Федеральное государственное бюджетное образовательное учреждение
высшего образования
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ
И РАДИОЭЛЕКТРОНИКИ (ТУСУР)

На правах рукописи

Рахманенко Иван Андреевич

АЛГОРИТМЫ И ПРОГРАММНЫЕ СРЕДСТВА ВЕРИФИКАЦИИ ДИКТОРА ПО ПРОИЗВОЛЬНОЙ ФРАЗЕ

Специальность 05.13.17 – «Теоретические основы информатики»

Диссертация на соискание ученой степени кандидата технических наук

Научный руководитель:
доктор технических наук, профессор
Р.В. Мещеряков

Томск 2017

Содержание

Введение

Актуальность темы. Задача автоматической верификации диктора является одной из наиболее сложных задач в области обработки речи. Возможность достоверно автоматически идентифицировать и верифицировать человека по его голосу позволила бы заменить обычные пароли, использовать в полной мере человеко-машинные интерфейсы, упростить разметку аудиостенограмм.

Голос, в отличие от сетчатки глаза или отпечатков пальцев, считается менее надежным идентифицирующим признаком, однако в некоторых случаях, требуется именно верификация по голосу. Простота применения, отсутствие необходимости в специальных регистрирующих устройствах, увеличение защищенности системы путем добавления дополнительного фактора верификации – все это дает неоспоримые преимущества при использовании голосовой верификации в реальных системах разграничения доступа.

Системы идентификации и верификации диктора по голосу нашли широкое применение в таких областях как дистанционное банковское обслуживание, биометрическая многофакторная верификация, криминалистическая экспертиза. Внедрение систем верификации по голосу планируется в банках ВТБ24 и Сбербанк, используется в таких зарубежных банках как Barclays, HSBC Holdings, Santander, TD Bank. Разработкой систем верификации диктора занимаются такие компании как Центр Речевых технологий, Microsoft, Nuance, Powervoice, Agnitio, VoiceVault и другие. Достаточно часто системы верификации диктора используются совместно с системами распознавания лица.

Решением проблем верификации диктора занимались такие ученые, как Сорокин В. Н., Матвеев Ю. Н., Симончик К. К., Пеховский Т.С., Новоселов С.А., Потапова Р. К., Рамишвили Г.С., Bonastre J. F., Campbell J. P., Campbell W. M., Rose R. C., Reynolds D. A., Quatieri T. F., Dunn R. B., Rosenberg A. E., Isobe T., Dehak N., Colibro D., Lei Y., Shum S.H., Stafylakis T., Kenny P., Xu L.,

McLaren M., Ferrer L., Richardson F., Variani E., Soong F. K, Garcia-Romero D., Martin A., Atal B. S.

Методы, используемые в современных системах верификации диктора далеко не идеальны, что накладывает на такие системы определенные ограничения. Некоторые методы верификации отлично работают в хороших акустических условиях при минимальном уровне шума, однако заметно теряют в точности распознавания в условиях малого соотношения сигнал/шум. Кроме того, существуют факторы, влияющие на точность подобных систем – голос человека может меняться с течением времени из-за различного физического и эмоционального состояния. Точность системы верификации диктора – одна из определяющих ее характеристик, необходимых для успешного применения. Современные системы не обладают той точностью, которая необходима для их внедрения и использования. С каждым годом требования к точности повышаются, мотивируя тем самым развитие существующих алгоритмов, методов и систем. Таким образом, задача создания алгоритмов, методов и систем автоматической верификации диктора по произвольной фразе, обладающих высокой точностью, является актуальной.

Цели и задачи исследования. Целью диссертационной работы является разработка и исследование алгоритмов и речевых признаков верификации диктора по произвольной фразе, повышающих точность верификации диктора по сравнению с известными подходами, методами и алгоритмами.

Для достижения поставленной цели сформулированы следующие основные задачи:

1) обзор существующих методов и алгоритмов верификации диктора по произвольной фразе, речевых признаков, используемых для верификации диктора;

2) разработка алгоритма верификации диктора с применением речевых признаков, полученных с помощью жадного и генетического алгоритмов отбора признаков.

3) разработка алгоритма генерации признаков, основанного на применении сверточной глубокой сети доверия;

4) разработка гибридного алгоритма верификации диктора по произвольной фразе на основе ансамбля классификаторов;

5) создание программного средства с применением полученных алгоритмов и параллельных вычислений на центральном и графическом процессорах;

6) оценка разработанных алгоритмов и программного средства на контрольных примерах и сравнение с аналогами.

Объектом исследования является процесс верификации диктора по произвольной фразе.

Предметом исследования являются алгоритмы и речевые признаки, используемые для верификации диктора по произвольной фразе.

Методы исследования. В диссертационной работе применялись методы теории вероятностей и математической статистики, методы оптимизации, интеллектуального анализа данных, цифровой обработки сигналов и обработки данных.

Достоверность результатов обеспечивается строгостью применения математических методов, результатами проведенных численных экспериментов с использованием реальных данных, а также путём сопоставления результатов, полученных в диссертации, с результатами, доступными в открытой печати.

Научная новизна полученных результатов. В диссертации получены следующие новые научные результаты.

1. Разработан оригинальный алгоритм верификации диктора, отличающийся от существующих применением речевых признаков, полученных с помощью жадного алгоритма отбора признаков.

2. Предложен алгоритм генерации признаков, основанный на применении сверточной глубокой сети доверия. Данный алгоритм отличается

от существующих расширенной архитектурой нейронной сети, выделяющей более высокоуровневые признаки и уменьшающей их количество.

3. Разработан гибридный алгоритм верификации диктора по произвольной фразе на основе ансамбля классификаторов. Отличительной особенностью алгоритма является применение в ансамбле классификаторов, использующих выходы первого и третьего скрытых слоев сверточной глубокой сети доверия.

Теоретическая значимость работы заключается в развитии алгоритмов и методов извлечения признаков из данных, алгоритмов верификации диктора по произвольной фразе. Алгоритм генерации признаков, основанный на применении сверточной глубокой сети доверия, может использоваться не только для выделения признаков из речевых данных, но и для выделения признаков из изображений. Также возможно применение полученного набора признаков для идентификации пола диктора и распознавания речи.

Практическая значимость работы подтверждается использованием полученных в ней результатов для решения практической задачи верификации диктора по произвольной фразе. Результаты внедрены в АО «ОЭЗ ТВТ «Томск», а также используются в учебном процессе на факультете безопасности ТУСУР.

Созданные алгоритмы и программное средство использованы в рамках мероприятия 1.3 ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014—2020 годы» (соглашение о предоставлении субсидии № 14.577.21.0172 от 27 октября 2015 г.; уникальный идентификатор RFMEFI57715X0172).

Результаты диссертационной работы были получены в рамках выполнения базовой части государственного задания Минобрнауки России, проект 8.9628.2017/8.9 на базе лаборатории медико-биологических исследований (ЛМБИ) ТУСУР.

На защиту выносятся приведенные ниже положения.

1. Разработанный алгоритм верификации диктора по произвольной фразе, отличающийся от существующих применением речевых признаков, полученных с помощью генетического и жадного алгоритмов, позволяет уменьшить равную ошибку 1-го и 2-го рода (EER) по сравнению со стандартным набором признаков. На некоторых данных полученный набор признаков позволяет уменьшить ошибку EER на 42,1 %.

Соответствует пункту 5 паспорта специальности: Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях. разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.

2. Предложенный алгоритм генерации признаков, основанный на применении глубокой нейронной сети доверия, позволяет выделять из речи высокоуровневые признаки и уменьшить их количество.

Соответствует пункту 5 паспорта специальности: Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.

3. Разработанный гибридный алгоритм верификации диктора по произвольной фразе на основе ансамбля классификаторов позволяет повысить точность верификации диктора применяя различные классификаторы и используя признаки, выделенные сверточной глубокой сетью доверия.

Соответствует пункту 5 паспорта специальности: Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.

4. Созданное программное средство верификации диктора, отличается от существующих применением алгоритмов обучения универсальной фоновой

модели (УФМ) на центральном и графическом процессорах, что позволяет уменьшить время обучения на 10 %.

Внедрение результатов диссертационного исследования. Результаты исследовательской работы были использованы при создании системы верификации диктора по произвольной фразе, используемой в АО «ОЭЗ ТВТ «Томск».

Разработанные алгоритмы и программное средство используются при изучении дисциплины «Программно-аппаратные средства обеспечения информационной безопасности» на кафедрах комплексной информационной безопасности электронно-вычислительных систем и безопасности информационных систем ТУСУР.

Апробация работы. Основные положения работы докладывались и обсуждались на следующих конференциях, семинарах:

– 1-ой Всероссийской акустической конференции, г. Москва, 2014 г.;

– 12th All-Ukrainian International Conference on Signal/Image Processing and Pattern Recognition UkrObraz'2014, г. Киев, Украина, 2014 г.;

– XI Международной конференция студентов и молодых ученых «Перспективы развития фундаментальных наук», г. Томск, 2014 г.;

– Международной научно-практической конференции «Электронные средства и системы управления», г. Томск, 2015 г.;

– Всероссийской научно-технической конференции студентов, аспирантов и молодых ученых «Научная сессия ТУСУР», г. Томск, 2013, 2014, 2016 г.;

– 18th International Conference on Speech and Computer SPECOM, г. Будапешт, Венгрия, 2016 г.;

– Томском IEEE семинаре «Интеллектуальные системы моделирования, проектирования и управления» г. Томск.

Публикации по теме диссертации. По результатам исследований опубликовано 10 печатных работ, из которых в рекомендованных ВАК РФ

периодических изданиях – 2. Две работы индексированы в базе научных публикаций SCOPUS.

Личный вклад автора. Постановка цели и задач научного исследования и подготовка материалов к печати велась совместно с научным руководителем. Автором самостоятельно разработаны алгоритм генерации признаков, алгоритм верификации диктора и программное средство, осуществлена постановка экспериментов и экспериментальные исследования, обработка полученных данных.

Структура и объем работы. Диссертационная работа состоит из введения, трех глав основной части, заключения, списка литературы из 138 наименований и 2-х приложений. Основная часть работы изложена на 111 страницах, в том числе содержит 34 рисунка и 21 таблицу.

Во введении приведено обоснование актуальности темы исследования, формулируется цель работы, представлены полученные автором основные результаты проведенных исследований, обозначена их научная новизна, теоретическая и практическая значимость, отражены основные положения, выносимые на защиту.

В первой главе описана постановка задачи, представлен обзор существующих речевых признаков, методов и алгоритмов верификации диктора по произвольной фразе.

Во второй главе приведен алгоритм верификации на базе Гауссовых смесей и универсальной фоновой модели; алгоритм верификации диктора с применением признаков, полученных с помощью жадного алгоритма отбора признаков; алгоритм генерации признаков, основанный на применении сверточной глубокой сети доверия; гибридный алгоритм верификации диктора по произвольной фразе на основе ансамбля классификаторов.

В третьей главе представлено описание и состав разработанного программного средства, внедрение разработанных алгоритмов и программного средства в деятельность АО «ОЭЗ ТВТ «Томск». Программное средство

включает в себя все необходимые модули для извлечения речевых признаков, обучения моделей дикторов и УФМ, а также проведения верификационных испытаний. Средство позволяет произвести отбор речевых признаков с помощью алгоритма жадного добавления-удаления и генетического алгоритма.

В заключении сформулированы основные научные и практические результаты.

1. Обзор существующих речевых признаков, методов и алгоритмов верификации диктора по произвольной фразе

1.1 Постановка задачи верификации диктора по произвольной фразе

Пусть имеется тестовый отрезок речи – *Y*, предполагаемый диктор – *S*. На данном отрезке *Y* может присутствовать голос одного или нескольких дикторов, могут быть различные шумы или тишина. Эти факторы могут влиять на результаты работы методов верификации, но решаются в рамках других задач – диаризации и фильтрации. Для решения же задачи верификации диктора, ограничимся условием, что на *Y* присутствует только речь одного диктора.

Задачу верификации диктора можно задать так: необходимо определить, присутствует ли речь диктора *S* на отрезке *Y*. Соответственно, зададим две гипотезы: на отрезке *Y* присутствует речь диктора *S* (гипотеза H_0), на отрезке *Y* отсутствует речь диктора *S* (гипотеза H_1). Для проверки данных гипотез оптимальным является использование теста отношения правдоподобия [1], который можно представить в виде формулы (1.1):

$$\frac{p(Y \mid H_0)}{p(Y \mid H_1)} \begin{cases} \geq \Theta - принять \quad H_0 \\ < \Theta - отклонить \quad H_0 \end{cases}, \tag{1.1}$$

где $p(Y \mid Hi)$, $i = 0, 1$ – функция плотности вероятности гипотезы H_i, или правдоподобие гипотезы H_i для данного речевого сегмента. Базовая задача системы верификации – задать методы, позволяющие вычислить значения функций $p(Y \mid H_0)$ и $p(Y \mid H_1)$.

Перед тем как вычислять значения данных функций, необходима предварительная обработка сигнала. К ней относят операции фильтрации и очистки от шума, выделение из сигнала характеристик, характерных для целевого диктора. Выходными данными для данного этапа будет являться

последовательность векторов признаков $X = \{\bar{x}_1,...,\bar{x}_T\}$, относящихся к различным временным промежуткам $t \in \{1,2,...,T\}$. Эти вектора можно использовать для вычисления правдоподобия гипотез H_0 и H_1. Представим гипотезу H_0 моделью λ_{hyp}, которая описывает предполагаемого диктора S в пространстве признаков x, а альтернативную гипотезу H_1 моделью $\lambda_{\overline{hyp}}$. Таким образом, логарифм отношения правдоподобия [1] можно вычислить как (1.2)

$$\Lambda(X) = \log p(X \mid \lambda_{hyp}) - \log p(X \mid \lambda_{\overline{hyp}}). \quad (1.2)$$

Получается, что оценить гипотезу H_0 возможно, так как логарифм правдоподобия $\log p(X|\lambda_{hyp})$ можно вычислить с использованием тренировочного набора данных $Train_S$ диктора S. Данный набор включает в себя аудиозаписи голоса диктора X_S, использующиеся для обучения модели λ_{hyp}. Однако оценить гипотезу H_1 затруднительно, так как она представляет собой бесконечное множество альтернатив, исключающих наличие диктора S на записи Y.

Существует два основных подхода для моделирования данной гипотезы [1]. Первый заключается в использовании множества отдельных моделей дикторов для представления альтернативной гипотезы [2, 3]. В этом случае, для каждого конкретного диктора используют свое множество альтернативных моделей, что является недостатком данного подхода в случае большого множества предполагаемых дикторов, использующих систему верификации. Возможен вариант создания систем с несколькими альтернативными моделями только для тех дикторов, голоса которых имеют близкие по величине признаки. Такие модели называют когортами [3].

Второй подход предполагает создание единой модели, обученной на речи нескольких дикторов. Такую модель называют общей моделью (general model), моделью мира (world model) или универсальной фоновой моделью (УФМ,

universal background model, UBM) [1]. Данная модель будет рассмотрена в разделе 1.3. Исследования, связанные с данной моделью, направлены на методы выбора дикторов для обучения модели. Преимущество этой модели заключается в том, что модель требуется обучить только один раз, используя ее впоследствии для вычисления функции правдоподобия [4, 5].

1.2 Обзор речевых признаков

Индивидуальность акустических характеристик голоса определяется тремя факторами: механикой колебаний голосовых складок, анатомией речевого тракта и системой управления артикуляцией [6].

Один из самых часто используемых признаков, используемых в научных работах, связанных с обработкой речи и распознаванием диктора, являются мел-кепстральные коэффициенты (Mel frequency cepstral coefficients, MFCC) [1, 2, 7-9]. По мнению автора, существуют другие признаки, которые могут содержать дополнительную информацию о дикторе, применение которой может улучшить точность распознавания. Следует провести обзор и применить для задачи верификации диктора другие признаки, используемые в обработке речи. К таким признакам можно отнести пары линейного спектра (line spectral pair, LSP), кепстральные коэффициенты перцептивного линейного предсказания (perceptual linear prediction cepstral coefficients – PLP), энергию, формантные частоты, частоту основного тона, вероятность вокализации (voicing probability), частоту пересечения нуля (zero crossing rate, ZCR), джиттер и шиммер [10-13].

Основной набор признаков, по сравнению с которым будем в дальнейшем производить сравнение полученных наборов признаков, это мел-частотные кепстральные коэффициенты. Метод мел-частотного кепстрального преобразования спектра был впервые представлен в работе [14]. Мел-кепстральные коэффициенты (МКК) используются в таких областях, как распознавание диктора, распознавание речи и многих других задачах,

связанных с обработкой речи. Наиболее часто используют 12, 13 или 14 МКК. Кроме того, часто используются дельта и двойные дельта коэффициенты, которые отражают изменения в мел-кепстральных коэффициентах во времени.

Несмотря на тот факт, что в спектре речи нет признаков, по которым можно было бы однозначно идентифицировать диктора [15], тем не менее, мел-кепстральные коэффициенты достаточно эффективно используются в задаче автоматической верификации диктора. Это возможно благодаря тому факту, что в спектре речи диктора отражается структура речевого тракта, которая позволяет отличаться голосам людей на физиологическом уровне.

Для вычисления МКК, после предварительной обработки сигнала и разбиения на отдельные отрезки - окна, производится дискретное преобразование Фурье (ДПФ). Частоты f, полученные после ДПФ, переводят к шкале мел f_{mel} с помощью преобразования (1.3) [16]:

$$f_{mel} = 1125\ln(1 + f/700) \tag{1.3}$$

Преобразование между частотами в герцах и в мелах является линейным до частоты 1000 Гц и логарифмическим выше данной частоты. Для выполнения данного преобразования создается набор треугольных фильтров и вычисляется логарифм энергии в каждой полосе частот данных фильтров [14] (Рисунок 1.2.1). Последним шагом извлечения МКК является выполнение обратного ДПФ.

Мел-кепстральные коэффициенты в качестве признаков, используемых для идентификации и верификации диктора, используются в [1, 2, 7].

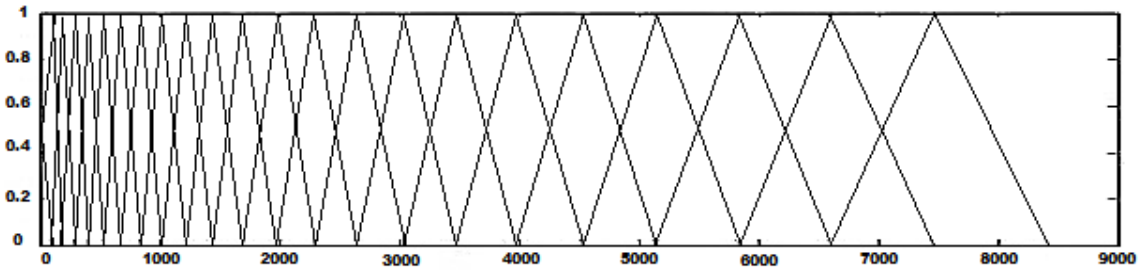

Рисунок 1.2.1 – Переход к шкале мел и наборы фильтров речевого сигнала

В [17] наиболее важными характеристиками голоса считаются формантные частоты. В [18] было показано, что четвертая форманта практически не зависит от типа фонемы и характеризует речевой тракт конкретного диктора. Формантами называют области концентрации энергии в спектре звука речи [19]. Таких областей может быть несколько, обозначаются они *F1*, *F2*, *F3* и т.д. Появление нескольких резонансных областей в самом первом приближении объясняется тем, что речевой тракт состоит из системы резонансных полостей [19]. Форманты можно выделить только для вокализованных звуков, соответственно для верификации диктора с применением формант отбирают речь с вокализованными звуками. Рассмотрим спектр слова “шевеля” (Рисунок 1.2.2). На нем выделены черными точками области концентрации энергии, которые и называются формантами.

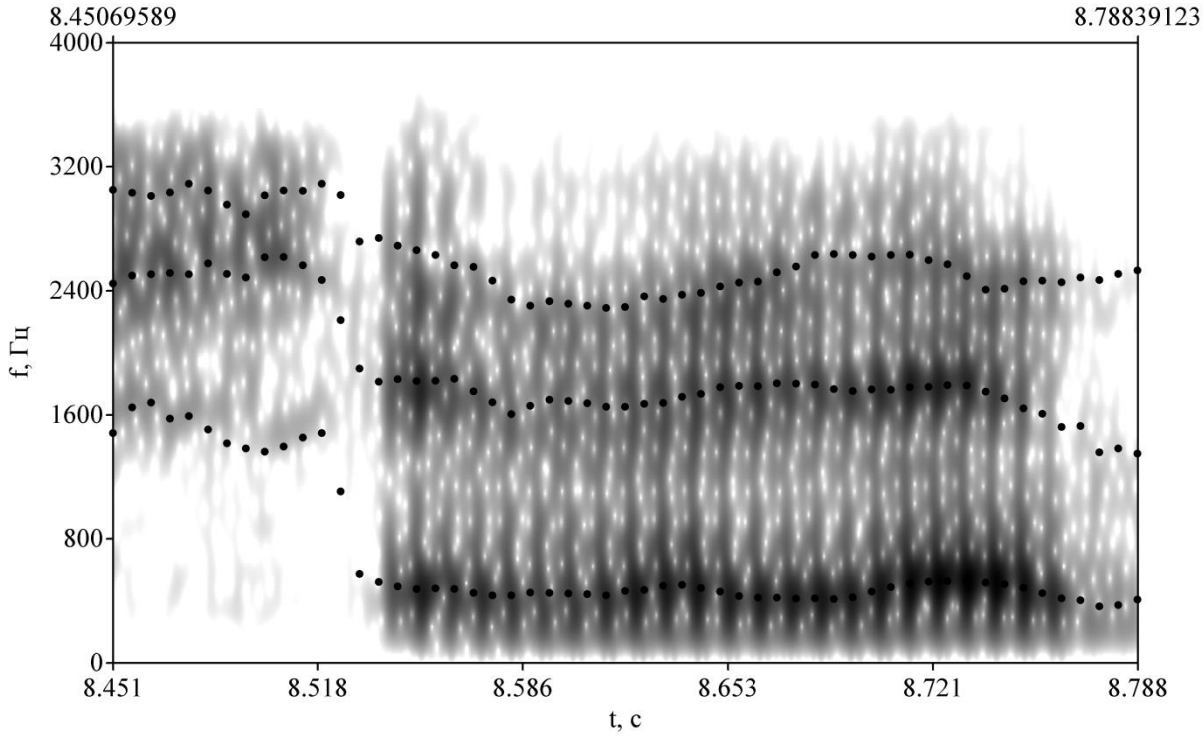

Рисунок 1.2.2 – Отображение формант в спектре слова “шевеля”

Для извлечения формант используются такие методы, как выбор спектральных пиков [20-24]; метод извлечения корней на основе коэффициентов линейного предсказания [25, 26]; метод анализа через синтез [22, 27]; дискретное вейвлет преобразование, объединенное с логарифмической мощностью спектра [28]; различные комбинированные методы [29-33]; методы с применением глубоких нейронных сетей [34]. Для верификации диктора используются частоты 3-й и 4-й форманты гласных [35], с 1-й по 5-ю форманты [36], треки первых трех формант [37], 8 формант по отдельности и совместно с другими признаками [38], от 7 до 9 формант [39], первые 3 форманты [33, 40, 41].

