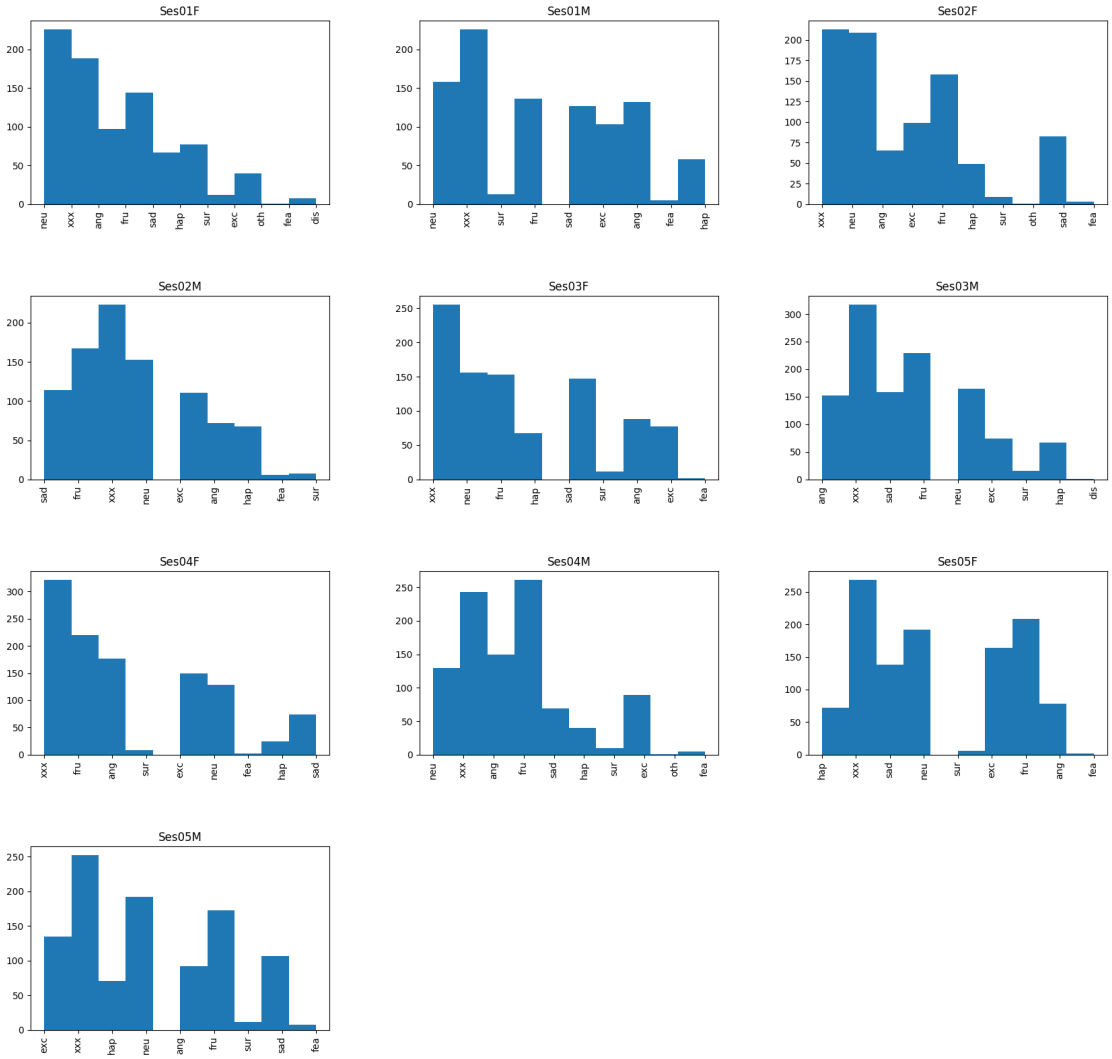

Рис. 21: IEMOCAP - Столбчатая диаграмма: количества аудио записей по эмоциям для каждого диктора

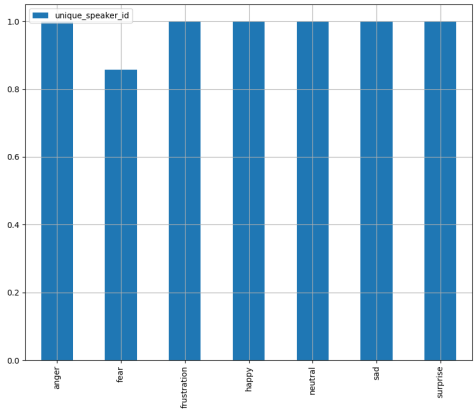

Рис. 22: IEMOCAP - Столбчатая диаграмма: количества уникальных дикторов по эмоциям в тренировочной выборке в абсолютных величинах

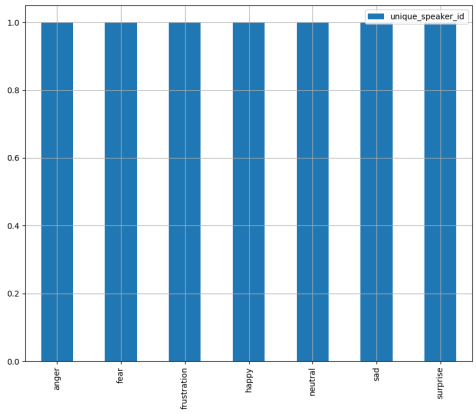

Рис. 23: IEMOCAP - Столбчатая диаграмма: количества уникальных дикторов по эмоциям в тестовой выборке в абсолютных величинах

Тренировочная выборка включает в себя данные 7 пользователей, в то время как тестовая выборка состоит из данных 3 пользователей. На основе этого корпуса было принято решение не создавать валидацинную выборку из-за небольшого числа дикторов. Важно отметить, что внутри каждой выборки каждый диктор представлен почти всеми эмоциями, что ясно видно на графиках 22 и 23.: только у одного диктора отсутствует эмоция страха в тренировочной выборки, у этого диктора не было этой эмоции в исходном наборе данных.

2.5. EmoV-DB

В EmoV-DB [22] англоязычные носители (2 женщины и 2 мужчин) читали предложения, выражая одну из эмоций (веселье, гнев, сонливость, отвращение и нейтральность). Записи для данных были выполнены в безэховых камерах, то есть они студийного качества. Предложения были взяты из базы данных CMU-arctic. Запи-

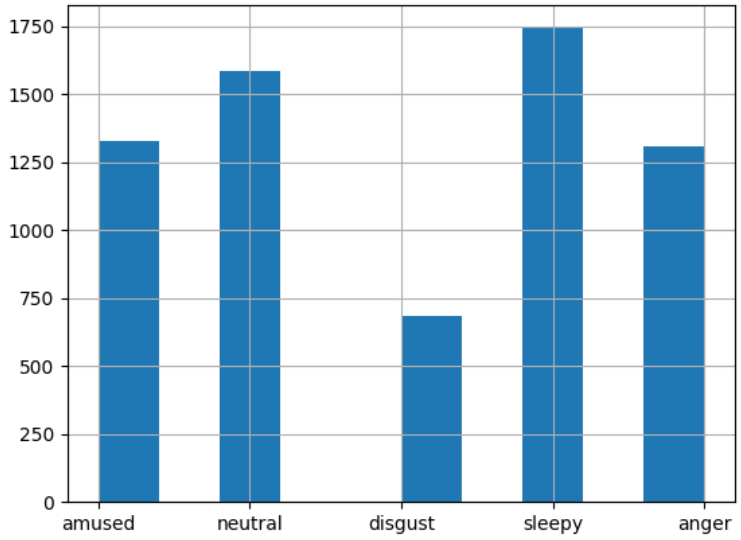

Рис. 24: EmoV-DB - Столбчатая диаграмма: количества аудио записей по эмоциям

си производились в несколько сессий длительностью около 30 минут с последующим перерывом от 5 до 15 минут, и сбор данных проходил в течение нескольких дней в зависимости от доступности актеров. Актерам было предложено записать каждый класс эмоции отдельно в различных сессиях. Предложения были вручную сегментированы(определение интервалов начала и конца каждого предложения) для некоторых дикторов.

Корпус EmoV-DB размечен на 5 классов. Он является не сбалансированным 24 из-за отсутствия определенных классов у некоторых дикторов, что видно в гистограмме 25. Для нашей задачи была откинуты аудиозаписи с меткой сонливость. Так метка веселье(amused) была трансформирована в радость(happy). После анализа этого набора данных было принято решение отдать его полностью в тренировочную выборку.