1.3 Метод верификации, основанный на применении Гауссовых смесей

Одним из самых популярных методов, используемых в сфере верификации диктора по голосу, является модель Гауссовой смеси (Gaussian Mixture Model, GMM, ГС) [1, 8, 9, 42]. Данная обобщенная вероятностная модель успешно применяется при решении задачи текстонезависимой верификации диктора, так как многомерное нормальное распределение способно представлять произвольные распределения. К числу таких сложных распределений можно отнести распределение МКК в записях речи. Применение ГС для текстонезависимого распознавания диктора впервые было описано в [42].

Применяя модель Гауссовой смеси, плотность вероятности смеси для D-мерного вектора характеристик *x* можно представить формулой (1.4) [1]:

$$p(x|\lambda) = \sum_{i=1}^{M} w_i p_i(x|\mu_i, \Sigma_i), \tag{1.4}$$

где плотность вероятности смеси $p(x|\lambda)$ представляет собой взвешенную сумму M D-мерных Гауссовых плотностей вероятности $p_i(x|\mu_i,\sum_i)$ с весами w_i, которые характеризуются вектором математических ожиданий μ_i и ковариационной матрицей Σ_i (1.5):

$$p_i(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'(\Sigma_i)^{-1}(x-\mu_i)\right\}. \tag{1.5}$$

При этом веса компонент смеси w_i удовлетворяют ограничению $\sum_{i=1}^{M} w_i = 1$. Таким образом, все параметры модели Гауссовой смеси можно представить как $\lambda = \{w_i, \mu_i, \Sigma_i\}$, где $i = 1,\ldots,M$. Кроме того, в большинстве систем используется не полная ковариационная матрица Σ_i, а диагональная

ковариационная матрица. Это обосновано тем, что Гауссова смесь с полной ковариационной матрицей может быть адекватно представлена Гауссовой смесью с диагональной ковариационной матрицей и большим количеством компонент смеси. Применение диагональной матрицы ковариации упрощает вычисления и повышает точность работы системы верификации [1].

Параметры модели максимального правдоподобия с использованием заданного набора обучающих векторов, как правило, принято оценивать с помощью EM алгоритма [43, 44]. Данный алгоритм последовательно уточняет параметры Гауссовой смеси, монотонно увеличивая правдоподобие модели. Для вычисления параметров модели, используется апостериорная вероятность для i-ой компоненты смеси $\Pr(i|x_t, \lambda)$, которая вычисляется как (1.6) [1]:

$$\Pr(i|x_t, \lambda) = \frac{w_i p_i(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k p_i(x_t|\mu_k, \Sigma_k)}, \quad (1.6)$$

где $p_i(x_t|\mu_i, \Sigma_i)$ вычисляется согласно выражению 1.5.

К преимуществам использования Гауссовой смеси можно отнести низкую вычислительную сложность и нечувствительность к временным аспектам речи. Последнее также можно отнести и к недостаткам, так как информация более высокого уровня, характеризующая особенности произношения диктора, не используется.

Необходимо отметить, что индивидуальные компоненты смеси могут моделировать некоторое множество акустических классов [2]. Данное множество представляет собой набор конфигураций голосового тракта диктора, что позволяет использовать их в целях верификации. При этом i-ый акустический класс представляется компонентой смеси λ_i. Акустические классы являются «скрытыми», так как в обучающих и контрольных данных они не размечены. Если предположить, что векторы признаков независимы друг от

друга, то Гауссова смесь описывает эти классы через плотность распределения наблюдаемых векторов признаков.

К системам, разработанным с применением модели Гауссовой смеси, относят: системы с применением GMM-SVM подхода [8, 9, 45-50]; системы, комбинирующие Гауссовы смеси и скрытые Марковские модели (HMM) [51-54]; системы, применяющие метод главных компонент или векторное квантование для верификации диктора [55-62]; а также множество других систем, использующих различные методы и приемы совместно с Гауссовыми смесями [1, 2, 63-68].

Дальнейшее развитие данной модели было основано на создании **универсальной фоновой модели** (Universal Background Model, UBM, УФМ), и адаптации моделей дикторов из данной универсальной модели. Универсальная фоновая модель – это большая модель Гауссовой смеси, обученная для представления дикторонезависимого распределения признаков. Для обучения данной модели используется речевой корпус, содержащий аудиозаписи большого количества дикторов. Системы, созданные с использованием УФМ, называются системами верификации диктора на основе модели Гауссовых смесей и универсальной фоновой модели (GMM-UBM). Одни из первых вариантов подобных систем предложены в [5, 69-71].

Существует несколько подходов, применяемых для получения УФМ. Возможно простое обучение модели на всей обучающей выборке с помощью EM (Expectation-Maximization) алгоритма. Кроме того, возможно обучение отдельных моделей для разных выборок с последующим объединением результатов в одну универсальную фоновую модель. Например, возможно объединение отдельных моделей, обученных на выборках с дикторами-мужчинами и дикторами-женщинами, или обучение отдельных моделей для записей на различные типы микрофонов. Также известны другие подходы, связанные с обучением моделей для когорт (групп) дикторов [2, 4, 72].

При создании УФМ необходимо помнить, что данные, используемые для обучения модели, должны быть сбалансированными по отношению к дальнейшему применению системы. Т.е. длительность обучающей выборки для дикторов мужчин и женщин должна быть примерно одинаковой для системы голосовой верификации. Аналогично, обучающие данные должны быть сбалансированы и по типу используемых при записи дикторов микрофонов. В [1] производится обучение УФМ с помощью объединения двух отдельных 1024-компонентных УФМ для дикторов-мужчин и дикторов-женщин. Для каждой из двух моделей был использован речевой материал длительностью 1 час.

В GMM-UBM системе для создания модели диктора производится адаптация параметров универсальной фоновой модели на обучающих данных конкретного диктора [73, 74]. Данная адаптация известна также как Байесово обучение или оценка апостериорного максимума (MAP). Это позволяет не только увеличить точность распознавания диктора по сравнению с неадаптированными моделями, но и ускорить оценку соответствия моделей. Также как и в EM-алгоритме, адаптация состоит из двух шагов, однако на втором шаге вычисленные параметры смешиваются с исходными параметрами, взятыми из УФМ, по определенному коэффициенту.

Чтобы вычислить необходимые для адаптации модели параметры, производится оценка весов n_i (1.7), математических ожиданий $E_i(x)$ (1.8) и дисперсии $E_i(x^2)$ (1.9), соответствующая E-шагу EM-алгоритма:

$$n_i = \sum_{t=1}^{T} \Pr(i|x_t, \lambda_{\text{УФМ}}) \tag{1.7}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|x_t, \lambda_{\text{УФМ}}) x_t \tag{1.8}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i|x_t, \lambda_{\text{УФМ}}) x_t^2 \tag{1.9}$$

В зависимости от количества появлений *i*-ой компоненты смеси в обучающих данных n_i, производится адаптация отдельных компонент УФМ (1.10-1.13). При адаптации модели диктора могут вычисляться один или несколько следующих параметров: вес *i*-ой компоненты смеси w_i, математическое ожидание *i*-ой компоненты смеси μ_i и дисперсии σ_i^2.

$$w_i = \left[\alpha_i \frac{n_i}{T} + (1 - \alpha_i) w_i\right] \gamma \tag{1.10}$$

$$\mu_i = \alpha_i E_i(x) + (1 - \alpha_i)\mu_i \tag{1.11}$$

$$\sigma_i^2 = \alpha_i E_i(x^2) + (1 - \alpha_i)\left(\sigma_i^2 + \mu_i^2\right) - \mu_i{}^2, \tag{1.12}$$

где α_i – коэффициент адаптации, γ – коэффициент масштабирования, учитывающий, что все веса w_i суммируются к 1, $E_i(x)$ - математическое ожидание *i*-ой компоненты смеси, вычисленное на обучающих данных, $E_i(x^2)$ - дисперсия *i*-ой компоненты смеси, вычисленная на обучающих данных.

Если компонента смеси в обучающих данных встречается редко, это приводит к снижению значимости новых параметров, полученных после E-шага EM-алгоритма, и увеличению значимости старых, взятых из УФМ. С помощью коэффициента релевантности *r* можно контролировать, насколько часто должны встречаться новые данные, чтобы заменить старые в смеси (1.13).

$$\alpha_i = \frac{n_i}{n_i + r}. \tag{1.13}$$

Так как адаптация параметров зависит от данных, не все Гауссоиды фоновой модели адаптируются во время обучения модели диктора. Если знать, какие компоненты ГС не адаптируются при обучении модели диктора, можно хранить гораздо меньше данных. Было обнаружено [1], что примерно 24% компонент ГС дикторов-мужчин и 14% компонент ГС дикторов-женщин не изменяются.

В [69] показано, что системы голосовой верификации, производящие адаптацию модели диктора из универсальной фоновой модели, имеют намного лучшую точность, чем системы, в которых модель диктора обучается отдельно от УФМ. Это можно объяснить тем, что УФМ покрывает большинство классов акустических событий, появляющихся в речи дикторов. Соответственно, во время адаптации модели диктора, часть таких событий, появившихся в речи диктора, изменяют и подстраивают компоненты смеси под конкретного диктора. Оставшаяся часть событий, не встречающаяся в обучающей выборке, копируется из УФМ. Таким образом, это добавляет устойчивости модели к тем акустическим событиям конкретного диктора, которые отсутствовали в обучающей выборке.

Тот факт, что модель диктора была адаптирована из универсальной фоновой модели, позволяет ускорить оценку схожести двух Гауссовых смесей. Во-первых, лишь несколько смесей вносят значимый вклад в значение правдоподобия, поэтому для оценки отношения правдоподобия берут только C самых значимых компонент. Во-вторых, вектора, близкие к определенной компоненте смеси общей фоновой модели, будут также близки и к компонентам смеси адаптированной модели диктора. Для универсальной фоновой модели с M компонентами, требуется лишь $M+C$ вычислений Гауссиан, вместо $2M$ вычислений.

1.4 Метод верификации, основанный на факторном анализе

Комбинированный факторный анализ (Joint Factor Analysis, JFA), демонстрирует выдающиеся результаты в решении задачи текстонезависимой верификации диктора [75-77]. В комбинированном факторном анализе отрезок речи диктора можно представить супервектором M (1.14), состоящим из суммы компонент, представляющих подпространство диктора и канала (сессии):

$$M = m + Vy + Ux + Dz\,, \tag{1.14}$$

где m – супервектор, независимый от диктора и сессии (обычно УФМ), V и D задают подпространство диктора (матрицу собственных векторов голоса и диагональные остатки), и U задает подпространство сессии (матрицу собственных векторов каналов). Вектора y, x, и z – зависимые от диктора и сессии факторы, в соответствующих подпространствах, каждая из которых считается случайной переменной с нормальным распределением $N(0, I)$.

Для верификации диктора применение комбинированного факторного анализа заключается в оценке подпространств (V, D, U) по размеченному речевому корпусу и оценке факторов диктора и сессии (x, y, z) по записи речи диктора. Таким образом, удалив составляющую сессии из формулы, приведенной выше, можно представить супервектор диктора как $s = m + Vy + Dz$. Оценка соответствия речевого сигнала диктора его модели вычисляется как разность рассчитанного правдоподобия тестового отрезка речи диктора с компенсированной относительно сессии моделью диктора (M - Ux).

Подход, используемый в комбинированном факторном анализе, заключается в определении единого пространства, вместо двух пространств дикторов и каналов. Это пространство одновременно включает в себя характеристики и диктора и канала. В данной модели нет разделения между эффектами влияния диктора и эффектами влияния канала в Гауссовой смеси.

Данный метод позволяет отследить все изменения, происходящие во время адаптации математических ожиданий УФМ для заданной последовательности окон рассматриваемого отрезка речи. Эта информация моделируется в пространстве малой размерности, называемом пространством полной изменчивости. Таким образом, в данном методе каждая произнесенная диктором фраза имеет соответствующий вектор M, заданный следующим образом (1.15):

$$M = m + Tw, \tag{1.15}$$

где m – дикторо- и каналонезависимый супервектор (например универсальная фоновая модель), T – квадратная матрица малого порядка и w – случайный вектор с нормальным распределением $N(0, I)$. Компоненты вектора w являются полными факторами, а сам вектор называется вектором идентичности или i-вектором (i-vector). Данный вектор является скрытой переменной, которая может быть задана апостериорным распределением с использованием статистики Баума-Велша.

Компоненты i-вектора отражают изменения в компонентах Гауссовой смеси универсальной фоновой модели (супервектор m), произошедшие после адаптации УФМ к заданной фразе диктора. При проведении верификации диктора, i-векторы являются соответствующими отрезку речи диктора признаками. После извлечения данных векторов, они подаются на вход классификатору.

Предположим, что имеется последовательность из L окон $\{y_1, y_2, \ldots, y_L\}$ и универсальная фоновая модель Ω, состоящая из C компонент смеси, заданных в пространстве признаков размерностью F. Тогда статистику Баума-Велша (1.16,1.17), необходимую для вычисления i-вектора можно получить как [78]

$$N_C = \sum_{t=1}^{L} P(c \mid y_t, \Omega) \tag{1.16}$$

$$F_C = \sum_{t=1}^{L} P(c \mid y_t, \Omega) y_t, \tag{1.17}$$

где $c = 1,\ldots,C$ – это индекс Гауссиана, и $P(c|y_t,\Omega)$ соответствует апостериорной вероятности компоненты смеси c, генерируемой вектором y_t. Кроме того, необходимо вычислить централизованную статистику Баума-Велша первого

порядка (1.18), основанную на математических ожиданиях универсальной фоновой модели:

$$\widetilde{F}_C = \sum_{t=1}^{L} P(c \mid y_t, \Omega)(y_t - m_c), \tag{1.18}$$

где m_c – математическое ожидание компоненты Гауссовой смеси c. Таким образом, i-вектор для заданного отрезка речи можно вычислить по формуле (1.19):

$$w = (I + T^t \Sigma^{-1} N(u) T)^{-1} . T^t \Sigma^{-1} \widetilde{F}(u). \tag{1.19}$$

Зададим $N(u)$ как диагональную матрицу размерности CF x CF, диагональные блоки которой равны $N_c I$ (c = 1,...,C), $\tilde{F}_c$ – супервектор размерности CF x 1, полученный объединением всей статистики Баума-Велша первого порядка $\tilde{F}_c$ для заданного отрезка речи u. Σ – это диагональная матрица ковариации размерности CF x CF, оцениваемая во время факторного анализа [79]. Данная матрица Σ отражает остаточную вариативность, не зафиксированную матрицей полной вариации T.

Для вычисления метрики, используемой для сравнения двух i-векторов, используют несколько методов, в их числе – машина опорных векторов с применением косинусного ядра и косинусное расстояние между i-векторами. Машина опорных векторов [80] является бинарным классификатором, который пытается найти наилучший линейный разделитель между позитивными и негативными образцами. Однако, можно использовать нелинейное разделение, заменив ядро машины, в данном случае используется косинусное ядро.

Второй метод расчета расстояния между i-векторами – вычисление косинусного расстояния. При расчете косинусного расстояния между целевым и тестовым диктором, результат сравнивается с порогом, определяющим

конечное решение. Преимущество данного метода – не требуется предварительного участия диктора с обучением. Разницу между двумя дикторами можно вычислять напрямую, без дополнительных вычислений и затрат, поэтому i-вектора можно рассматривать в качестве признаков, используемых для верификации.

Для компенсации влияния канала необходимо обработать пространство полных факторов. Преимущество использования полных факторов в данном случае – более экономные вычисления, так как супервекторы Гауссовой смеси используют большую размерность. В [78] было протестировано три методики компенсации влияния канала. Первый подход – нормализация внутриклассовой ковариации [81], где производится ее инвертирование. Второй подход – линейный дискриминантный анализ. Данная методика пытается задать новые оси таким образом, чтобы нормализовать внутриклассовые изменения, вызванные влиянием канала, т.к. записи должны принадлежать одному диктору. Последний подход называется проекцией атрибутов помех [82]. Данный подход используется для поиска подходящей матрицы проекции, которая позволила бы удалить помехи из пространства диктора. Согласно данному подходу вычисляется матрица малого порядка R, основанная на собственных векторах с наибольшими собственными значениями внутриклассовой ковариации, вычисленной на множестве i-векторов. Новые i-вектора при этом проецируются на ортогональное пространству канала, пространство диктора.

При создании современных систем, использующих i-векторы для верификации диктора, наиболее часто применяется вероятностный линейный дискриминантный анализ (Probabilistic LDA) для извлечения i-векторов [83-89]. Системы, использующие подход, основанный на извлечении из аудиозаписей i-векторов, активно применяются при проведении соревнований NIST SRE [90]. В соревнованиях NIST SRE-2016 были представлены системы [91-93].

1.5 Методы верификации с применением глубоких нейронных сетей.

Одной из современных тенденций стало применение глубоких нейронных сетей (ГНС) в системах верификации диктора по голосу. Глубокие нейронные сети используются как для извлечения статистик Баума-Велша [94, 95], так и для извлечения новых признаков, которые формируются нейронной сетью в скрытом слое с меньшим количеством нейронов (Bottleneck Features, BNF). Возможно применение ГНС в качестве отдельного классификатора, обученного с целью верификации диктора [96, 97]. Кроме того, возможно применение ГНС как для извлечения BNF, предварительно обучив ГНС для распознавания речи, так и для извлечения признаков, полученных из выходного слоя ГНС [98-100]. В том числе, ГНС используют для противодействия спуфинг атакам на системы верификации диктора [101].

Обычно для рассмотренных целей применяют нейронные сети прямого распространения, которые намного больше (более тысячи нейронов в скрытом слое), и намного глубже (5-7 скрытых слоев) традиционных нейронных сетей. Для обучения ГНС применяют алгоритм обратного распространения ошибки и метод стохастического градиентного спуска.

В работе [102] применение ГНС позволило получить прирост точности распознавания диктора для микрофонной речи. Было применено два подхода: извлечение признаков, полученных с помощью ГНС, и моделирование признаков с помощью ГНС. Моделирование признаков осуществляется таким образом, что вместо универсальной фоновой модели, которая обычно используется для извлечения i-векторов, применяется ГНС.

Нетрадиционная схема применения ГНС предложена в [103]. В данной работе шумоподавляющие автоэнкодеры (Denoising Auto-Encoders, ША) используются для минимизации внутриклассовой вариативности входных i-векторов. Данный подход позволяет из заданного на входе ША i-вектора диктора, полученного с определенными характеристиками сессии, получить на

выходе усредненный i-вектор диктора. Усредненные i-векторы затем подаются на вход системе, использующей вероятностный линейный дискриминантный анализ.

Были получены результаты [100] в условиях соревнований DAC2015, в соответствии с которыми ошибка EER была уменьшена на 55% благодаря задействованной ГНС. В [98] применение ГНС, обученной для распознавания речи позволило уменьшить ошибку EER на 30 %, по условиям NIST SRE 2012 [104].

В [105] применение ША позволило уменьшить равную ошибку 1-го и 2-го рода EER на 32 % по сравнению с базовой PLDA системой (условия NIST 2012 [104]). Также в данной работе отмечаются следующие недостатки методов верификации диктора с применением ГНС: проявляется зависимость точности верификации от языка, на котором говорит диктор, а также имеется низкая устойчивость к изменениям в условиях, в которых производится аудиозапись.

1.6 Выводы

1. Выполнена постановка задачи верификации диктора по произвольной фразе. Заданы условия для проведения верификации диктора, определены гипотезы, требующие проверки, рассмотрены подходы для моделирования альтернативной гипотезы.

2. Выполнен обзор и произведен анализ речевых признаков, используемых в области верификации диктора по голосу. Сделан вывод, что кроме мел-кепстральных коэффициентов, которые наиболее часто используются для верификации диктора по голосу, существуют другие признаки, которые могут улучшить точность верификации. Данные признаки позволяют получить дополнительную информацию о дикторе. Следовательно, необходимо найти такой набор признаков, который позволит достичь наилучшей точности верификации.