Другие наборы данных

В ходе работы были так же рассмотрены наборы данных, от которых пришлось отказаться:

1. ESD: имеют плохую разметку по эмоциям, много нейтральных записей помечены метками эмоций

2. EmoReact: не имеет разметки по дикторам, эмоции по большой части визуальные, а не звуковые

3. MSP-AVW Whisper: хороший набор данных содержащий предложения произнесенные как обычным голосам, так и шёпотом

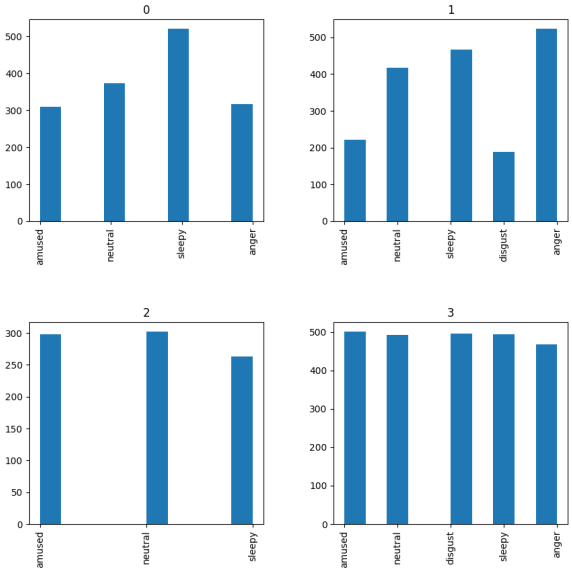

Рис. 25: EmoV-DB - Столбчатая диаграмма: количества аудио записей по эмоциям для каждого диктора

4. SEWA: содержит реакции на рекламу то есть представляет специфичный домен и не достаточно экспрессивен

5. FAU-Aibo: эмоциональный корпус на немецком языке

2.6. Выводы

В данной главе были рассмотрены и проанализированы пять наборов данных: VoxCeleb, CREMA-D, MSP-Podcast, IEMOCAP и EmoV-DB. Каждый из этих наборов данных имеет свои особенности. На основе рассмотренных данных был создан корпус для проведения экспериментов в данной работе. Корпус был разделен на тренировочную, валидационную и тестовую выборку с сохранением распределения.

3. Эксперименты

В данной главе будет верифицирована проблема неустойчивость моделей распознавания пользователей на стилистических данных, будут описаны эксперименты по дообучению и состязательному обучению для улучшения задачи верификации и будет приведен анализ полученных результатов.

3.1. Неустойчивость моделей верификации на стилистических данных

Одним из результатов статьи [20]: в векторах из кодировщика модели верификации дикторов содержится стилистическая информация. Их эксперименты были проведены на наборе данных IEMOCAP и IViE с использование архитектур I-vector и X-vector.

Во-первых, был проведен эксперимент подтверждающий результаты статьи [20]. В качестве модели верификации была выбрана предобученная архитектура GE2E из репозитория [4]. В качестве данных для верификации был выбран набор данных ESD. Каждый аудио файл был закодирован с помощью предобученного кодировщика. После к каждому из векторов был применен метод уменьшения размерностей LDA. Затем был построен классификатор SVM с ядром RBF, классификация проводилась на 5 эмоций из датасета ESD. Была достигнута метрика ROC_AUC 0.92 (Рис. 26). Так же были проведены эксперименты с другими методами уменьшения размерностей PCA, t-SNE и классификатором на основе дерева решений, они показали похожие или меньшие метрики качества классификации.

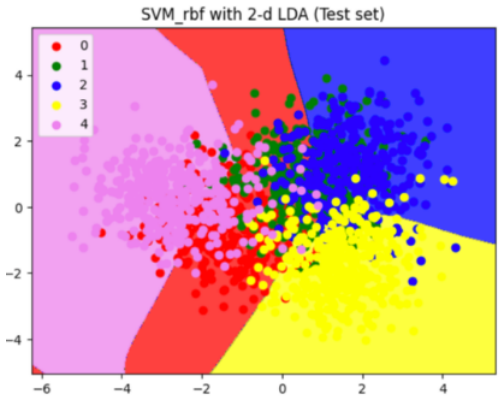

Рис. 26: Визуализация результатов классификации тестовой выборки из ESD

В статье [27] была показана проблема неустойчивость архитектуры X-vector к стилистическим данным на этапе верификации. Проблема была показана на 3 наборах данных: CREMA-D, IEMOCAP, MSP-Podcast. Более подробно про эту статью описано в обзоре литературы.

Во-вторых, был проведен ряд экскрементов по верификации результатов статьи

[27]. Проблема была верифицирована на двух архитекторах предобученных кодировщиках GE2E [4] и X-vector [24](библиотека SpeechBrain [19]). Эксперименты были проведены на 4 наборах данных: CREMA-D, EmoV-DB, IEMOCAP, ESD. По результатам экспериментов были подтверждены вывод из статьи[27]: кодировщик кодирует информацию о стиле, которая является шумом для систем верификаций пользователя и к данному шуму модели неустойчивы. В лучшем случае система показала относительную деградацию на 6% по сравнению с нейтральным множеством на верификации (IEMOCAP - x-vector). В худшем случае система показала относительную деградацию на 670% по сравнению с нейтральным множеством на верификации (EmoV-DB - x-vector). Более того, результаты подтверждены на архитектуре GE2E и двух новых наборах данных: EmoV-DB и ESD. Так же была показана деградация системы при верификации на данных с сонливым стиле (Таб. 6). Результаты приведены в таблица 5, 6, 7, 8.

в лучшем случае деградация составила на 6% в худшем случае на 670%

Enroll \ Test	anger	disgust	fear	happy	sad	neutral
X-vector	18.7004	17.3876	19.1008	17.0984	16.5823	9.03233
GE2E	21.1308	19.0636	25.9377	21.075	21.6014	8.69139

Таблица 5: CREMA-D: EER для систем верификации диктора в %

Enroll \ Test	anger	sleepy	amused	neutral
X-vector	8.33704	14.0116	10.3425	2.09357
GE2E	9.9634	17.9499	12.7033	4.73289

Таблица 6: EmoV-DB: EER для систем верификации диктора в %

Enroll \ Test	anger	sad	fustration	happy	neutral
X-vector	44.8706	44.3086	42.6204	43.8407	39.9824
GE2E	41.7695	43.1321	40.5765	41.2647	37.158

Таблица 7: IEMOCAP: EER для систем верификации диктора в %

Enroll \ Test	anger	happy	surprise	sad	neutral
X-vector	8.50798	9.80824	10.1413	7.67539	4.82752
GE2E	7.11146	9.12226	8.47855	8.26897	2.62037

Таблица 8: ESD: EER для систем верификации диктора в %

3.2. Конфигурация экспериментов

В следующих двух подразделах будут описаны две группы экспериментов, направленных на увлечение устойчивости системы верификации дикторов на эмоциональных данных. В первой группе экспериментов будет представлен наивный способ борьбы с деградацией: дообучение на эмоциональных данных. Во второй группе экспериментов представлен метод состязательного обучения в основе которого лежит слой обратных градиентов.

В основе всех экспериментов в качестве кодировщика выбрана архитектура ECAPA-TDNN [7]. Эта архитектура показала хорошие метрики на соревнованиях [23] [2] по верификации пользователей, на данный момент ECAPA считается одной из SOTA архитектур. Модель предобучена [26] на VoxCeleb1+VoxCeleb2. Модель обучалась на данных длинной 3 секунды, Её размер порядка 20 миллионов параметров.

Всех эксперименты проводились на наборе данных составленном в главе 2. Данные были обрезаны случайно до 3 секунд или заполнены тишиной до 3х секунд. Над данными была проведенная базовая обработка: вырезании пауз, приведение к одному каналу и к одной частоте дискретизации.

В AAM-Softmax слое использовались параметры margin 0.2 и scale 32.0, такой выбор был сделан на основе соревнования [2]. В этом слое содержалось порядка 250000 обучаемых параметров.

Обучение останавливалась в двух случаях: когда на протяжение несколько десятка эпох метрика EER на валидации росла, оставалась на том же уровне или после большого числа эпох(больше 400). Валидация считалась каждые 10 эпох, чекпоинты моделей так же считались каждые 10 минут. Лучший чекпоинт выбирался по метрики на валидационном множестве.

Для ускорения подсчета метрик на тестовых наборах данных были применены знания параллельного программирования. На каждый из наборов данных создавался процесс, для больших наборов данных(oxCeleb1 тест, MSP-Podcast тест) создавался по два процесса - происходила парализация по дикторам.