3. Рассмотрены современные методы верификации диктора по произвольной фразе. К основным методам относят метод верификации основанный на применении Гауссовых смесей (GMM-UBM системы), метод верификации, основанный на факторном анализе с применением i-векторов, а также методы верификации диктора, основанные на применении глубоких нейронных сетей.

4. С целью повышения точности систем верификации диктора по произвольной фразе, применяют несколько подходов. К ним относят разработку новых или объединение уже существующих методов выделения признаков из речи, разработку новых методов верификации, а также разработку новых методов построения решающих правил (классификаторов).

2. Алгоритмы и программные средства верификации диктора по произвольной фразе

2.1 Алгоритм верификации на базе Гауссовых смесей и универсальной фоновой модели

Рассмотрим базовый алгоритм верификации на базе Гауссовых смесей и универсальной фоновой модели (Рисунок 2.1.1). В данный алгоритм включены четыре последовательных шага. Первым шагом алгоритма является извлечение признаков из аудиозаписей речи, используемой для обучения как УФМ, так и моделей дикторов. На данном этапе обычно производится обработка сигнала, разбиение всего сигнала на отрезки и другие действия, в зависимости от типа извлекаемых признаков.

Следующим шагом алгоритма является обучение универсальной фоновой модели. Данная модель является Гауссовой смесью, обученной на речевом материале большого количества дикторов (раздел 1.3). Результатом данного шага алгоритма является обученная Гауссова смесь $\lambda_{УФМ}$.

После обучения УФМ производится обучение моделей дикторов, зарегистрированных в системе, с помощью адаптации от УФМ.

Последним шагом алгоритма является тестирование моделей. На данном этапе вычисляется логарифм отношения правдоподобия между заданной моделью диктора и универсальной фоновой моделью для заданного речевого сигнала.

Рисунок 2.1.1 – Блок-схема алгоритма верификации диктора на базе Гауссовых смесей и универсальной фоновой модели

Рассмотрим подробнее процесс извлечения речевых признаков (Рисунок 2.1.2). На первом шаге процесса извлечения речевых признаков из аудиозаписи производится разделение ее на короткие временные отрезки (окна) – маленькие части речевого аудиосигнала. Данные окна обрабатываются по отдельности, обработка всего сигнала целиком не производится. Длина такого окна составляет 20 мс, а смещение, по которому сигнал разбивается на окна, составляет 10 мс [1]. После этого, производится предобработка сигнала (фильтр верхних частот) и умножение на оконную функцию Хэмминга [16]. Из аудиозаписей извлекаются такие признаки, как мел-кепстральные коэффициенты, пары линейного спектра, кепстральные коэффициенты

перцептивного линейного предсказания, энергия сигнала, формантные частоты, частота основного тона, вероятность вокализации (максимум автокорреляционной функции спектра в окне), частота пересечения нуля, джиттер и шиммер. Для извлечения речевых признаков из аудиозаписей голоса диктора в данной работе была использована библиотека openSMILE [106]. Диаграмма процесса извлечения признаков из звукового сигнала, а также извлекаемые признаки, изображены на Рисунке 2.1.2.

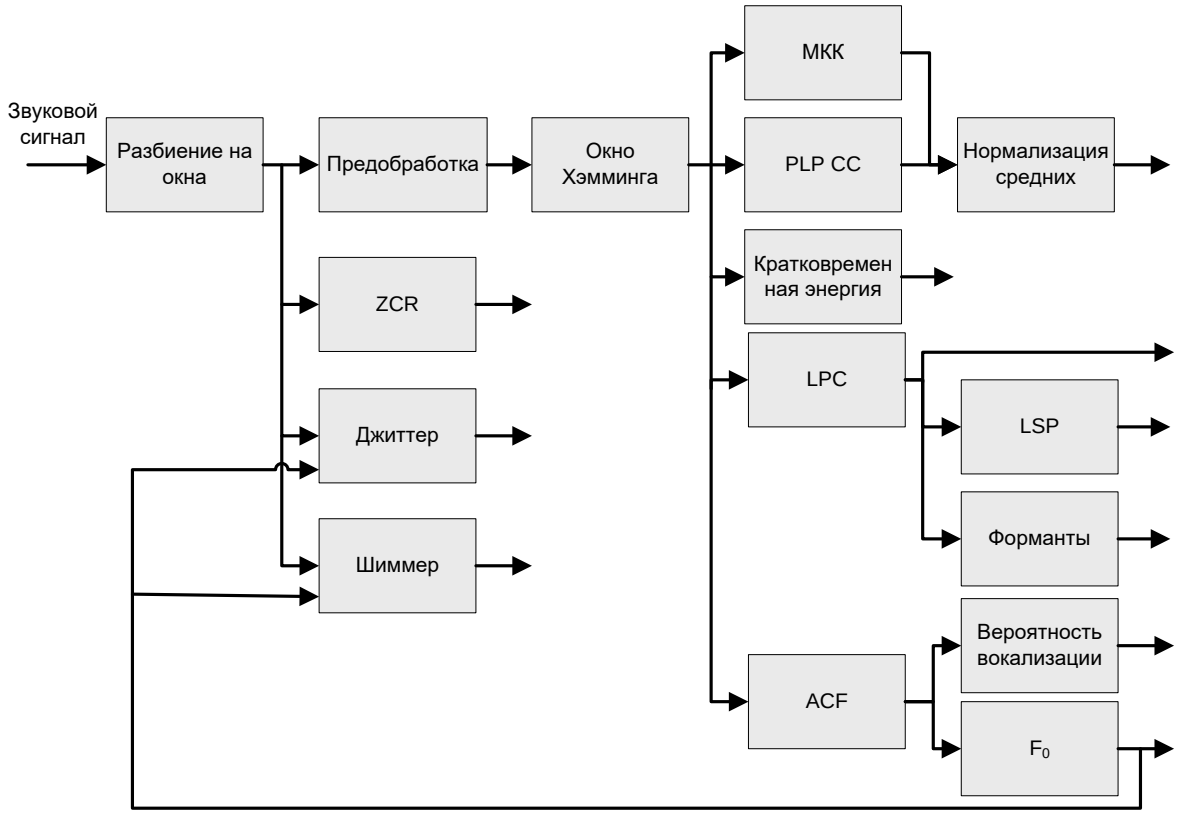

Рисунок 2.1.2 – Диаграмма процесса извлечения признаков из звукового сигнала

Рассмотрим подробнее алгоритм обучения УФМ (Рисунок 2.1.3). Как и в [1], УФМ была обучена с помощью EM-алгоритма. Основная идея данного алгоритма заключается в том, что используя начальную модель λ, произвести оценку новой модели λ', чтобы $p(X|\lambda') \geq p(X|\lambda)$. Новая модель становится начальной моделью для следующей итерации алгоритма и процесс повторяется пока не будет достигнуто максимальное количество итераций [107].

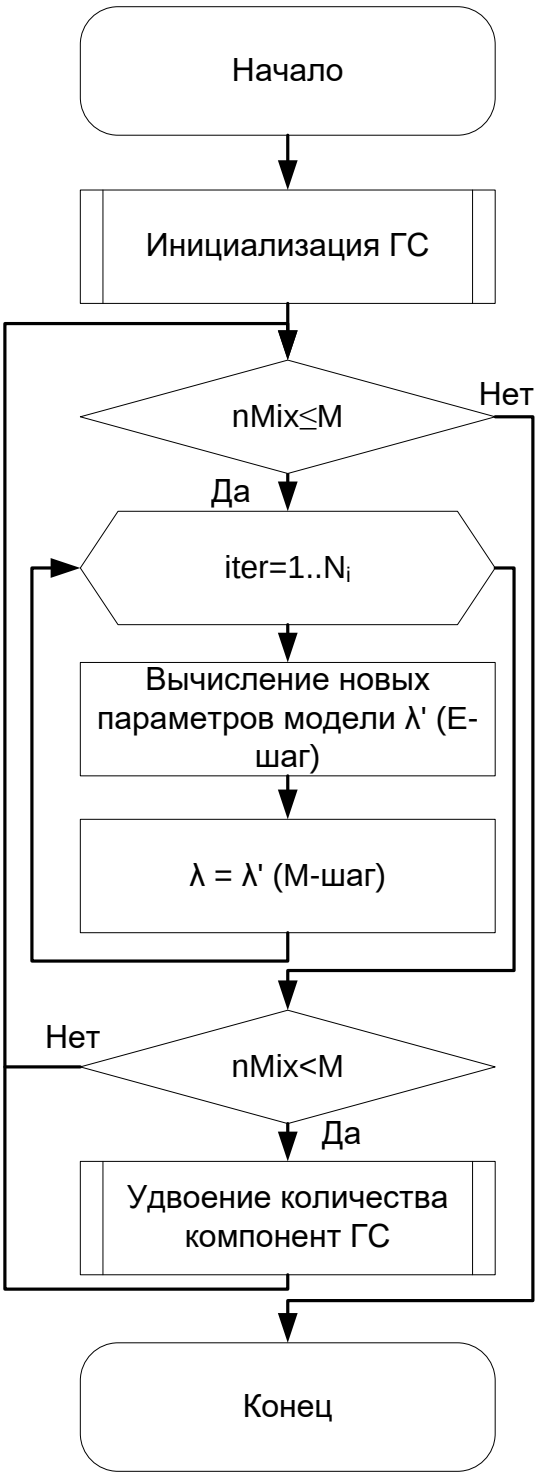

Рисунок 2.1.3 – Блок-схема алгоритма обучения УФМ

На каждой итерации EM-алгоритма происходит оценка параметров Гауссовой смеси (2.1.1-2.1.3), тем самым монотонно увеличивая правдоподобие модели:

$$w_i' = \frac{1}{T}\sum_{t=1}^{T} \Pr(i|x_t, \lambda) \quad (2.1.1)$$

$$\mu_i' = \frac{\sum_{t=1}^{T} \Pr(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} \Pr(i|x_t, \lambda)} \quad (2.1.2)$$

$$\sigma_i^{2\prime} = \frac{\sum_{t=1}^{T} \Pr(i|x_t, \lambda) x_t^2}{\sum_{t=1}^{T} \Pr(i|x_t, \lambda)} - \mu_i{}^{2\prime}. \quad (2.1.3)$$

Апостериорная вероятность для i-ой компоненты смеси $\Pr(i|x_t, \lambda)$ вычисляется как (2.1.4):

$$\Pr(i|x_t, \lambda) = \frac{w_i p_i(x_t|\mu_i, \Sigma_i)}{\sum_{k=1}^{M} w_k p_i(x_t|\mu_k, \Sigma_k)}, \quad (2.1.4)$$

где $p_i(x_t|\mu_i, \Sigma_i)$ вычисляется согласно выражению 1.5. Инициализация Гауссовой смеси производится таким образом: вес единственной компоненты смеси устанавливается равным 1, μ_i и Σ_i – математическое ожидание и дисперсия, вычисленные по обучающим данным. Обучение Гауссовой смеси УФМ производится с последовательным расщеплением и удвоением количества компонент смеси. Таким образом, начинается обучение со смеси, состоящей из одной компоненты, последовательно доходя до M компонент. При этом производится N_i EM-шагов. Данное удвоение производится согласно [107].

Для обучения моделей дикторов была использована оценка апостериорного максимума (MAP адаптация) [1]. Данная методика используется в области верификации диктора для адаптации модели диктора из универсальной фоновой модели. Также как и в EM-алгоритме, адаптация состоит из двух шагов (Рисунок 2.1.4). На первом шаге производится вычисление необходимых для адаптации модели параметров $Pr(i|x_t,\lambda)$, n_i, $E_i(x)$ (согласно выражениям 1.6-1.8). Однако, на втором шаге вычисленные

параметры смешиваются с исходными параметрами, взятыми из УФМ, используя коэффициент α_i (выражения 1.10-1.13).

Рисунок 2.1.4 – Блок-схема алгоритма адаптации модели диктора

На заключительном этапе алгоритма верификации диктора осуществляется тестирование модели диктора λ_S. На этом этапе для каждого тестового речевого сегмента определяется оценка вероятности принадлежности речевого сигнала X модели диктора S - λ_S или фоновой модели $\lambda_{УФМ.}$ Затем вычисляется логарифм отношения правдоподобия (2.1.5):

$$\Lambda(X) = \log p(X \mid \lambda_S) - \log p(X \mid \lambda_{УФМ}), \qquad (2.1.5)$$

где X – тестируемый отрезок речи, λ_S – модель предполагаемого диктора, $\lambda_{УФМ}$ - универсальная фоновая модель. Предполагаемый диктор принимается или отвергается системой на основе заданного порога принятия решения Θ (формула 1.1). Если полученное значение больше порога, считаем, что на заданном речевом сигнале присутствует речь диктора S, иначе – речь диктора S отсутствует (Рисунок 2.1.5).

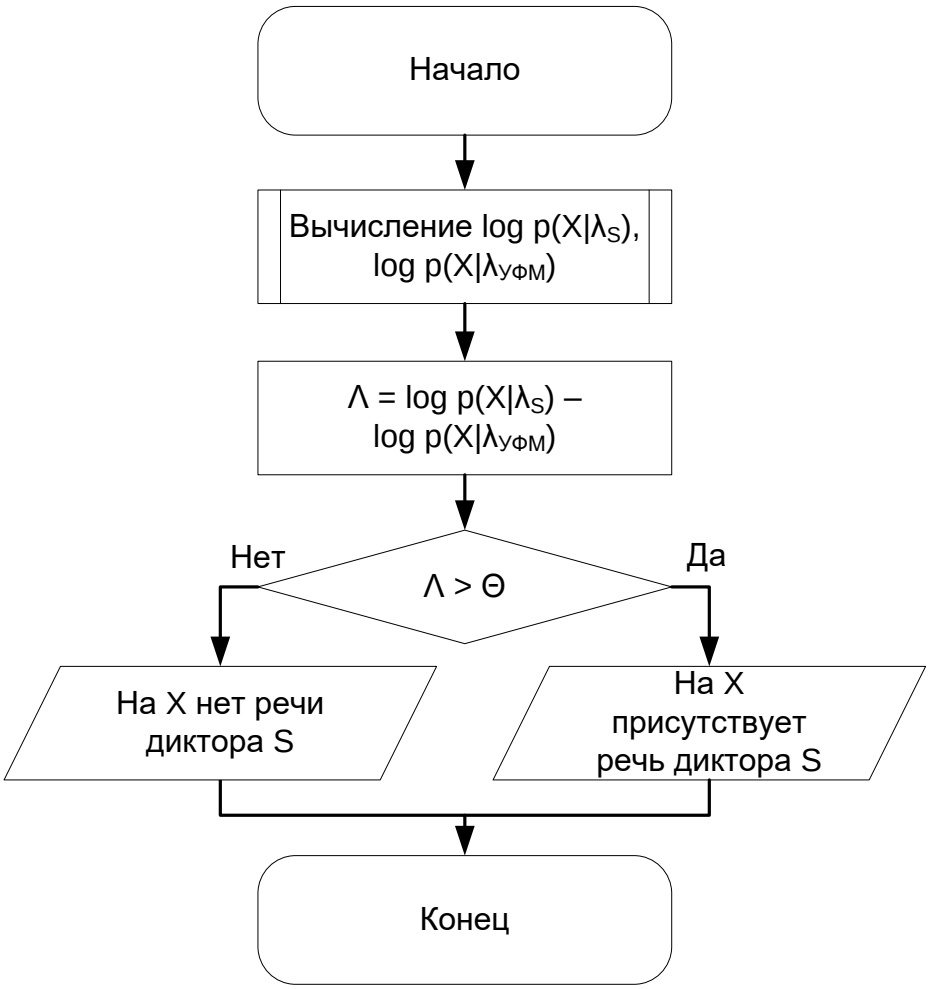

Рисунок 2.1.5 – Блок-схема алгоритма тестирования модели диктора

Для оценки системы голосовой верификации используются две различных метрики: равная ошибка 1-го и 2-го рода EER и минимальная функция стоимости обнаружения (minimum detection cost function, minDCF) с параметрами SRE 2008 [108]. Равная ошибка 1-го и 2-го рода EER вычисляется таким образом: для всех тестовых данных вычисляется логарифм отношения правдоподобия $\Lambda(X)$ и подбирается такой порог Θ, чтобы количество ошибок 1-го и 2-го рода для заданных тестов было одинаково.

Функция стоимости обнаружения вычисляется как взвешенная сумма вероятности отказа целевому диктору P_{fa} и вероятности пропуска самозванца P_{miss} (2.1.6). Соответственно, минимум данной функции определяется по полученным оценкам данных вероятностей (по ошибке 1-го и 2-го рода).

$$DCF = 0{,}1P_{miss} + 0{,}01P_{fa}. \quad (2.1.6)$$

В рамках данной работы была использована ГС, состоящая из M = 256 компонент. Автором было замечено, что при проведении экспериментов на используемых в работе речевых корпусах EER не уменьшается при увеличении компонент смеси. Модели дикторов были получены с помощью MAP адаптации, с адаптацией только векторов математических ожиданий и фактором релевантности r = 10. Согласно [1] были получены результаты, согласно которым адаптация математических ожиданий компонент УФМ, без адаптации весов и дисперсий, дает наилучший результат.

Экспериментальная оценка. Для оценки точности системы верификации диктора на базе модели Гауссовой смесей и универсальной фоновой модели, были проведены эксперименты с применением речевого корпуса, включающего записи речи 25 дикторов-мужчин и 25 женщин [109]. Данный речевой корпус содержит записи произнесенных без предварительной подготовки предложений, взятых из художественной литературы, или поговорок. Суммарная длина записей речи для каждого диктора составляет не менее 6 минут, включая 50 сегментов различной длины. Каждый диктор был записан на микрофон в условиях нешумной аудитории. Аудиозаписи имеют следующие параметры: частота дискретизации 8000 Гц, разрядность 16 бит.

Весь речевой корпус, состоящий из записей речи 50 дикторов, был разделен на две части – одна для обучения УФМ (состоит из записей 30 дикторов), вторая – для обучения и тестирования моделей дикторов (состоит из записей оставшихся 20 дикторов). Обе части включают в себя равное количество дикторов мужчин и женщин.

Для MAP адаптации моделей дикторов использовались 40 речевых сигналов. Оставшиеся 10 сигналов каждого диктора применялись для

тестирования системы верификации. В сумме, было произведено 4000 тестов для каждого набора признаков, по 10 положительных (тестируется целевой диктор) и 190 отрицательных (тестируется диктор-нарушитель) для каждого диктора. Схема процесса верификации, используемая при тестировании, изображена на Рисунке 2.1.6. GMM-UBM система, описанная в текущем разделе, была создана с применением библиотеки MSR Identity Toolbox [107]. Результаты экспериментов представлены в Таблице 2.1.1.

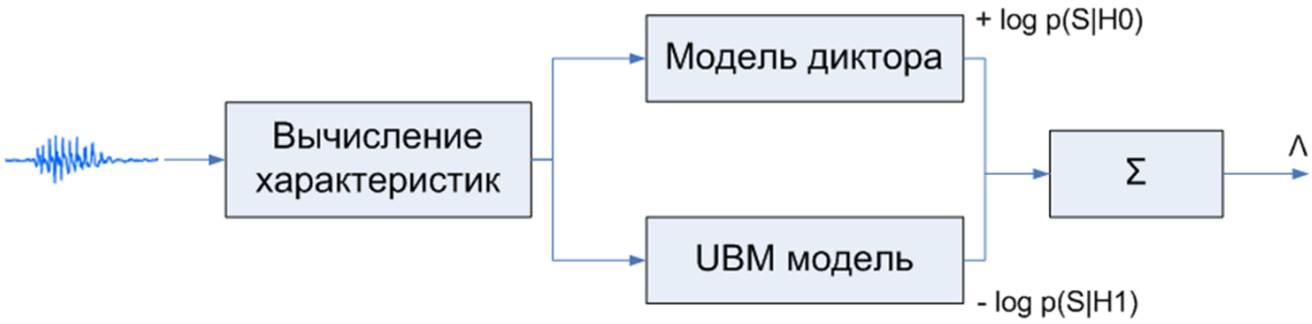

Рисунок 2.1.6 – Схема процесса верификации речевого сегмента

Как видно из Таблицы 2.1.1, наилучшие результаты были получены, используя вектор признаков, состоящий из 14 мел-кепстральных коэффициентов и вероятности вокализации с равной ошибкой 1-го и 2-го рода EER = 0,763 %. Минимальное значение DCF было получено с помощью набора признаков, состоящего из 14 мел-кепстральных коэффициентов, их дельт и вероятности вокализации.

Таблица 2.1.1 – Результаты оценки точности GMM-UBM системы верификации диктора с применением различных наборов признаков

Набор признаков	% EER	minDCF*100
MFCC+V_p	**0,763**	0,805
MFCC+Δ+V_p	1,000	**0,699**
MFCC+Δ+ΔΔ+V_p	1,000	0,803
MFCC	1,000	0,925
MFCC+Shimmer	1,000	1,007
MFCC+Δ	1,052	0,825
MFCC+JitterDDP	1,131	1,003
MFCC+Zcr	1,131	1,031
MFCC+F_0	1,157	1,161

На Рисунке 2.1.7 изображены графики кривых компромиссного определения ошибки (DET кривые) [108] для двух наборов признаков – МКК и МКК с вероятностью вокализации. Данная кривая была получена согласно выражению 1.1, путем изменения порога принятия решения. Каждая точка кривой соответствует полученным ошибкам 1-го и 2-го рода при фиксированном пороге.

Было выяснено (Таблица 2.1.1), что при добавлении некоторых признаков к стандартному вектору, состоящему из 14 МКК, результаты верификации были хуже, чем при использовании вектора только из МКК. Кроме того, можно заметить, что при добавлении вероятности вокализации к вектору признаков, происходит уменьшение ошибки EER или уменьшение minDCF. Таким образом, можно сделать вывод, что добавление вероятности вокализации в вектор признаков улучшает эффективность работы GMM-UBM системы верификации диктора.

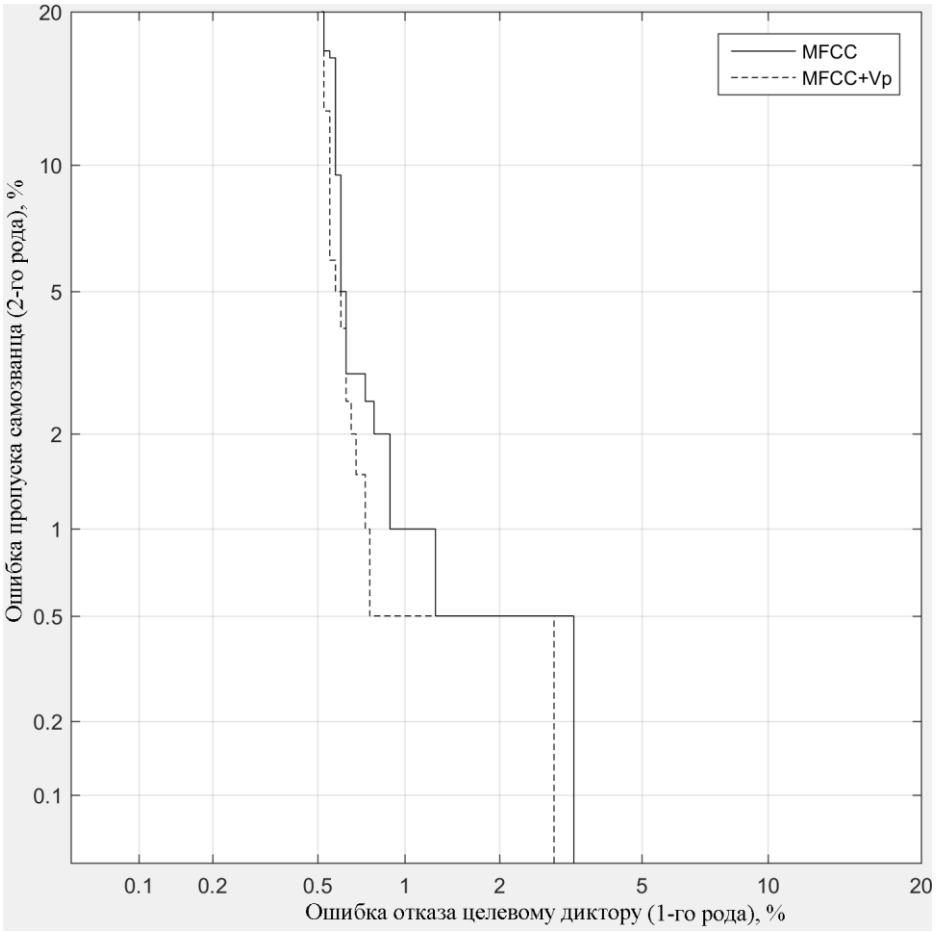

Рисунок 2.1.7 – Кривые компромиссного определения ошибки (DET кривые) для МКК и МКК с вероятностью вокализации

В данном разделе был рассмотрен алгоритм верификации диктора на базе Гауссовых смесей и универсальной фоновой модели, а также разработана система верификации диктора, основанная на данном алгоритме. Была произведена оценка системы как с применением стандартных наборов признаков, включающих в себя мел-кепстральные коэффициенты и их дельты, так и других признаков. Наилучшие показатели точности системы верификации были получены при использовании вектора из 14 мел-кепстральных коэффициентов и вероятности вокализации, равная ошибка 1-го и 2-го рода EER составляет 0,763 %.

2.2 Исследование признаков с применением генетического алгоритма и жадного алгоритма добавления-удаления

Полученный в предыдущем разделе набор признаков нельзя считать наилучшим по точности, т.е. дающим наименьшую ошибку EER, так как были рассмотрены только некоторые наборы признаков (Таблица 2.1.1), выбранные автором вручную. Выбрать наилучший набор признаков методом полного перебора в данном случае невозможно, так как для используемого количества признаков $n = 94$, общее количество непустых подмножеств составляет 2^n-1, что вычислительно слишком сложно.

Для решения данной задачи, обычно используются различные методы отбора признаков, разделяемые на три категории – методы фильтрации, оберточные методы и встроенные методы [110]. В данном случае необходимо использовать оберточные методы [111], так как методы фильтрации напрямую не применимы – используемые признаки не позволяют сразу провести классификацию, а для метода классификации с применением Гауссовых смесей отсутствуют встроенные методы отбора признаков. К оберточным относят такие методы, как метод последовательного добавления признаков, метод последовательного сокращения признаков [112], поиск в глубину (метод ветвей и границ), поиск в ширину (метод группового учёта аргументов), генетический алгоритм, случайный поиск с адаптацией [113], различные эволюционные алгоритмы [114].

Отбор признаков позволяет снизить переобучение модели в тех случаях, когда их используется слишком много. Для этого методы отбора сохраняют наиболее информативные признаки. Если используется слишком мало признаков, то методы отбора позволяют добавить информативные признаки, увеличив тем самым точность системы.

В рамках данной работы были использованы два известных метода отбора признаков, а именно метод жадного добавления-удаления (Add-del) и

генетический алгоритм (ГА) [115]. Данные методы отбора признаков позволяют получить положительные результаты оптимизации, затратив небольшое количество итераций, и, следовательно, времени. Время, затрачиваемое на тестирование и оценку точности системы с заданным набором признаков, зависит от количества используемых признаков, и в среднем составляет около 250 секунд. Соответственно, алгоритмы, требующие проведения большого количества итераций не рассматривались.

Жадный алгоритм (ЖА) добавления-удаления признаков [116], включает в себя две жадные стратегии, т.е. производится поочередное добавление и удаление признаков из текущего множества. Сначала алгоритм добавления Add последовательно добавляет признаки (*method* = 0), до тех пор, пока не начнет увеличиваться ошибка EER и еще $d = 3$ шагов с увеличением ошибки. После этого начинает работу алгоритм жадного удаления Del (*method* = 1), который удаляет избыточные признаки. Блок схема алгоритма жадного добавления-удаления изображена на Рисунке 2.2.1.

Основной цикл данного алгоритма ограничен *maxSteps* итерациями, поэтому производится выход из цикла либо по достижении этого значения, либо, когда прошло d шагов с увеличением ошибки err_i. Выходом алгоритма является набор признаков F.

Автором был опробован модифицированный алгоритм (Рисунок 2.2.2), позволяющий проводить цикличную смену алгоритмов Add и Del (изначально алгоритм сменяется только один раз), однако применение этого подхода привело к постоянному добавлению и удалению одних и тех же признаков из набора без улучшения точности системы верификации.

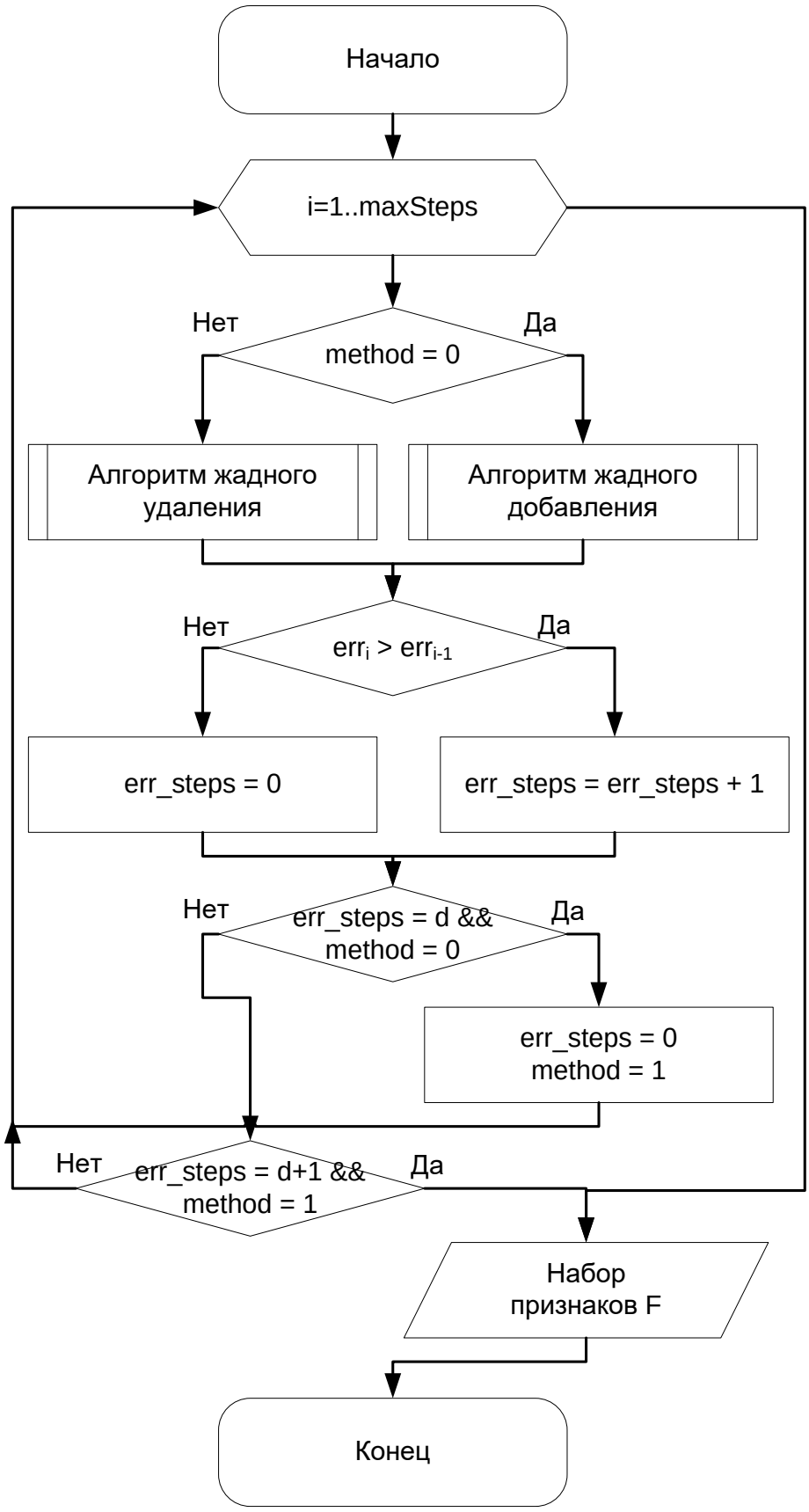

Рисунок 2.2.1 – Блок-схема алгоритма жадного добавления-удаления признаков

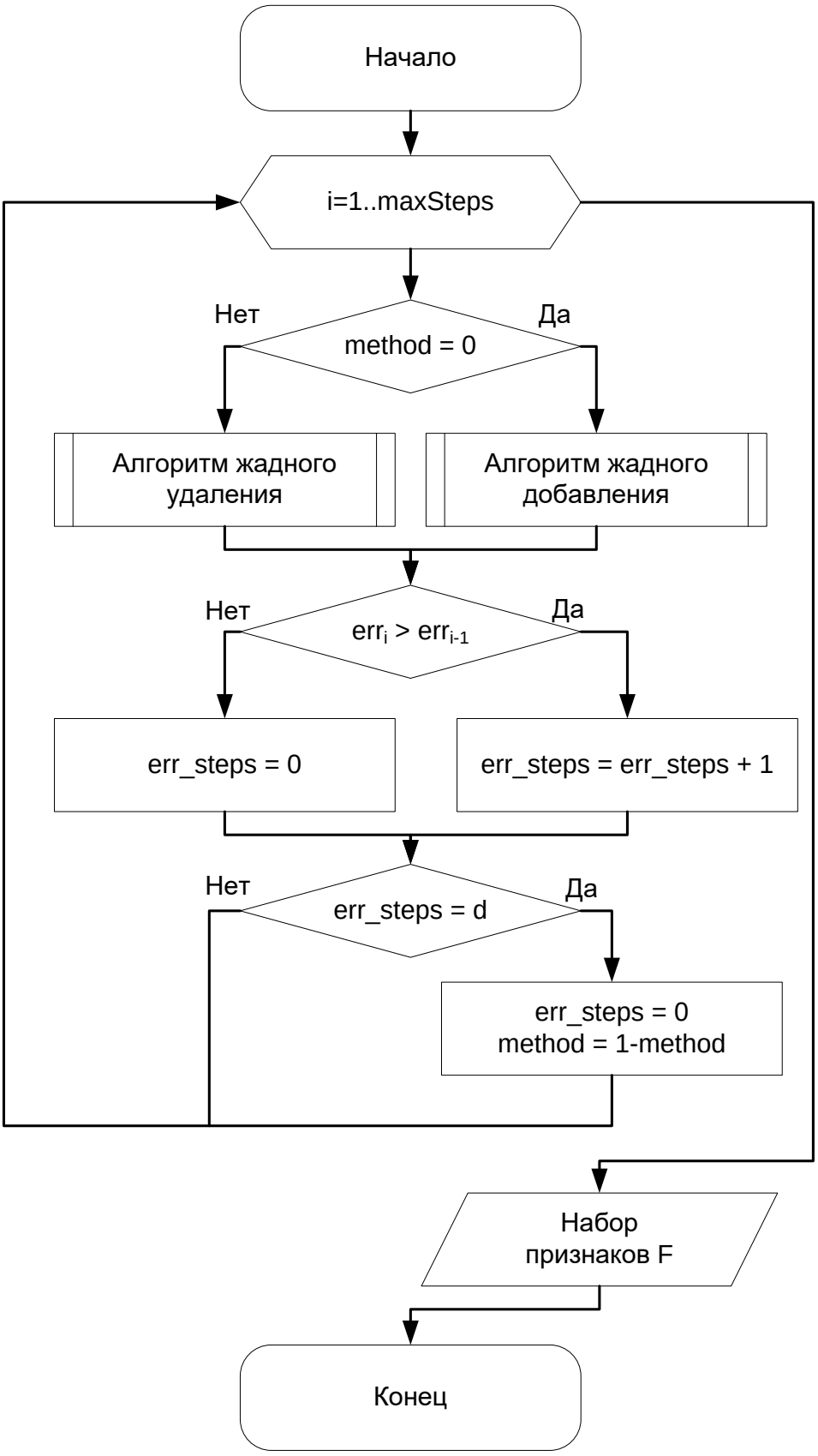

Рисунок 2.2.2 – Блок-схема модифицированного алгоритма жадного добавления-удаления признаков

Рассмотрим алгоритм жадного добавления признаков. Данный алгоритм основан на последовательном добавлении признаков, начиная с пустого множества признаков *features* = {}. На каждой итерации алгоритма (Рисунок 2.2.3) производится добавление только одного признака в исходное множество, причем добавляется такой признак *feature*[i], при добавлении которого точность верификации диктора наибольшая, т.е. минимальна ошибка EER и функция стоимости обнаружения minDCF.

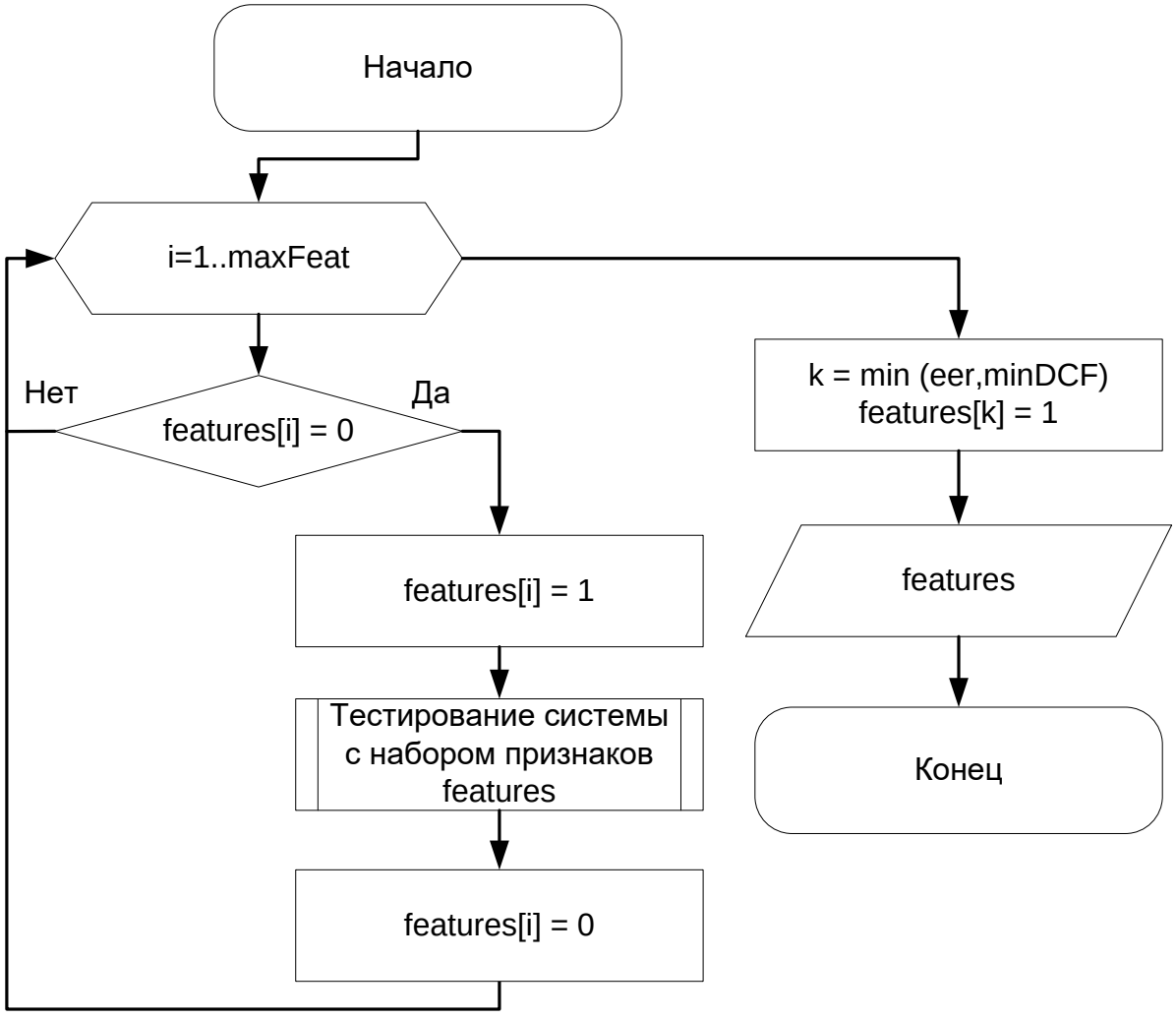

Рисунок 2.2.3 – Блок-схема итерации алгоритма жадного добавления признаков

В отличие от алгоритма жадного добавления признаков, алгоритм жадного удаления основан на последовательном удалении признаков, начиная с исходного множества признаков *features* = {$feature_1$,..$feature_{Max}$}. На каждой итерации алгоритма (Рисунок 2.2.4) производится удаление одного признака из множества *features*. Перебор удаляемых признаков производится таким образом, чтобы найти признак $feature_i$, минимизирующий ошибку EER и функцию стоимости обнаружения minDCF.

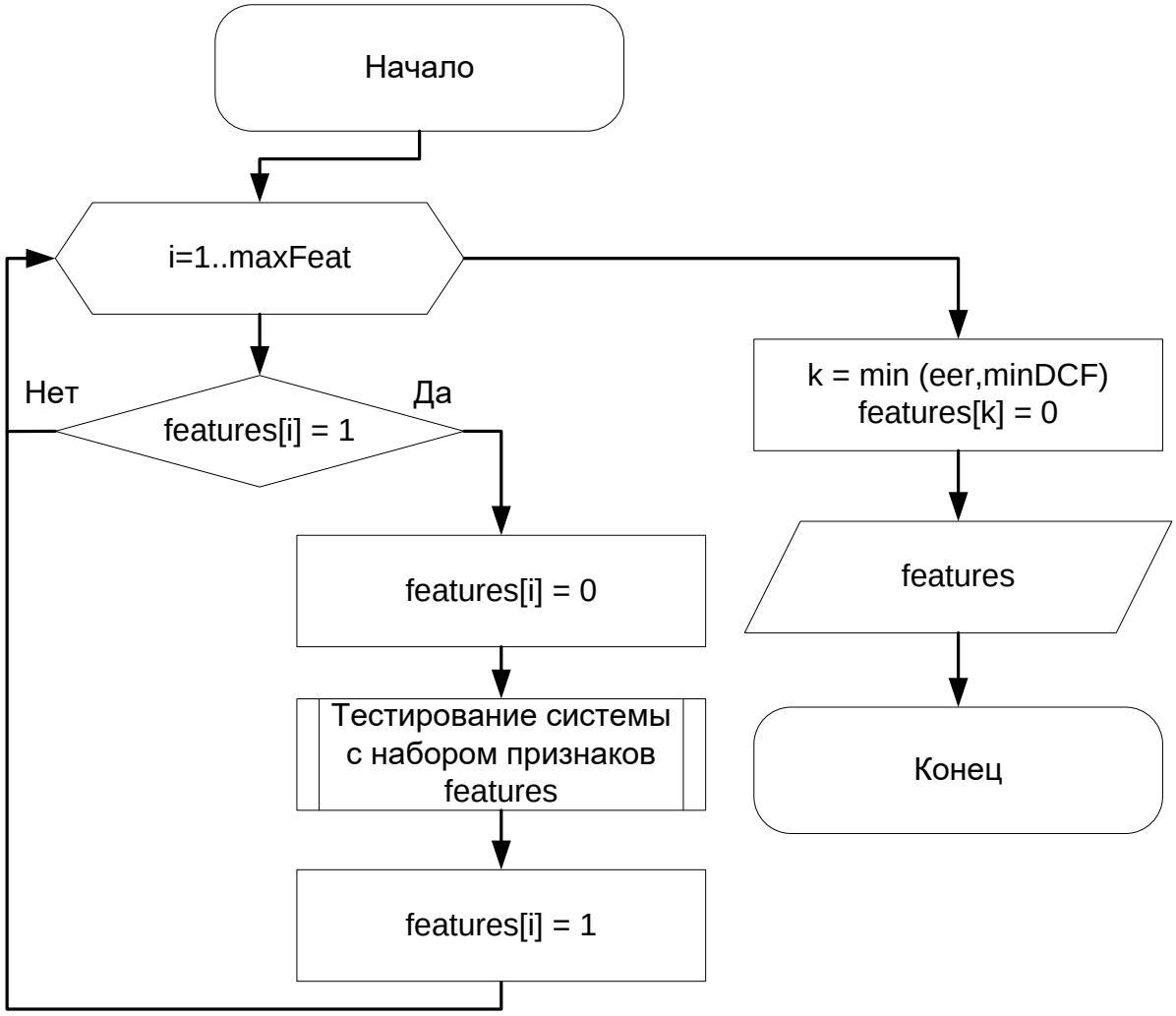

Рисунок 2.2.4 – Блок-схема итерации алгоритма жадного удаления признаков

Генетический алгоритм [117] осуществляет поиск наилучшего набора признаков с использованием методов естественной эволюции. Случайным образом формируется несколько наборов признаков, называемых индивидами,

которые объединяются в популяцию. К полученным индивидам случайным образом применяются операции мутации и скрещивания (кроссовера), таким образом получая новые индивиды. В конце каждой итерации генетического алгоритма производится отбор лучших индивидов, для которых значение целевой функции является наилучшим.