Процесс тестирования был устроен следующем образом. Каждый раз в регистрационное множество бралось 3 аудио записи, они усреднялись для образования центройды. Такой подход очевидно улучшает все метрики и при этом является естественным ограничением для промышленного применения. Все метрики считались в стиле cross-validation по диктора, внутри каждого диктора производилась cross-validation по регистрационному множеству. В качестве метрики близости векторов было выбрано косинусное сходство.

Все данные логировались с помощью библиотеке tensorboard.

3.3. Дообучение

В этой группе экспериментов я перебирал несколько параметров: количество замороженных слоев и скорость обучения. Результаты представлены в таблице 10. Общее устройство обучения показано на Рис. 27.

Рис. 27: Архитектура ECAPA-TDNN для дообучения [7]

% замороженных параметров кодировщика	Скорость обучения	CREMA-D (тест)	MSP-Podcast (тест)	IEMOCAP (тест)	VoxCeleb1 (тест)
Без дообучения	-	10.07	3.59	23.46	**0.44**
28%	$1e-5$	8.01	**2.37**	**11.01**	1.09
	$5e-6$	8.47	2.67	11.59	0.97
42%	$1e-5$	**7.45**	2.59	11.16	0.92
	$5e-6$	8.68	2.63	12.84	0.99
86%	$1e-7$	11.01	3.66	23.00	0.549
	$1e-8$	11.21	3.83	23.24	0.495

Таблица 9: Результаты экспериментов по дообучению. Метрика EER в %.

Результаты демонстрируют, что модель смогла обучиться под эмоциональный домен составленного тренировочного набора данных. Но на менее экспрессивном VoxCeleb1 (тест) метрика стала хуже, этот результат соответствует переобучению на тренировочный набор данных. В целом можно заключить, что наша система будет работать лучше для эмоциональных людей, но в данной работе этот результат не соответствует поставленной цели.

3.4. Состязательное обучение: слой обратных градиентов

В данной группе экспериментов реализовано состязательное обучение на основе слоя обратных градиентов. Общая архитектура обучения показана на Рис. 28.

Параметр λ отвечает за интенсивность обучение весов кодировщика со стороны классификатора эмоций. В то время как параметр β отвечает за интенсивность обучение весов кодировщика со стороны классификатора дикторов.

Результаты не сильно отличаются от результатов предыдущих экспериментов по дообучению: модель так же смогла обучиться под эмоциональный домен, но с эффектом переобучения.

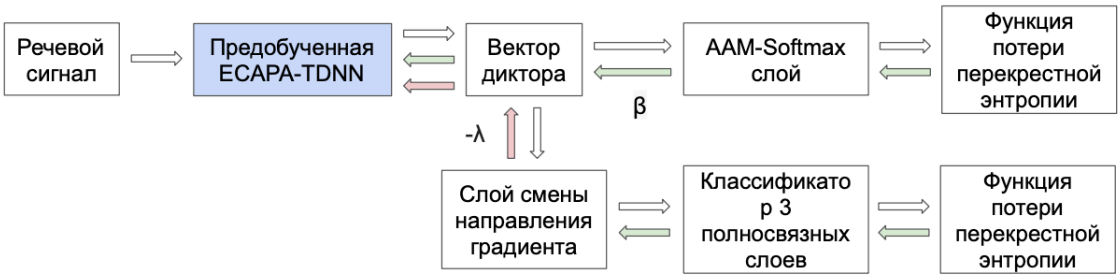

Рис. 28: Архитектура ECAPA-TDNN для состозятельного обучение

% замороженных параметров кодировщика	Скорость обучения	Вес функции потерь диктора β	Вес функции потерь эмоций λ	CREMA-D (тест)	MSP-Podcast (тест)	IEMOCAP (тест)	VoxCeleb1 (тест)
Без дообучения	-	-	-	10.07	3.59	23.46	**0.44**
42%	$1e-5$	0.95 1.0	0.1 0.1	8.07 **7.73**	**2.35** 2.53	10.43 **9.54**	1.11 1.03
42%	$5e-6$	0.95 1.0	0.1 0.1	8.26 8.53	2.72 2.59	13.31 12.29	0.83 0.82
86%	$1e-6$	1.0	0.1	8.45	2.64	12.98	0.68
86%	$5e-7$	1.0	0.1	10.31	2.68	17.00	0.625

Таблица 10: Результаты экспериментов со слоем обратных градиентов. Метрика EER в %.