Несмотря на то, что генетический алгоритм позволяет достаточно быстро получить некоторый результат, его недостатками являются медленная сходимость и сложный подбор параметров. Рассмотрим подробнее данный алгоритм отбора признаков.

Перед выполнением основного цикла ГА необходимо задать начальную популяцию $Pop = \left\{X_1, X_2, \dots, X_{N_{Pop}}\right\}$ размером N_{Pop} особей. Геномы X_i для каждой особи i задаются случайным образом, так, чтобы бинарный ген $x_{ji} \in \{0,1\}$. Если ген $x_{ji} = 0$, значит соответствующий признак отсутствует в данной особи, если $x_{ji} = 1$, то присутствует. Общее количество генов в геноме соответствует количеству признаков, участвующих в отборе.

После этого начинается основной цикл ГА (Рисунок 2.2.5), в ходе которого начальная популяция Pop дополняется новыми особями с применением процедур скрещивания Pop_c и мутации Pop_m. Скрещивание производится попарно для N_c особей, таким образом, что для каждой пары особей образуется еще пара новых индивидов-потомков, с генами, взятыми от их родителей. В данном случае было использовано одноточечное скрещивание с выбором родителей методом рулетки, точка скрещивания выбирается случайным образом [118].

При проведении операции мутации производится инвертирование генов N_m особей основной популяции Pop, таким образом, чтобы каждый из генов инвертировался с вероятностью P_m. Таким образом образуется N_m особей-мутантов, формирующих популяцию Pop_m.

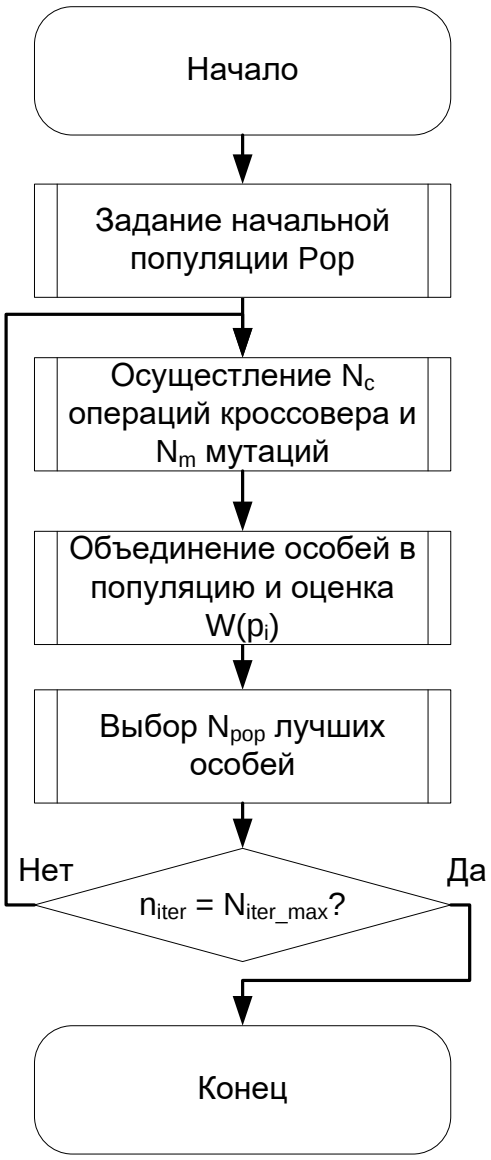

Рисунок 2.2.5 – Блок-схема генетического алгоритма отбора признаков

Далее производится объединение особей в единую популяцию $Pop_{new} = \{Pop, Pop_c, Pop_m\}$ и оценка приспособленности W индивидов pi: $W(p_i)$. В данной работе функция приспособленности W должна отражать точность системы верификации диктора с набором признаков индивида p_i. В качестве функции приспособленности была выбрана функция (2.2.1):

$$W(p_i) = 100EER(p_i) + minDCF(p_i) \quad (2.2.1)$$

Таким образом, приоритетным для данной функции является минимизация EER, однако, при равной ошибке EER у двух особей, будет отобрана особь с меньшим параметром minDCF.

Отбор особей в полученной популяции Pop_{new} осуществлялся методом ранжирования. Для этого особи популяции p_i отсортировываются по значению функции приспособленности $W(p_i)$ в порядке возрастания, и в начальную популяцию отбираются первые N_{Pop} особей. Перед проведением отбора из популяции удаляются особи с одинаковым генотипом.

Заканчивается основной цикл ГА проверкой номера текущей итерации. Если цикл достиг заданного максимального количества итераций $n_{iter} = N_{iter\_max}$, то работа алгоритма заканчивается, иначе начинается новая итерация цикла.

Экспериментальная оценка. В данной работе были использованы следующие параметры ГА: размер популяции N_{Pop} = 10, число особей, подвергающихся скрещиванию N_c = 8, число особей-мутантов N_m = 8, вероятность мутации гена P_m = 0,1, количество итераций ГА N_{iter_max} = 40.

При проведении эксперимента были использованы следующие алгоритмы: жадный алгоритм Add-del, генетический алгоритм (40 итераций), ЖА Add-del, основанный на наилучшем индивиде, полученном по результатам ГА, и ЖА Add-del, основанный на наилучшем наборе признаков из Таблицы 2.1.1 (MFCC+V_p). Полученные результаты представлены в Таблице 2.2.1.

Анализ Рисунка 2.2.6 показал, что алгоритм Add-del достаточно быстро уменьшает ошибку EER, однако ниже 2 % данная ошибка не опускается. Задав в качестве начального набора признаков вектор, состоящий из 14 МКК и вероятности вокализации, минимум ошибки EER был достигнут на 10 шаге ЖА (Рисунок 2.2.7).

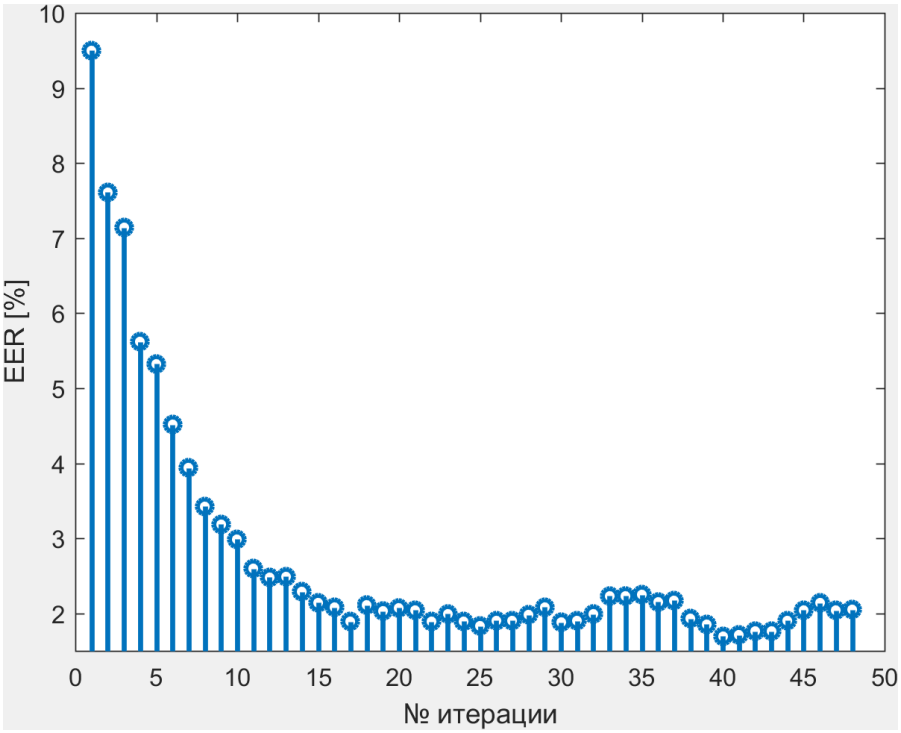

Рисунок 2.2.6 – График изменения ошибки EER для ЖА Add-Del

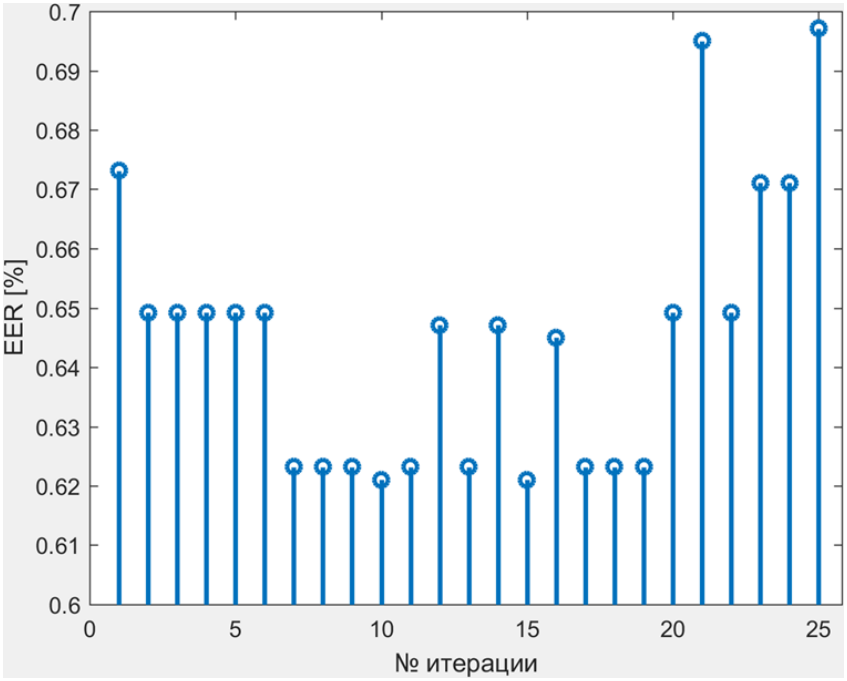

Рисунок 2.2.7 – График изменения ошибки EER ЖА Add-Del для начального вектора признаков MFCC+Vp

Генетический алгоритм был ограничен 40 итерациями, так как уже после третьей итерации наилучший генотип давал EER = 1 %, без дальнейшего уменьшения данного показателя. При этом в последующих итерациях несколько уменьшилась функция minDCF. Это может говорить о том, что алгоритм быстро сходится к локальному оптимуму. Попытки перезапуска ГА с целью получения случайных начальных популяций, не позволили выйти из локального оптимума.

Применение алгоритма Add-Del после генетического алгоритма, позволило уменьшить количество используемых признаков с 50 до 37, однако не дало улучшений в точности системы, EER остался равен 1 %.

Проведя анализ результатов применения алгоритмов отбора признаков (Таблица 2.2.1) было обнаружено, что наилучшую точность системы верификации дает набор из 28 признаков, полученных методом жадного добавления-удаления, для которого за основу был взят вектор из 14 мел-кепстральных коэффициентов и вероятности вокализации. Из данного вектора был исключен двенадцатый мел-кесптральный коэффициент, а также добавлены десять дельта и два двойных дельта мел-кестральных коэффициентов, а также 1 коэффициент линейного предсказания и 1 коэффициент LSP (Рисунок 2.2.8). Графики кривых компромиссного определения ошибки для двух наборов признаков – МКК и полученного набора из 28 признаков представлены на Рисунке 2.2.9.

Таблица 2.2.1 – Результаты применения алгоритмов отбора признаков

Набор признаков	Количество признаков	% EER	minDCF*100
ЖА Add-del (MFCC+V_p)	28	**0,579**	0,623
MFCC+V_p	15	0,763	0,805
ГА	50	1,000	**0,539**
ГА, ЖА Add-del	37	1,000	0,593
MFCC (базовый)	14	1,000	0,925
ЖА Add-del	22	2,079	1,827

13MFCC	10 Δ MFCC	2 ΔΔ MFCC	1 LPC	1 LSP	Vp

Рисунок 2.2.8 – Набор признаков, полученный с помощью ЖА Add-del

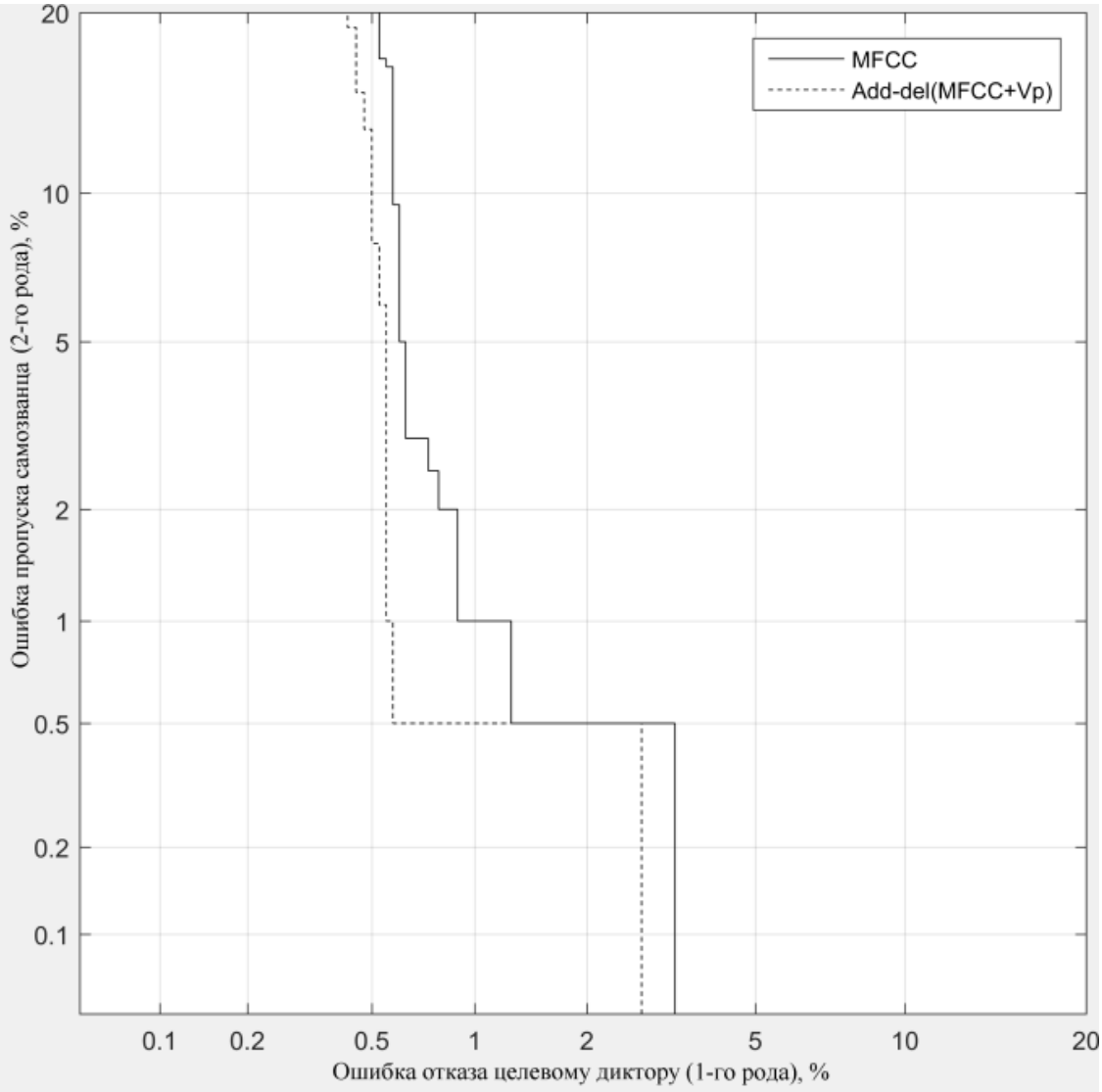

Рисунок 2.2.9 – Кривые компромиссного определения ошибки (DET кривые) для МКК и набора признаков, полученного с помощью ЖА Add-del

Полученный алгоритм верификации, использующий наилучший набор признаков из Таблицы 2.2.1, был также протестирован и сравнен согласно

условиям [119]. При проведении данного эксперимента был использован речевой корпус CHAINS [120]. Этот корпус содержит аудиозаписи 36 дикторов высокого качества. Запись каждого диктора производилась двумя сессиями, перерыв между которыми составляет от 2 до 3 месяцев. В работе было использовано только условие Solo (normal).

Опишем методику проведения эксперимента [119], которая была повторена в данной работе. Точность системы верификации определяется усредненной оценкой для 10 тестов (10-кратная кросс-валидация) с доверительным уровнем 95%, во время каждого из которых разбиение выборок происходит случайным образом. Все дикторы, входящие в речевой корпус, разделяются произвольно на 3 группы: 8 целевых дикторов, 4 диктора-нарушителя, 24 диктора, используемых только для обучения УФМ. Для каждого из дикторов использована обучающая выборка длительностью 60 секунд, также выбранная случайным образом. Для тестирования было выбрано 30 секунд речи каждого диктора, не использованной для обучения.

После проведения эксперимента были получены следующие результаты (Таблица 2.2.2). Авторы статьи [119] использовали два набора признаков – 15 мел-кепстральных коэффициентов с их дельтами и коэффициенты pykfec [121] (pyknogram frequency estimates frequency coefficient). По сравнению с представленными авторами статьи результатами (первые две строки Таблицы 2.2.2.), с полученным с помощью ЖА набором признаков была получена 100% точность верификации. Для набора из 14 мел-кепстральных коэффициентов была получена средняя точность верификации 99.4±0.36%. Отличие данной оценки можно объяснить разницей в количестве компонент ГС, у авторов статьи использовалась ГС с 64 компонентами, в данной работе – 256 компонент.

Таблица 2.2.2 – Сравнение точности систем верификации с различными наборами признаков на речевом корпусе CHAINS.

Набор признаков	Количество признаков	Точность системы [%]
15MFCC+Δ	30	98±2
Pykfec (Pyknogram)	40	99±1
ЖА Add-del(MFCC+V_p)	**28**	**100**
14MFCC	**14**	**99.4±0.36**

Таким образом, можно говорить об успешности применения набора признаков, полученного с помощью ЖА для решения задачи повышения точности системы верификации диктора по произвольной фразе. Блок-схема алгоритма верификации, использующего данный набор признаков, представлена на Рисунке 2.2.10.

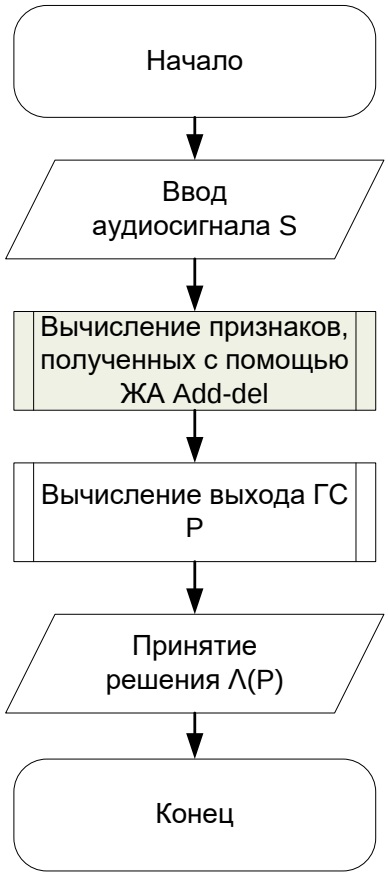

Рисунок 2.2.10 – Блок-схема алгоритма верификации диктора с применением признаков, полученных с помощью ЖА

2.3 Алгоритм генерации признаков, основанный на применении сверточной глубокой сети доверия

Одним из современных подходов, используемых для извлечения из данных комплексных, высокоуровневых признаков, являются методы глубокого обучения. Данные методы сначала извлекают низкоуровневые признаки, на основе которых строят признаки более высокого уровня, затем данный процесс может повторяться на еще более высоких уровнях представления. Тем самым данные методы пытаются построить представление об изучаемых объектах, извлекая признаки на основе иерархического подхода.