3.5. Выводы

В ходе работы была подтверждена проблема неустойчивости моделей верификации к эмоциональным данным на этапе верификации: все модели показывают деградацию от 6% до 670% в относительных величинах. Поставленные эксперименты по улучшению модели верификации не помогли достичь цели работы: модели переобучились на составленный набор данных и не смогли превзойти базовую систему без дообучения на нейтральной тестовой выборке.

4. Предложения по улучшению

В данной главе будет предложен метод для улучшения результатов данной работы.

Одной из основных причин сложившейся ситуации, в которой модели не удалось превзойти базовую модель на нейтральных данных, является ограниченность доступных данных. В данном домене существует несколько небольших наборов данных с хорошей разметкой и большой датасет - VoxCeleb, который, не был размечен по эмоциям. Для решения проблемы неустойчивости моделей на эмоциональных данных была предпринята попытка объединить множество маленьких датасетов, но это не привело к желаемым результатам.

В свете этого, предлагается улучшить результаты исследования путем разметки большого датасета VoxCeleb с использованием предобученной модели распознавания эмоций [25]. Корпус VoxCeleb составлен из интервью знаменитостей и является менее экспрессивным набором данных, чем составлен корпус из моих экспериментов, но при этом в нём есть эмоциональные аудио записи. Такой подход уже был применен в статье [9] на набор данных VoxCeleb1, там этот подход превзошел остальные методы.

Разметка VoxCeleb хорошей моделью распознавания эмоций позволит расширить тренировочную выборку предложенную в этой работе и, таким образом, возможно повысить точность моделей верификации дикторов.

Заключение

В заключении данной работы приведены основные результаты и выводы, основанные на проведенных исследованиях.

В ходе исследования был составлен набор данных, включающий информацию о 9 различных эмоциях и 1460 дикторах. Этот набор данных был использован для проведения дообучения и состязательного обучения в рамках данной работы.

Важным результатом является подтверждение проблемы неустойчивости классических моделей верификации при работе с экспрессивными данными. Наши эксперименты подтвердили, что такие модели могут демонстрировать деградацию в качестве на уровне от 6% до 670% в относительных величинах.

Для улучшения результатов, мы реализовали дообучение и состязательное обучение с использованием слоя смены направления градиента. Это позволило значительно улучшить качество модели на эмоциональных наборах данных. Однако, в ходе исследования была выявлена проблема переобучения на менее экспрессивных данных. Был проведен анализ причины переобучения и предложен метод борьбы с переобучением.

Таким образом, работа не позволила достичь значительных результатов в улучшении модели верификации пользователя путем разделения физиологической и эмоциональной информации в векторном представлении. Однако, дальнейшие исследования и разработки необходимы для решения проблемы переобучения и повышения устойчивости моделей при работе с экспрессивными данными.

Список литературы

[1] Meng Zhong, Zhao Yong, Li Jinyu, and Gong Yifan. Adversarial speaker verification // ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) / IEEE. — 2019. — P. 6216–6220.

[2] Chen Zhengyang, Han Bing, Xiang Xu, Huang Houjun, Liu Bei, and Qian Yanmin. Build a SRE Challenge System: Lessons from VoxSRC 2022 and CNSRC 2022 // arXiv preprint arXiv:2211.00815. — 2022.

[3] Chung Joon Son, Nagrani Arsha, and Zisserman Andrew. Voxceleb2: Deep speaker recognition // arXiv preprint arXiv:1806.05622. — 2018.

[4] CorentinJ. Implementation of Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis // github. — 2021. — Access mode: `https://github.com/CorentinJ/Real-Time-Voice-Cloning/wiki/Pretrained-models`.

[5] Tong Fuchuan, Zheng Siqi, Zhou Haodong, Xie Xingjia, Hong Qingyang, and Li Lin. Deep Representation Decomposition for Rate-Invariant Speaker Verification // arXiv preprint arXiv:2205.14294. — 2022.

[6] Snyder David, Garcia-Romero Daniel, Povey Daniel, and Khudanpur Sanjeev. Deep neural network embeddings for text-independent speaker verification. // Interspeech. — 2017. — Vol. 2017. — P. 999–1003.

[7] Desplanques Brecht, Thienpondt Jenthe, and Demuynck Kris. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification // arXiv preprint arXiv:2005.07143. — 2020.