В основном, при проведении верификации диктора используются низкоуровневые признаки, например те же мел-кепстральные коэффициенты, однако предпринимаются попытки использования более высокоуровневых признаков, например применение Bottleneck Features, построение i-векторов на основе низкоуровнего представления и т.д. Основываясь на том, каким образом мозг обрабатывает поступающие в него визуальные и аудио-сигналы, можно предположить, что применение подобных признаков позволит улучшить точность систем верификации диктора. В данной работе для извлечения признаков более высокого уровня была использована сверточная глубокая сеть доверия (Convolutional Deep Belief Network, СГСД).

Основным отличием сверточной глубокой сети доверия [122, 123] от обычной глубокой сети доверия [124] является применение в качестве слоев сети сверточной ограниченной машины Больцмана (CRBM - Convolutional Restricted Boltzmann Machine, СОМБ) [122, 125]. Данная нейронная сеть (Рисунок 2.3.1) представляет собой детектор признаков, состоящий из трех слоев – видимого слоя *V*, слоя детекции *H* и слоя свертки *P*.

Применение дополнительного слоя свертки позволяет уменьшить детализацию подаваемых на следующий скрытый слой данных, что позволяет выделять более крупные особенности в признаках. Это также позволяет

уменьшить вычислительную нагрузку на последующих слоях и отфильтровывать случайные шумы.

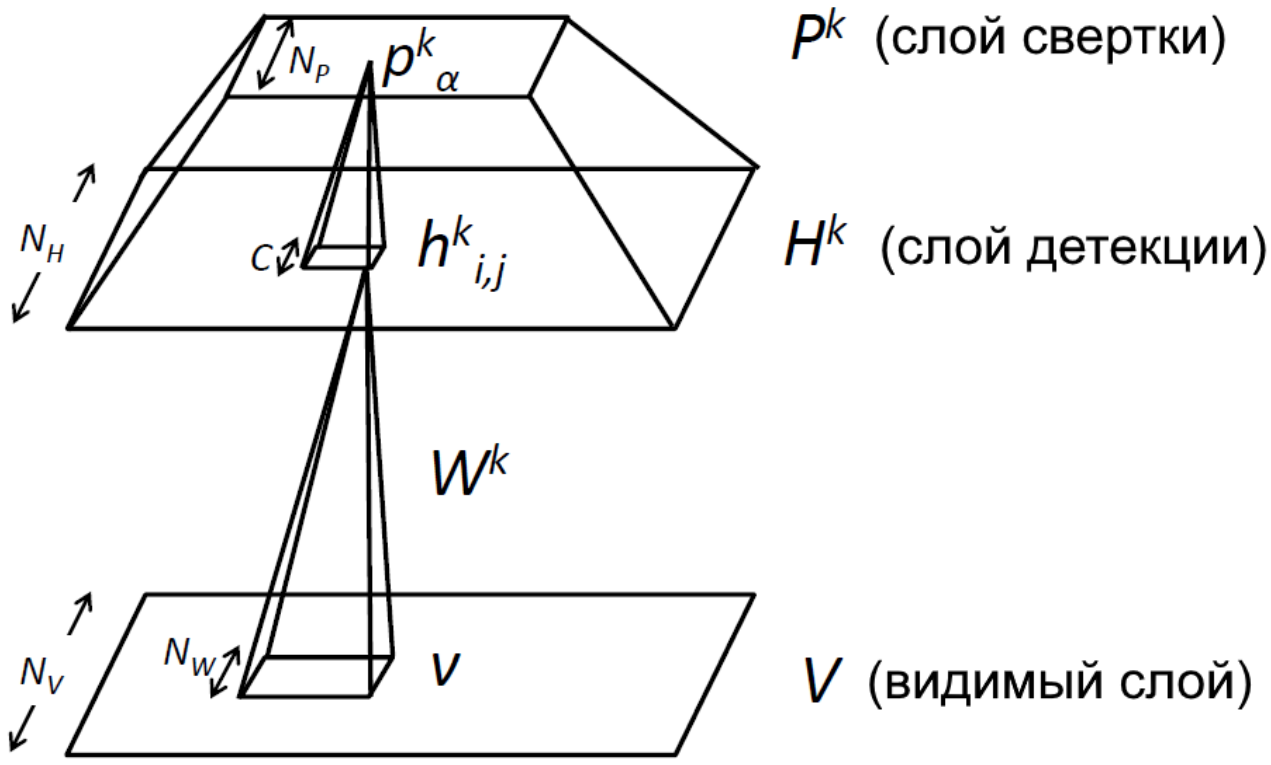

Рисунок 2.3.1 – Сверточная ограниченная машина Больцмана

Представим, что входной слой сети состоит из матрицы вещественных нейронов размерности $N_V \times Ch$, где N_V – количество окон, на которые разбивается аудиосигнал, Ch – количество каналов спектра. Для создания скрытого слоя используются K фильтров размерности $N_W \times Ch$ с весами W^K, которые также называют базами. Скрытый слой состоит из K групп матриц размерности $N_H \times Ch$ с нейронами из группы k, разделяющими общие веса W^k ($N_H = N_V$-N_W+1). Также для каждой группы фильтров задается общее смещение b_k и общее смещение для видимого слоя c. Тогда функцию энергии СОМБ (2.3.1) можно задать как [123]:

$$\mathrm{E}(\mathrm{v},\mathrm{h}) = \frac{1}{2}\sum_{i=1}^{N_V} v_i^2 - \sum_{k=1}^{K}\sum_{j=1}^{N_H}\sum_{r=1}^{N_W} h_j^k W_r^k v_{j+r-1} - \sum_{k=1}^{K} b_k \sum_{j=1}^{N_H} h_j^k - c\sum_{i=1}^{N_V} v_i \qquad (2.3.1)$$

Зададим совместные и условные функции распределения вероятностей для данной сети (2.3.2-2.3.4):

$$\mathrm{P}(\mathrm{v},\mathrm{h}) = \frac{1}{Z}\exp(-E(v,h)) \qquad (2.3.2)$$

$$\mathrm{P}\left(h_j^k = 1|v\right) = sigmoid(\ (\widetilde{W}_j^k *_v v)_j + b_k) \qquad (2.3.3)$$

$$\mathrm{P}(v_i|h) = Normal(\textstyle\sum_k(\widetilde{W^k} *_f h^k)_i + c, 1), \qquad (2.3.4)$$

где $*v$ – "действительная" свертка, $*f$ – "полная" свертка, $\widetilde{W}_j^k \triangleq W_{n_W-j+1}^k$ [123]. Для заданного вектора размерности m и ядра размерности n, где $m>n$, валидная свертка дает $(m\text{-}n\text{+}1)$-мерный вектор, полная свертка дает $(m\text{+}n\text{-}1)$-мерный вектор. Так как все нейроны скрытого слоя условно независимы от другого слоя, вывод в сети можно осуществить используя семплирование по Гиббсу.

Сверточная глубокая сеть доверия представляет собой композицию простых сверточных ограниченных машин Больцмана, благодаря чему скрытый слой каждой подсети служит видимым слоем для следующей. Из-за этого можно произвести быструю послойную процедуру обучения без учителя, в которой для оценки градиента относительное расхождение [126] применяется к каждой подсети по очереди, начиная с первой пары слоев. На видимый слой сети подаются данные из обучающего набора, последующие скрытые слои принимают на вход данные с выхода предыдущих.

Во время обучения на тренировочных данных СГСД может обучаться вероятностно отстраивать свои входы. После того, как будет произведено обучение основной структуры сети, ее можно дообучить для классификации объектов.

Общий алгоритм обучения СГСД можно задать так:

1. Представить два нижних слоя (входной и первый скрытый) как СОМБ. Произвести обучение СОМБ входных данных из видимого слоя V и получить матрицу ее весовых коэффициентов W, которая будет описывать связи между двумя нижними слоями сети.
2. Произвести вычисления, пропустив через уже обученную СОМБ входные данные V и получить данные скрытого слоя H' на выходе после активации узлов первого скрытого слоя.
3. Повторять шаги 1 и 2, используя в качестве входных данных H' для всех последующих пар слоев до тех пор, пока не будут обучены самые верхние слои.

Экспериментальная оценка. Для проведения экспериментов необходимо задать структуру и параметры СГСД, а после этого обучить данную сеть. Была использована следующая структура сверточной глубокой сети доверия: сеть состояла 3-х слоев, первый и второй слой состоят из 300 баз, третий из 60. Входной слой состоит из 80 нейронов (Ch = 80), на вход подаются данные спектрограммы аудиозаписи, полученные с уменьшением размерности с помощью метода главных компонент [127]. Данные, подаваемые на видимый слой, выбираются окнами по 20 мс со смещением 10 мс. Для каждой базы в скрытых слоях была использована размерность фильтра N_W = 6, коэффициент свертки равен 3. Параметры для первого и второго слоя сети были взяты из [123]. Параметры для третьего слоя выбраны автором самостоятельно.

В результате обучения СГСД были получены три обученных слоя сети, выходы каждого из которых можно использовать в качестве признаков для проведения верификации диктора. Рассмотрим полученные выходы сети на примере фразы “Со спокойным мужеством Скайлс ожидал всего в этом безумном городе”, произнесенной мужчиной и женщиной (Рисунки 2.3.2-2.3.9). Исходя из визуального представления функций активации нейронов, видны

сильные отличия – для диктора-женщины активируется гораздо больше нейронов.

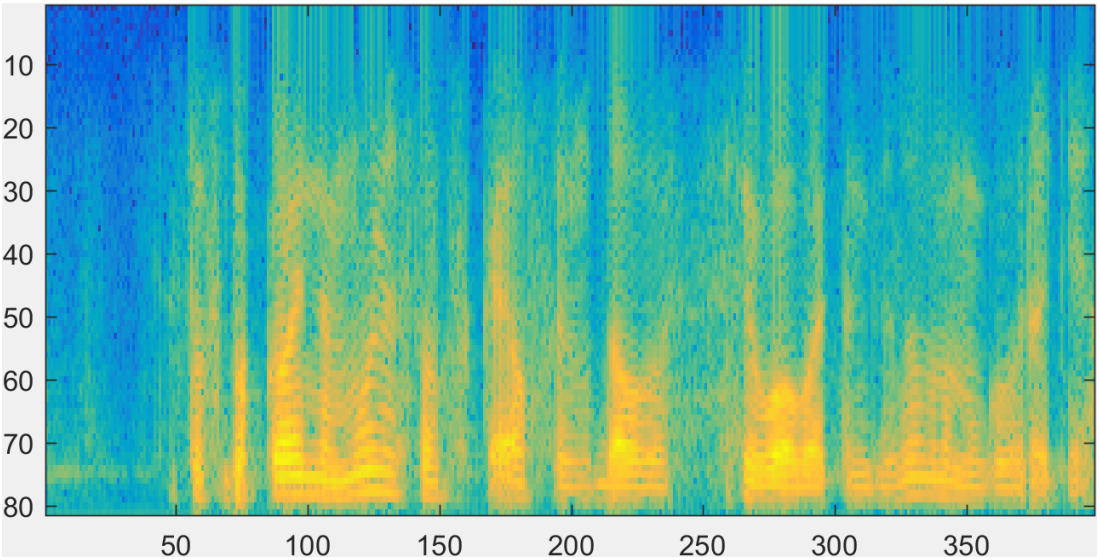

Рисунок 2.3.2 – Спектрограмма фразы диктора-мужчины

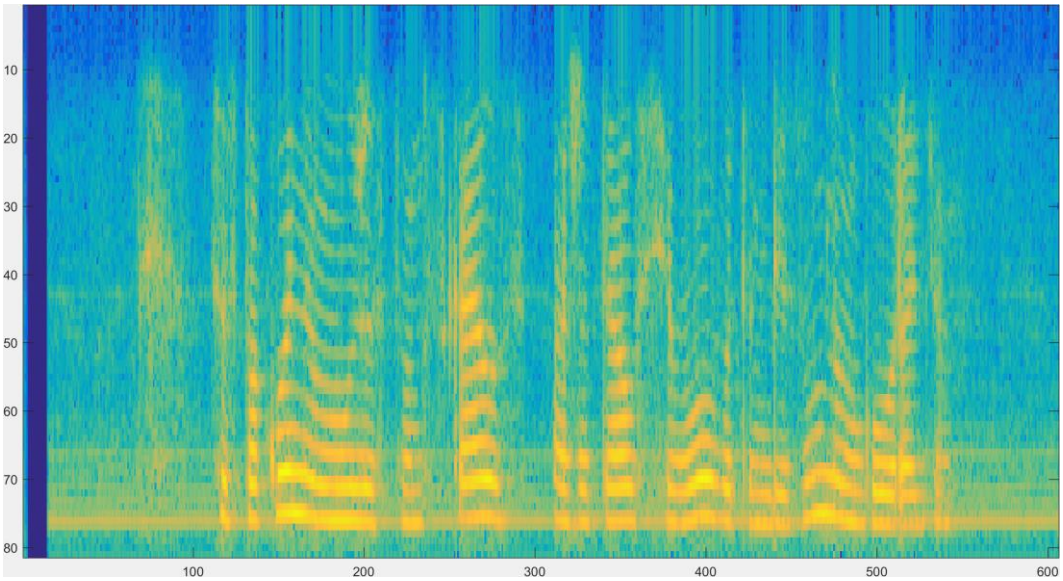

Рисунок 2.3.3 – Спектрограмма фразы диктора-женщины

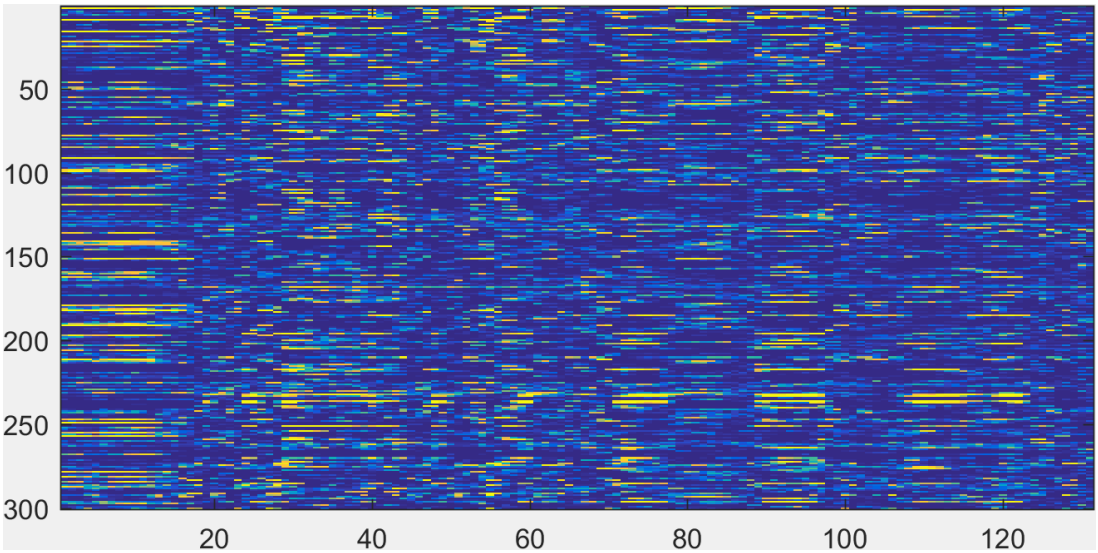

Рисунок 2.3.4 – Выходные значения нейронов первого скрытого слоя обученной СГСД для фразы диктора-мужчины

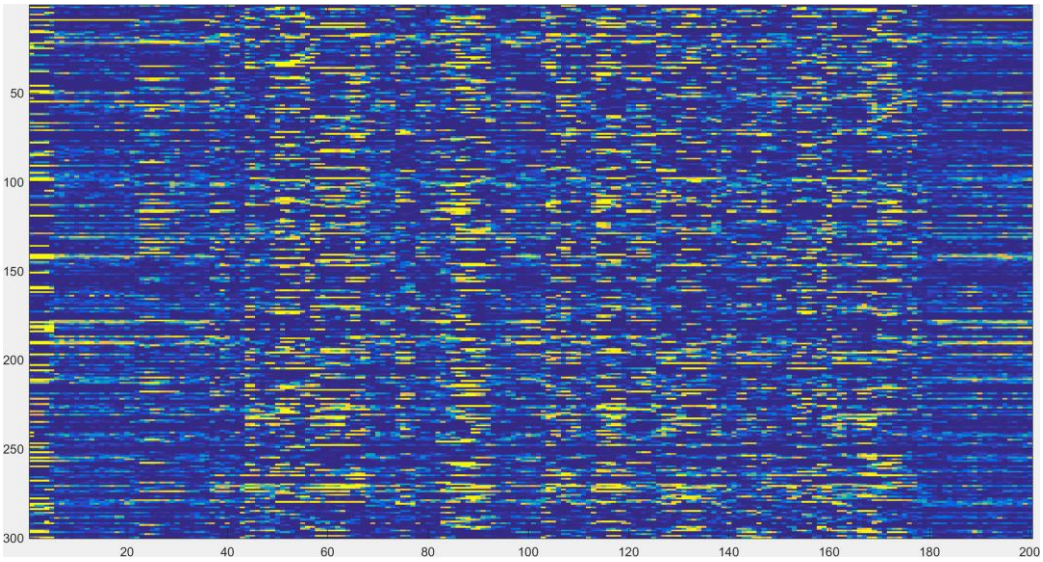

Рисунок 2.3.5 – Выходные значения нейронов первого скрытого слоя обученной СГСД для фразы диктора-женщины

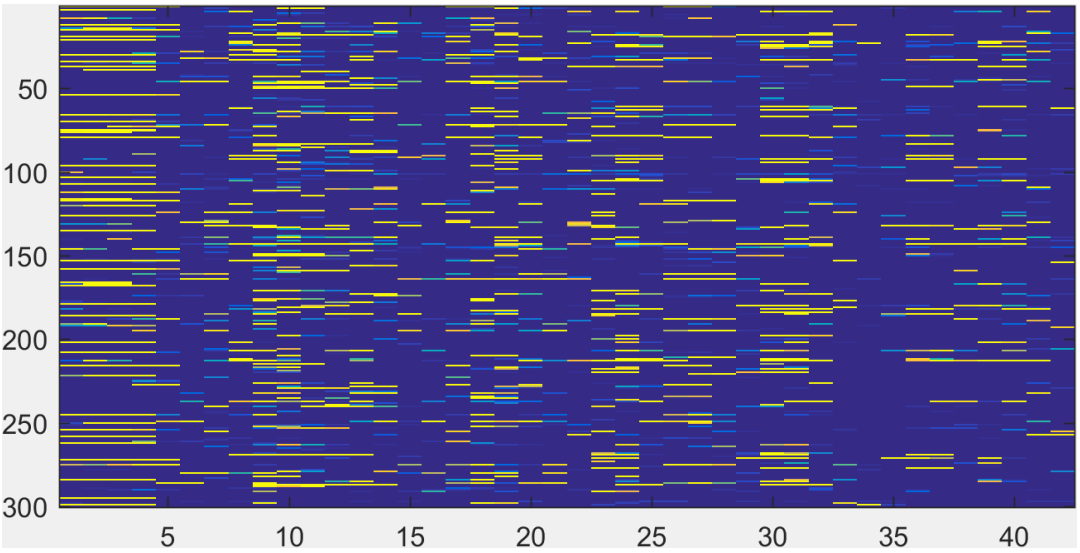

Рисунок 2.3.6 – Выходные значения нейронов второго скрытого слоя обученной СГСД для фразы диктора-мужчины

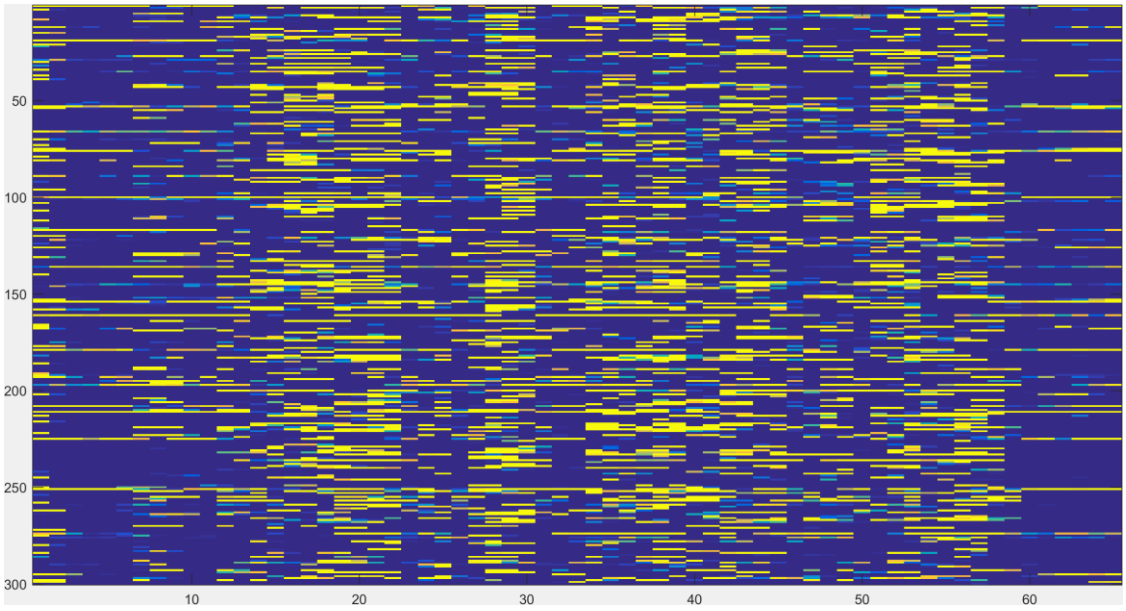

Рисунок 2.3.7 – Выходные значения нейронов второго скрытого слоя обученной СГСД для фразы диктора-женщины