[8] Mun Sung Hwan, Han Min Hyun, Kim Minchan, Lee Dongjune, and Kim Nam Soo. Disentangled Speaker Representation Learning via Mutual Information Minimization // 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC) / IEEE. — 2022. — P. 89–96.

[9] Kang Woo Hyun, Mun Sung Hwan, Han Min Hyun, and Kim Nam Soo. Disentangled speaker and nuisance attribute embedding for robust speaker verification // IEEE Access. — 2020. — Vol. 8. — P. 141838–141849.

[10] Albanie Samuel, Nagrani Arsha, Vedaldi Andrea, and Zisserman Andrew. Emotion recognition in speech using cross-modal transfer in the wild // Proceedings of the 26th ACM international conference on Multimedia. — 2018. — P. 292–301.

[11] Dehak Najim, Kenny Patrick J, Dehak Réda, Dumouchel Pierre, and Ouellet Pierre. Front-end factor analysis for speaker verification // IEEE Transactions on Audio, Speech, and Language Processing. — 2010. — Vol. 19, no. 4. — P. 788–798.

[12] Wan Li, Wang Quan, Papir Alan, and Moreno Ignacio Lopez. Generalized end-to-end loss for speaker verification // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) / IEEE. — 2018. — P. 4879–4883.

[13] Hu Jie, Shen Li, and Sun Gang. Squeeze-and-excitation networks // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — P. 7132–7141.

[14] Busso Carlos, Bulut Murtaza, Lee Chi-Chun, Kazemzadeh Abe, Mower Emily, Kim Samuel, Chang Jeannette N, Lee Sungbok, and Narayanan Shrikanth S. IEMOCAP: Interactive emotional dyadic motion capture database // Language resources and evaluation. — 2008. — Vol. 42. — P. 335–359.

[15] Lotfian Reza and Busso Carlos. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings // IEEE Transactions on Affective Computing. — 2017. — Vol. 10, no. 4. — P. 471–483.

[16] Nagrani Arsha, Chung Joon Son, and Zisserman Andrew. Voxceleb: a large-scale speaker identification dataset // arXiv preprint arXiv:1706.08612. — 2017.

[17] Naini Abinay Reddy, Rao Achuth, and Ghosh Prasanta Kumar. Whisper to Neutral Mapping Using I-Vector Space Likelihood and a Cosine Similarity Based Iterative Optimization for Whispered Speaker Verification // 2022 National Conference on Communications (NCC) / IEEE. — 2022. — P. 130–135.

[18] Peddinti Vijayaditya, Povey Daniel, and Khudanpur Sanjeev. A time delay neural network architecture for efficient modeling of long temporal contexts // Sixteenth annual conference of the international speech communication association. — 2015.

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

[20] Williams Jennifer and King Simon. Disentangling Style Factors from Speaker Representations. // Interspeech. — 2019. — P. 3945–3949.

[21] Snyder David, Garcia-Romero Daniel, Sell Gregory, Povey Daniel, and Khudanpur Sanjeev. X-vectors: Robust dnn embeddings for speaker recognition // 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) / IEEE. — 2018. — P. 5329–5333.

[22] Adigwe Adaeze, Tits Noé, Haddad Kevin El, Ostadabbas Sarah, and Dutoit Thierry. The emotional voices database: Towards controlling the emotion dimension in voice generation systems // arXiv preprint arXiv:1806.09514. — 2018.

[23] Zheng Yu, Chen Yihao, Peng Jinghan, Zhang Yajun, Liu Min, and Xu Minqiang. The speakin system description for cnsrc2022 // arXiv preprint arXiv:2209.10846. — 2022.

[24] speechbrain. Implementation of X-vector pretrained on the VoxCeleb // huggingface. — 2021. — Access mode: `https://huggingface.co/speechbrain/spkrec-xvect-voxceleb`.

[25] speechbrain. Emotion Recognition with wav2vec2 base on IEMOCAP // huggingface. — 2022. — Access mode: `https://huggingface.co/speechbrain/emotion-recognition-wav2vec2-IEMOCAP`.

[26] speechbrain. Implementation of ECAPA-TDNN pretrained on the VoxCeleb // huggingface. — 2022. — Access mode: `https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb`.

[27] Pappagari Raghavendra, Wang Tianzi, Villalba Jesus, Chen Nanxin, and Dehak Najim. x-vectors meet emotions: A study on dependencies between emotion and speaker recognition // ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) / IEEE. — 2020. — P. 7169–7173.