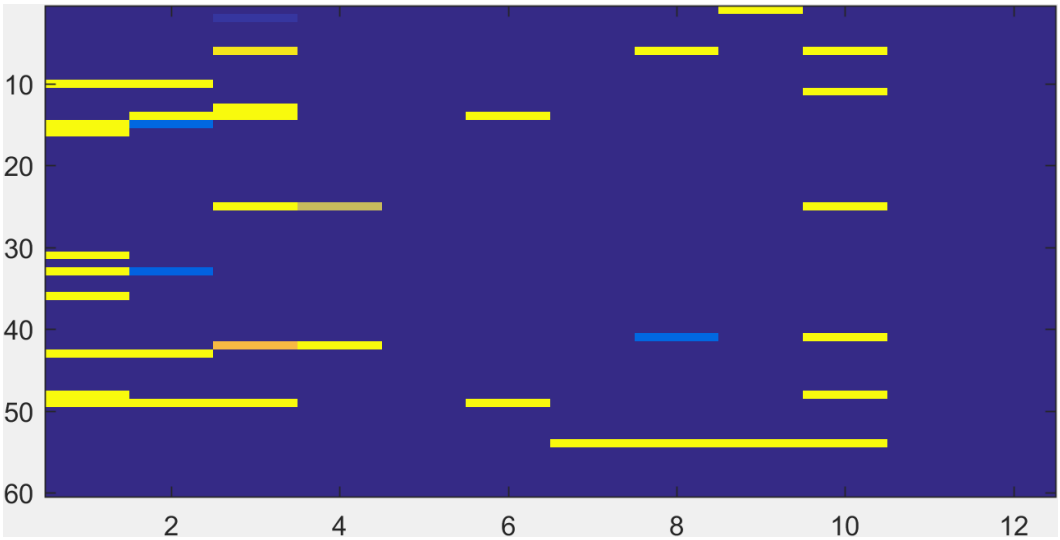

Рисунок 2.3.8 – Выходные значения нейронов третьего скрытого слоя обученной СГСД для фразы диктора-мужчины

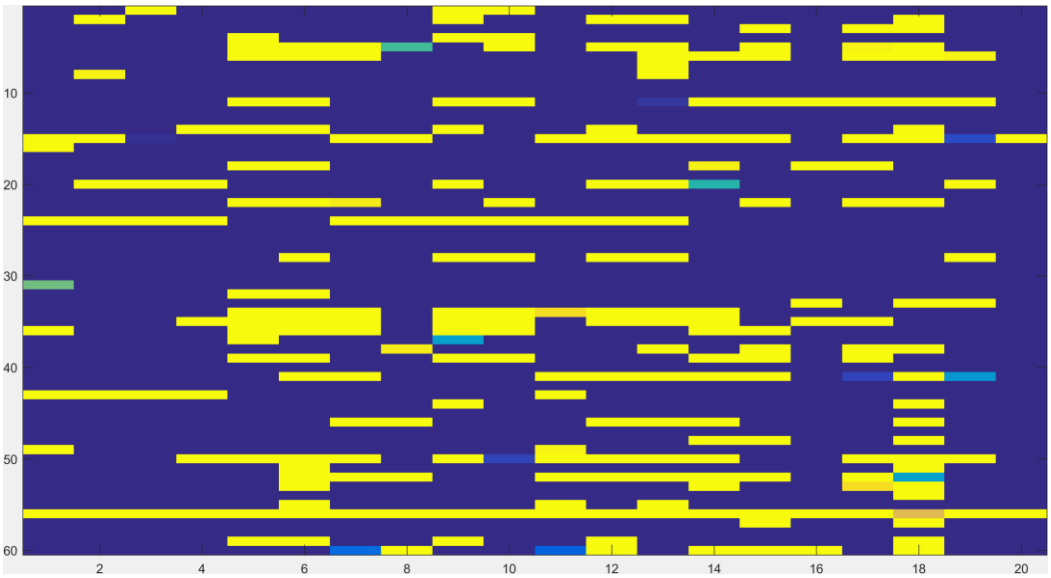

Рисунок 2.3.9 – Выходные значения нейронов третьего скрытого слоя обученной СГСД для фразы диктора-женщины

На Рисунке 2.3.10 изображены паттерны трех случайным образом взятых фильтров первого скрытого слоя обученной СГСД. Можно увидеть, что после обучения данные фильтры пытаются найти определенные спектральные шаблоны, соответствующие различным голосам или фонемам. На более высоких уровнях на основе фильтров первого уровня строится

высокоуровневое представление речи дикторов. Следовательно, вполне возможным является применение данных признаков для распознавания речи или идентификации пола говорящего.

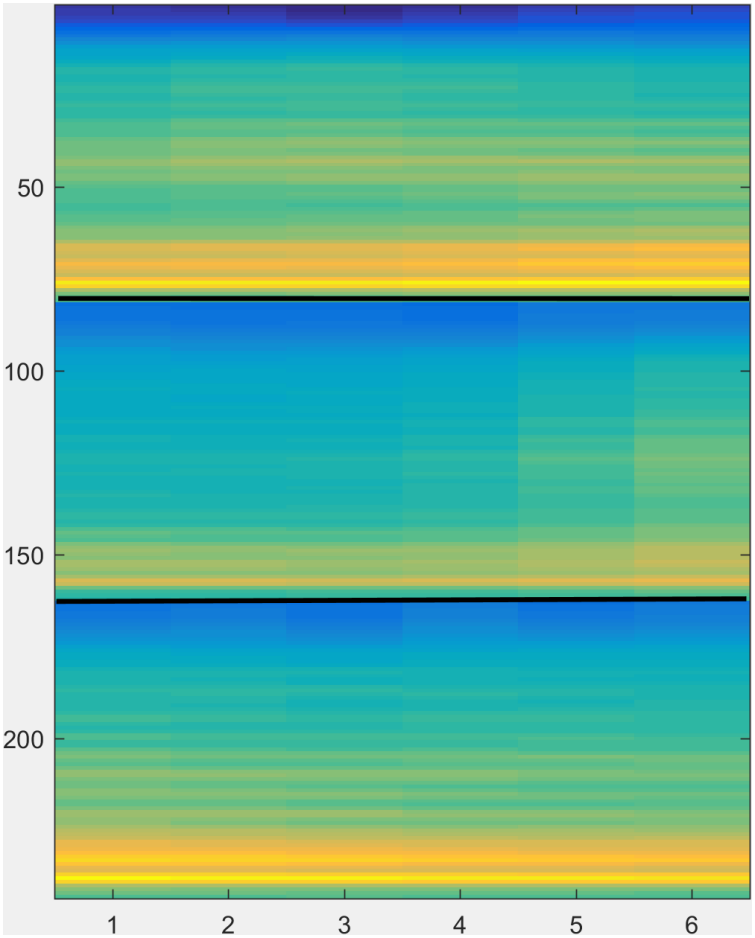

Рисунок 2.3.10 – Паттерны трех случайным образом взятых фильтров первого скрытого слоя обученной СГСД

Для оценки системы верификации с полученными признаками, подадим их на вход модели Гауссовой смеси, которую будем использовать в качестве классификатора. Для тестирования точности системы верификации были использованы те же параметры, что и в разделе 2.1.

Рассмотрим результаты тестирования системы верификации (Таблица 2.3.1). По полученным результатам можно сделать вывод, что ни один из полученных наборов признаков не дает точности верификации системы больше, чем стандартный набор признаков. Применение комбинации из

нескольких классификаторов, использующихся для верификации диктора, также не дало увеличения точности.

Следует помнить о том, что методы глубокого обучения работают лучше на большом объеме обучаемых данных, поэтому малый объем обучающей выборки (30 дикторов в данном случае) – одна из возможных причин малой точности системы. Однако другой причиной может быть использование классификатора, который не дает возможности показать лучшие результаты. Данный вывод будет проверен в следующем разделе.

Таблица 2.3.1 – Результаты тестирования признаков СГСД

Набор признаков	% EER	minDCF*100
СГСД1	2,000	0,997
СГСД2	3,500	1,740
СГСД3	10,000	5,765
СГСД1+СГСД2	2,000	1,197
СГСД1+СГСД3	2,000	1,121
СГСД2+СГСД3	3,289	1,926
СГСД1+СГСД2+СГСД3	2,000	1,327
MFCC (базовый)	1,000	0,925
ЖА Add-del(MFCC+V_p)	**0,579**	**0,623**

2.4 Гибридный алгоритм верификации диктора по произвольной фразе на основе ансамбля классификаторов

Одним из способов решения проблемы переобучения модели при решении задачи классификации является создание ансамбля классификаторов. Несомненно, данный метод повышает сложность модели, но когда решающим фактором является точность классификатора, а в случае применения систем верификации диктора по голосу она важна, сложность отходит на второй план.

Для построения ансамбля классификаторов были использованы следующие классификаторы: Гауссова смесь, с применением полученного

алгоритма верификации, описанного в разделе 2.2, и классификаторы, обученные на признаках из сверточной глубокой сети доверия, представленных в разделе 2.3. В качестве классификаторов были взяты стандартные классификаторы, реализованные в пакете Matlab: машина опорных векторов (Support Vector Machine, SVM), алгоритм ансамблевой классификации AdaBoost M1 и классификатор, основанный на линейном дискриминантном анализе (Linear Discriminant Analysis, LDA). В рамках работы также использовались наивный Байесовский классификатор и метод бэггинга деревьев принятия решений (TreeBag), однако результаты работы данных классификаторов оказались неудовлетворительными и в данной работе они не описаны.

Машина опорных векторов [128] (метод опорных векторов) является одним из популярных методов, используемых в задачах машинного обучения и задаче верификации диктора по голосу. Машина опорных векторов является двухклассовым классификатором, построенным на суммах функции ядра $K(\cdot,\cdot)$ (2.4.1), где t_i идеальные выходы (1 или -1, в зависимости от принадлежности опорного вектора классу 1 или -1), $\sum_{i=1}^{L}\alpha_i t_i = 0$, $\alpha_i > 0$. Вектора x_i – опорные вектора.

$$f(x) = \sum_{i=1}^{L}\alpha_i t_i K(x, x_i) + d\ , \qquad (2.4.1)$$

Этот метод позволяет построить гиперплоскость в многомерном пространстве, разделяющую два класса, например, признаки целевого диктора и признаки дикторов из референтной базы. Гиперплоскость вычисляется с использованием не всех векторов признаков, а только специально выбранных. Эти вектора и называются опорными [6].

Выбор ядра является одним из ключевых моментов при использовании машины опорных векторов. Для задачи верификации диктора по голосу

используются как обычные ядра – линейное ядро, радиально-базисное ядро, так и специальные ядра, например ядро Фишера [82]. В данной работе было использовано обычное линейное ядро.

Линейный дискриминантый анализ (ЛДА) [129] – метод, используемый в статистике, распознавании образов и машинном обучении для поиска линейной комбинации признаков, позволяющих разделить два и более класса объектов. Данный метод можно применять для решения задачи линейной классификации и уменьшения размерности признаков. В данной работе ЛДА использовался для решения задачи двухклассовой классификации.

Пусть есть набор наблюдений x – обучающий набор данных, для каждого из которых известен класс y. В ЛДА предполагается, что функции совместной плотности распределения вероятностей $p(x|y=1)$ и $p(x|y=0)$ – имеют нормальное распределение с параметрами $(\mu_0,\sum_0)$ и $(\mu_1,\sum_1)$, ковариационные матрицы равны $\sum_0 = \sum_1 = \sum$. Тогда задача классификации сводится к сравнению скалярного произведения с порогом c (2.4.2, 2.4.3):

$$w \cdot x > c \tag{2.4.2}$$

$$w = \Sigma^{-1}(\mu_1 - \mu_0) \tag{2.4.3}$$

Алгоритм AdaBoost [130] использует объединение взвешенных выходов так называемых “слабых” классификаторов (простых классификаторов, для которых вероятность правильной классификации немногим более 50 %), для получения более точного решения задачи классификации. При этом часть “слабых” классификаторов может корректировать неверные решения, выданные другими “слабыми” классификаторами. Данный метод меньше переобучается по сравнению с другими методами классификации, однако на точность метода сильно влияют выбросы и шумы в данных. В данной работе в качестве “слабых” классификаторов использовалось 50 деревьев принятия решений. Подробнее с алгоритмом AdaBoost можно ознакомиться в [131-133].

Экспериментальная оценка. При проведении экспериментов была оценена точность системы верификации, использующей рассмотренные классификаторы по отдельности. Для этого каждый из классификаторов был обучен, используя в качестве признаков выходы одного из трех слоев СГСД, цифра в названии классификатора означает номер скрытого слоя, генерирующего признаки (Таблица 2.4.1). Для обучения данных классификаторов использовался подход “один против всех”. Векторам признаков, принадлежащих диктору, для которого обучается модель, присваивался класс “1”, всем остальным – “0”. При этом данные, используемые для обучения УФМ, не были включены в обучающую выборку.

Таблица 2.4.1 – Точность отдельных классификаторов, использующих признаки СГСД

Тип классификатора	% EER	minDCF*100
SVM1	6,500	1,456
SVM2	5,157	1,204
SVM3	11,500	5,167
ADABOOST1	2,289	1,457
ADABOOST2	2,000	0,964
ADABOOST3	4,789	3,043
LDA1	2,000	1,421
LDA2	1,973	1,462
LDA3	7,078	4,856

Для создания ансамбля классификаторов использованы классификаторы для всех трех типов признаков СГСД. Для итоговой оценки тестовой аудиозаписи была использована взвешенная сумма нескольких классификаторов. Для определения весов был использован генетический алгоритм. Итоговый ансамбль, показавший наибольшую точность верификации EER = 0,21 % (Рисунок 2.4.1), состоит из следующих классификаторов: Гауссовой смеси, обученной на векторе признаков, который был получен с

помощью жадного алгоритма Add-del; классификаторов LDA и AdaBoost M1, обученных на признаках с первого скрытого слоя СГСД; машины опорных векторов SVM, обученной на признаках из третьего скрытого слоя СГСД.

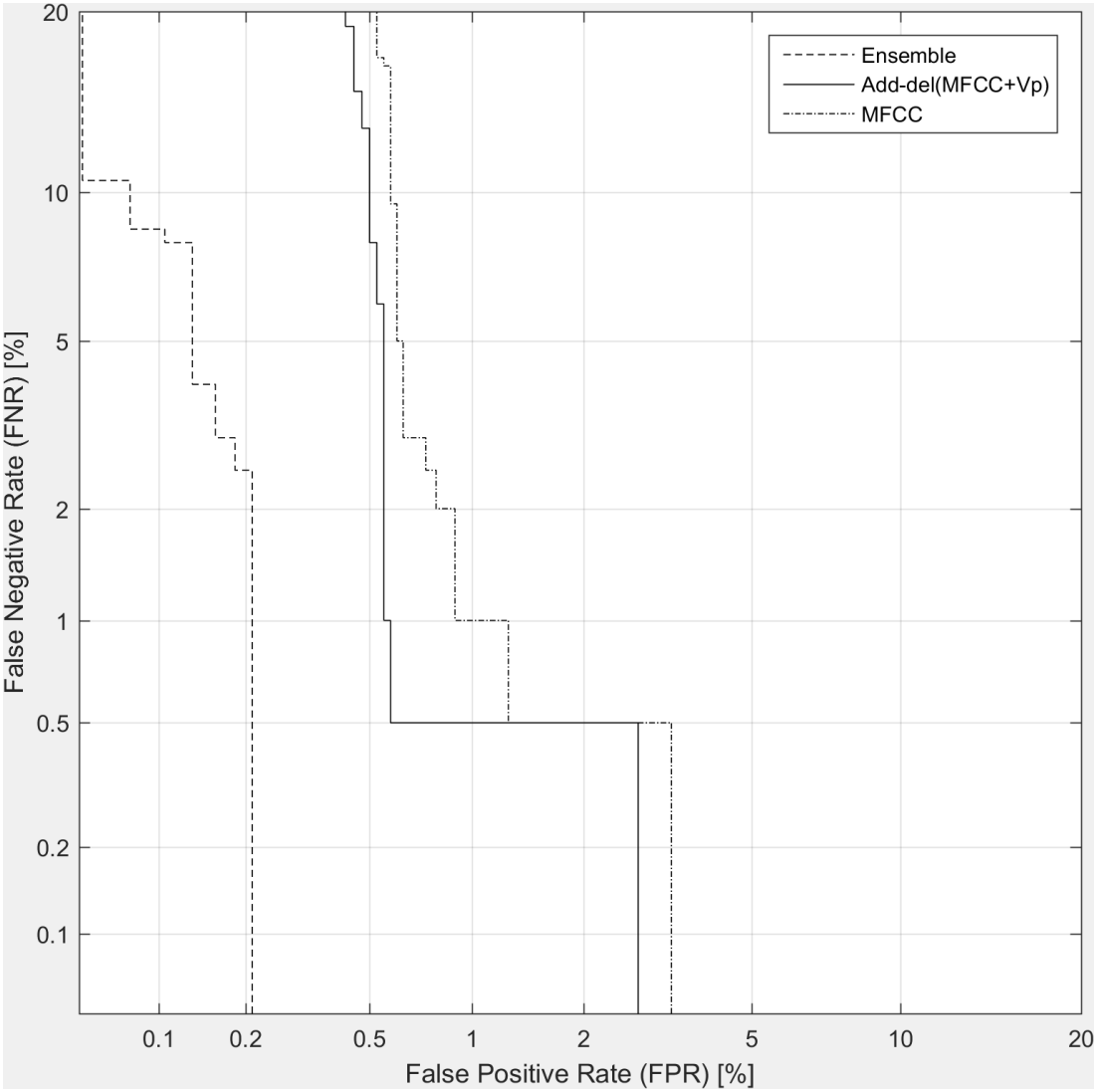

Рисунок 2.4.1 – Кривые компромиссного определения ошибки (DET кривые) для полученного ансамбля классификаторов, набора признаков, полученного с помощью жадного алгоритма Add-del и МКК

Полученные результаты показывают, что применение представленного ансамбля классификаторов позволяет решить задачу повышения точности системы верификации диктора по произвольной фразе. Блок-схема гибридного алгоритма верификации на основе ансамбля классификаторов представлена на Рисунке 2.4.2.

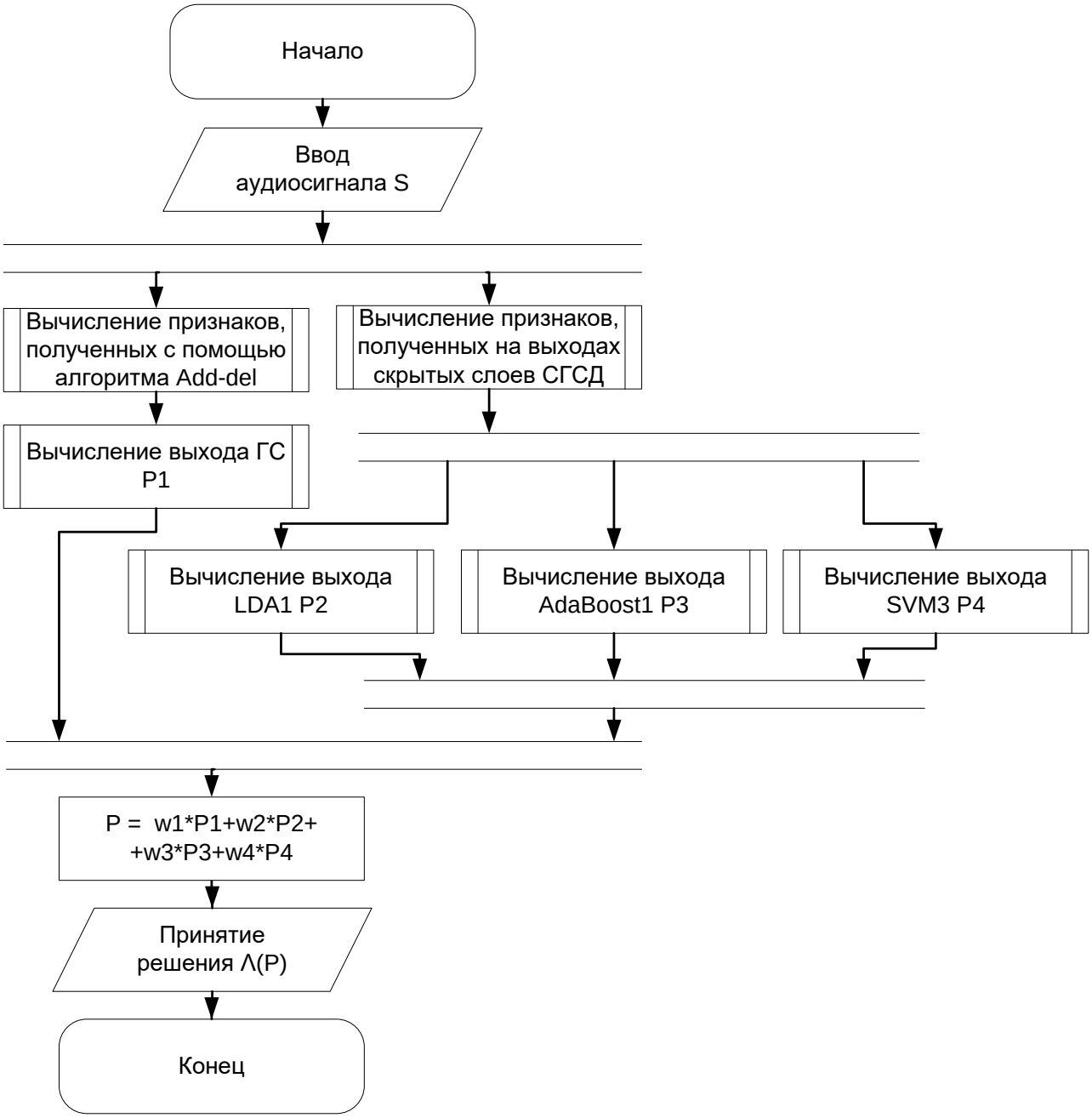

Рисунок 2.4.2 – Блок-схема гибридного алгоритма верификации диктора по произвольной фразе на основе ансамбля классификаторов

Наибольший вклад в финальную оценку тестовой аудиозаписи дают в порядке убывания: классификатор на ГС, классификатор LDA, классификатор SVM, классификатор AdaBoost M1. Следует отметить, что в повышении точности также участвует третий слой СГСД, который выделяет признаки более высокого уровня. Если убрать данный классификатор, произойдет

увеличение ошибки EER до 0,5 % (Рисунок 2.4.3). Отсюда следует, что СГСД позволяет выделять признаки, повышающие точность системы верификации диктора по произвольной фразе (Таблица 2.4.2).

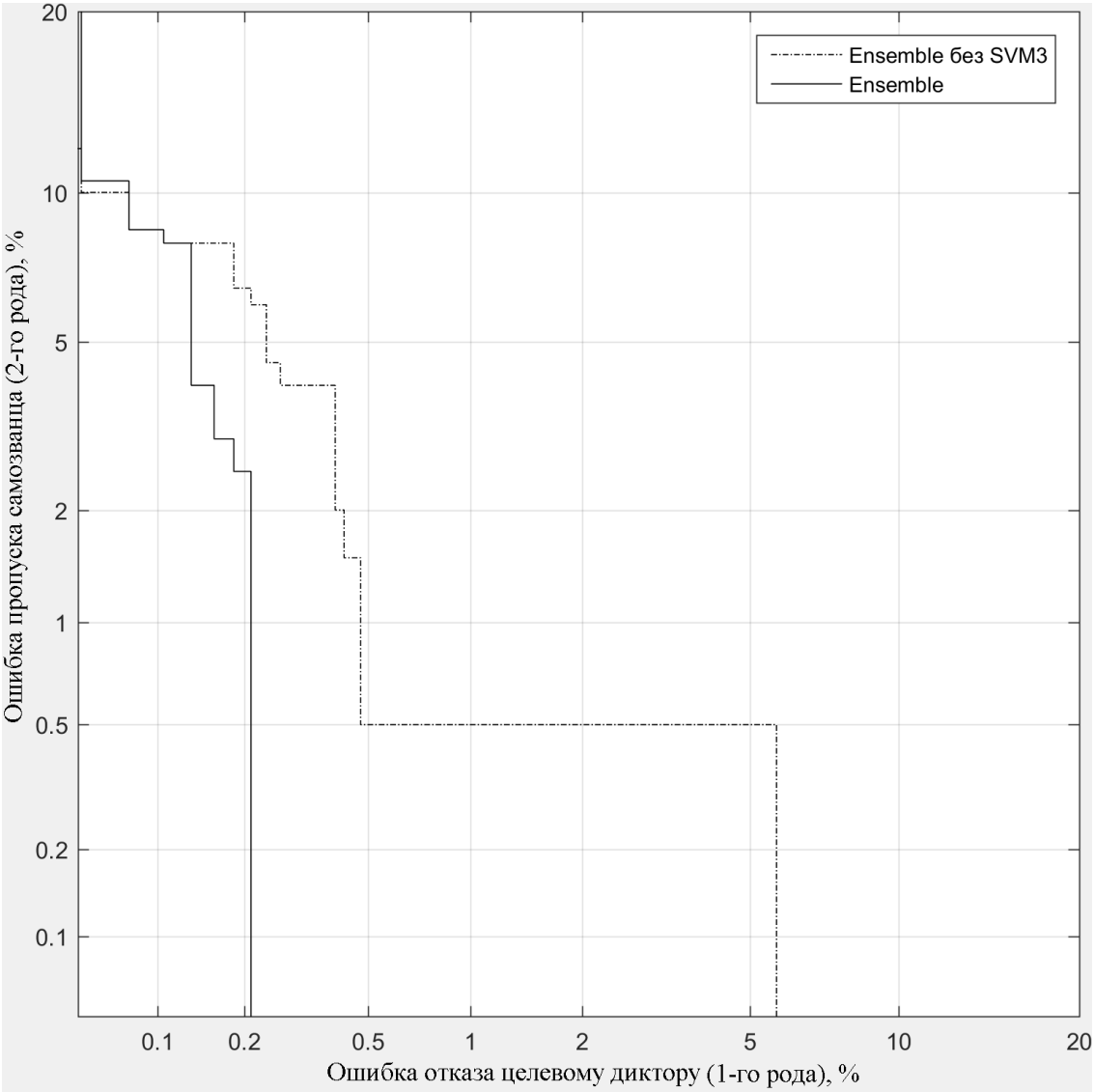

Рисунок 2.4.3 – Кривые компромиссного определения ошибки (DET кривые) для полученного ансамбля классификаторов, и ансамбля, в котором был исключен классификатор SVM3

Таблица 2.4.2 – Сравнение точности ансамблей классификаторов и Гауссовой смеси, использующей признаки полученные жадным алгоритмом Add-del и мел-кепстральные коэффициенты

Тип классификатора	% EER	minDCF*100
Ensemble	**0,210**	**0,208**
Ensemble(без SVM3)	0,500	0,519
ГС, Add-del(MFCC+Vp)	0,579	0,623
ГС, MFCC	1,000	0,925

2.5 Выводы

1. Разработан оригинальный алгоритм верификации диктора, отличающийся от существующих применением речевых признаков, полученных с помощью жадного алгоритма отбора признаков. На речевом корпусе, состоящем из аудиозаписей 50 дикторов, было получено уменьшение ошибки EER на 42,1 %, по сравнению с базовым алгоритмом, в котором применяются мел-кепстральные коэффициенты. На речевом корпусе CHAINS была достигнута 100 % точность верификации.

2. Разработан алгоритм генерации признаков, основанный на применении сверточной глубокой сети доверия. Данный алгоритм отличается от существующих расширенной архитектурой нейронной сети, выделяющей более высокоуровневые признаки и уменьшающей их количество. Применение признаков, полученных с помощью созданного алгоритма, возможно для повышения точности систем верификации диктора, распознавания речи или идентификации пола говорящего.

3. Разработан гибридный алгоритм верификации диктора по произвольной фразе на основе ансамбля классификаторов. Отличительной особенностью алгоритма является применение в ансамбле классификаторов, использующих признаки, извлеченные из аудиозаписей с помощью сверточной глубокой нейронной сети доверия. На рассмотренном речевом корпусе точность верификации была повышена на 79 %, по сравнению с базовым алгоритмом.

3. Программное средство для верификации диктора по произвольной фразе

3.1 Состав программного средства

На основе представленных алгоритмов было создано программное средство. Данное программное средство предназначено для проведения автоматической верификации или идентификации диктора, при этом включает в себя все необходимые модули для извлечения речевых признаков, обучения моделей дикторов и УФМ, а также проведения верификационных испытаний. Система, описанная в работе, была создана с применением библиотеки MSR Identity Toolbox [107].

Рассмотрим структуру программного средства (Рисунок 3.1.1).

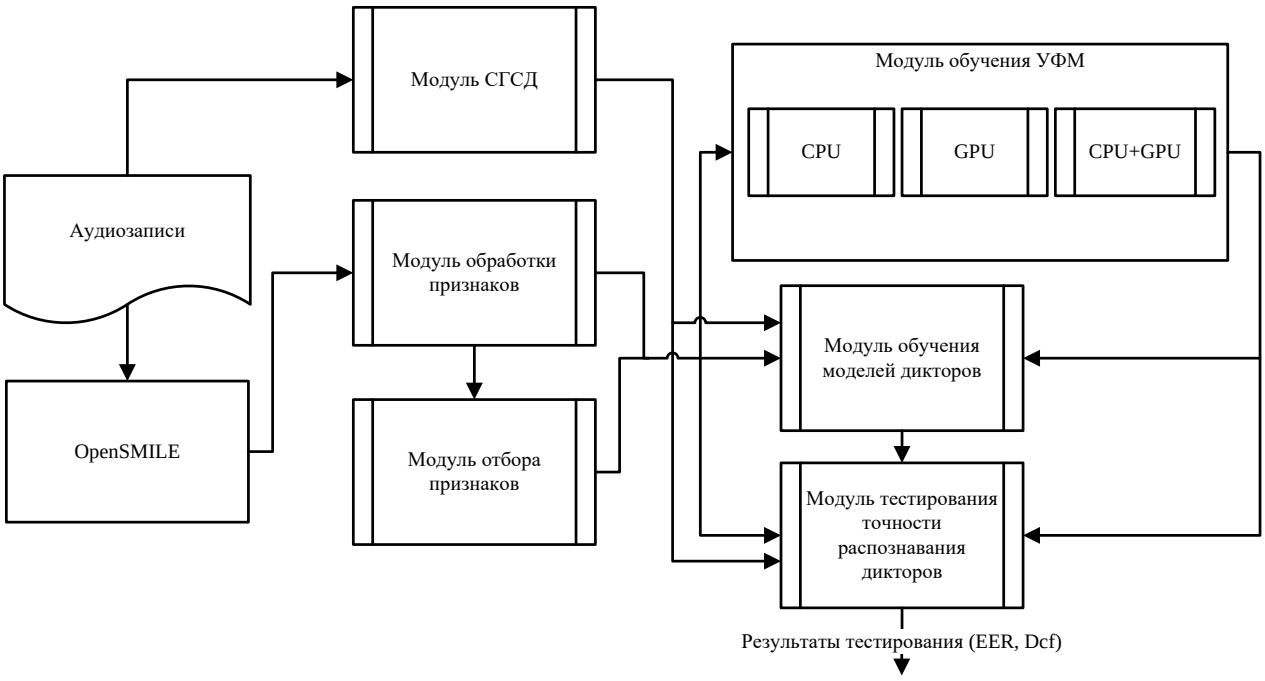

Рисунок 3.1.1 – Структура программного средства

Для извлечения речевых признаков из аудиозаписей голоса диктора был использована библиотека openSMILE [106] и разработанный модуль сверточной глубокой сети доверия. С помощью openSMILE из аудиозаписей

извлекаются такие признаки, как мел-кепстральные коэффициенты, пары линейного спектра, кепстральные коэффициенты перцептивного линейного предсказания, энергия, формантные частоты, частота основного тона, вероятность вокализации, частота пересечения нуля, джиттер и шиммер. Полный вектор признаков, вычисляемый для одного окна длиной в 20 мс, состоит из 94 признаков. С помощью СГСД извлекаются признаки следующей размерности: из первого и второго слоя извлекается вектор по 300 признаков, из третьего слоя – по 60.

Для проведения экспериментов реализовано несколько модулей – модуль обработки признаков, который позволяет отобрать необходимые в исследовании речевые признаки; модуль СГСД, позволяющий обучать сеть и выделять с помощью нее признаки; модуль обучения УФМ; модуль обучения моделей дикторов и модуль тестирования точности распознавания дикторов. Модуль обучения УФМ был реализован в нескольких вариациях – с вычислениями на центральном процессоре, с вычислениями на процессоре видеокарты и комбинированный вариант.

Для проведения серии экспериментов по отбору речевых признаков разработаны модули, реализующие алгоритм жадного добавления-удаления и генетический алгоритм. Отбор признаков позволяет снизить переобучение модели и сохранить при этом наиболее информативные признаки.

Одной из ключевых особенностей реализованного программного средства является предложенный алгоритм комбинированных вычислений на центральном процессоре и процессоре видеокарты [134]. При этом в аналогах производятся вычисления либо только на центральном процессоре [107, 135], либо только на процессоре видеокарты [136].

Для существующих систем верификации диктора используются базы речевых данных в несколько сотен часов. При этом обучение УФМ может длиться не одну неделю на современном центральном процессоре, а существенное увеличение размера базы становится практически невозможным

[136]. Для ускорения процесса обучения УФМ можно использовать параллельные алгоритмы, в том числе с применением вычислений на графическом процессоре видеокарты.

Наиболее трудоемким по количеству затрачиваемого времени и вычислений является обучение УФМ, однако обучение данной модели хорошо распараллеливается. Это возможно благодаря разделению последовательности входных обучающих векторов на отдельные блоки, каждый из которых вычисляется отдельно (1.4), (1.6), а затем суммируется. При этом полученная сумма не изменится, как и в случае если входные данные на блоки не разбивались бы (2.1.1-2.1.4). То есть, возможно выполнение расчетов отдельных частей в разных потоках на разных данных, соответственно каждый из этих потоков независим и работает со своими данными.

Для выполнения одновременных вычислений на центральном (ЦП) и графическом (ГП) процессорах часть блоков данных перемещается в память видеокарты, а затем в отдельном потоке запускаются необходимые вычисления. Для хранения в памяти параметров модели Σ, μ, и w требуется 8*C*D байт, в данном случае количество компонент смеси C = 256, количество используемых признаков D = 28, итого 57344 байт. Для вычислений используются блоки данных размером 50000 векторов по 8*D байт, итого 11,2×106 байт. Для промежуточных вычислений необходимы блоки 2*8*C*D байт, 8*50000*C байт, итого ≈ 102,5×106 байт.

Результаты испытаний программного средства. Проведены эксперименты с применением речевого корпуса, включающего записи речи 25 дикторов-мужчин и 25 женщин. Данный речевой корпус содержит записи произнесенных без предварительной подготовки предложений, взятых из художественной литературы, или поговорок. Суммарная длина записей речи для каждого диктора составляет не менее 6 мин, включая 50 сегментов различной длины. Каждый диктор был записан на микрофон в условиях небольшого шума, частота дискретизации 8000 Гц, разрядность 16 бит.

Весь речевой корпус, состоящий из записей речи 50 дикторов, разделен на обучающую выборку для УФМ, состоящую из записей 30 дикторов, и выборку, использующуюся для обучения и тестирования моделей дикторов, состоящую из записей оставшихся 20 дикторов. Количество дикторов-мужчин и дикторов-женщин во всех выборках одинаково. Общий объем данных, используемых для обучения УФМ, составляет 162,28 Мб.

Эксперименты проводились с использованием 4-ядерного процессора Intel Core i7-3630QM, видеокарты nVidia GeForce GT 640M с 2 Гб DVRAM.

Для сравнения эффективности работы разработанного программного средства, и в том числе модуля обучения УФМ, было проведено несколько экспериментов по определению скорости работы (времени обучения УФМ) модулей. Результаты оценки времени обучения УФМ с разбиением на различные размеры блоков представлены в Таблице 3.1.1. При проведении оценки зафиксировано количество используемых потоков процессора, равное 4.

Таблица 3.1.1 - Время обучения УФМ в зависимости от размера блоков обучающих данных

Размер блока данных (семплов)	Время обучения на ЦП, с	Время обучения на ГП, с	Время обучения на ЦП и ГП, с
5000	168,7424	165,7454	135,4970
10000	165,5472	128,2411	107,9179
25000	162,0763	116,2236	104,0332
50000	159,0245	111,4273	100,2596

Наименьшее время обучения УФМ было получено при использовании параллельных вычислений на центральном процессоре и видеокарте, которое составляет 100,2596 с.

Было произведено сравнение скорости вычислений в зависимости от количества запущенных потоков процессора, зафиксировав размер блоков обучающих данных на 50000 векторов (Рисунок 3.1.2). Можно отметить, что независимо от количества потоков, комбинированные вычисления на процессоре и видеокарте выполняются быстрее, чем только на центральном процессоре. При использовании более 4 потоков и комбинированных вычислений на ЦП и ГП время вычислений далее не уменьшается, при вычислениях на ЦП – уменьшается, но незначительно.

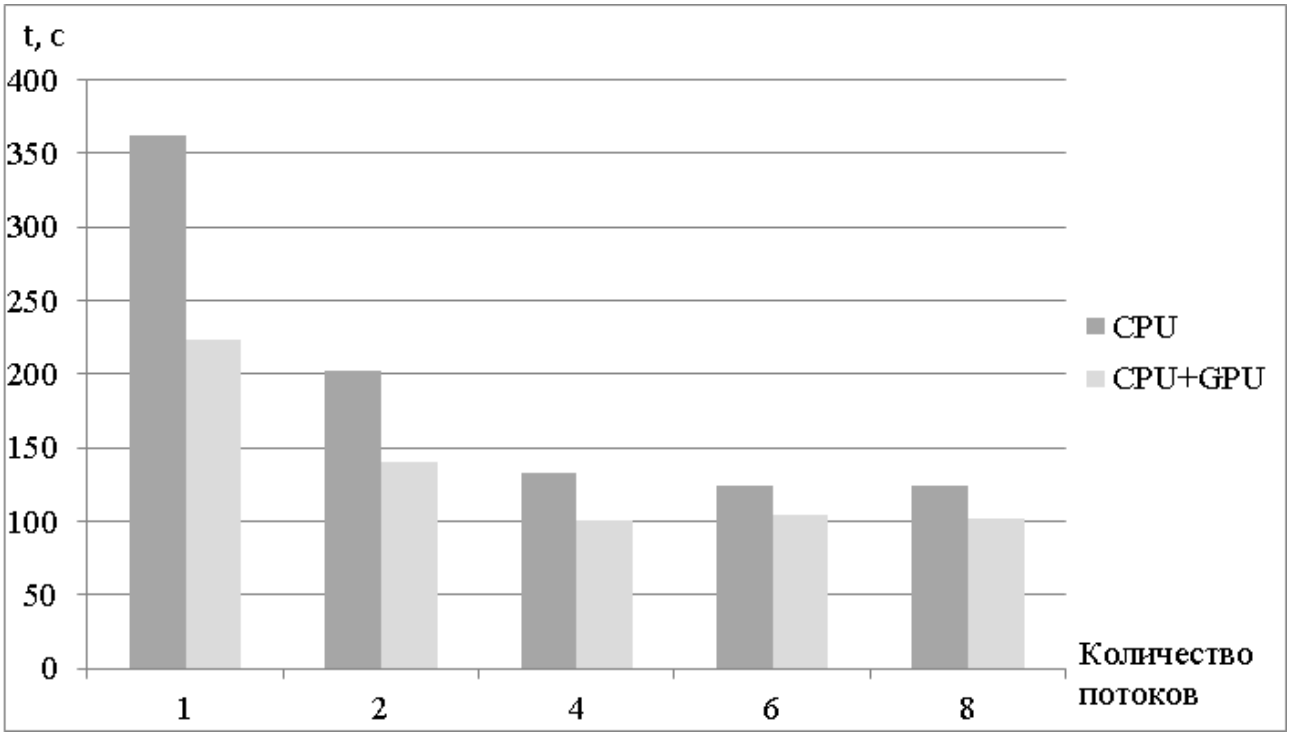

Рисунок 3.1.2 – Время обучения УФМ в зависимости от количества используемых потоков

Таким образом, можно сделать вывод, что реализованный модуль обучения УФМ с комбинированными вычислениями на центральном процессоре и процессоре видеокарты по сравнению с обучением УФМ на центральном процессоре позволяет уменьшить время работы на 36,95%, по сравнению с обучением на процессоре видеокарты позволяет уменьшить время работы на 10%.

3.2 Внедрение результатов диссертационной работы

Для оценки эффективности разработанных алгоритмов и программного средства, использующего рассмотренные алгоритмы верификации диктора по произвольной фразе, была произведена оценка в сравнении с аналогами. В качестве аналогов были использованы программные средства MSR Identity Toolbox [107] и ALIZE [137].

При внедрении был использован следующий подход: в течение трех месяцев программные средства использовались для верификации диктора по произвольной фразе с целью контроля доступа к рабочим компьютерам сотрудников предприятия. Количество зарегистрированных пользователей, участвовавших в проведении апробации – 34. Среди пользователей присутствовали лица возрастом от 21 до 48 лет. Для осуществления записи речи дикторов использовались отдельно подключаемые микрофоны бюджетного ценового сегмента.

Перед началом работы с системой производилась регистрация пользователя – запись речевого материала длиной около 60 секунд. Для этого ему предлагалось прочитать набор фраз, использовавшихся при оценке алгоритмов из главы 2. После этого в каждой из используемых систем было произведено создание моделей диктора. Необходимые для регистрации пользователей универсальная фоновая модель и сверточная глубокая сеть доверия были заранее обучены на основе рассмотренного в работе речевого корпуса.

Для проведения верификации пользователю необходимо произнести в микрофон выбранную случайным образом фразу. После этого данная фраза по очереди подавалась на вход системам для получения результата верификации. Таким путем проводилась верификация легального пользователя, для оценки отказов (ошибки 1-го рода). Для оценки ошибки 2-го рода производилась имитация нарушителя - запись подавалась на вход системам для верификации

моделей других пяти дикторов, выбранных случайным образом. Полученные при этом показатели были зафиксированы для дальнейшего анализа.

Для сравнения были использованы следующие критерии:

1. количество ошибок 1-го рода (отказов);
2. количество ошибок 2-го рода;
3. среднее время обучения модели диктора;
4. среднее время проведения верификации для одной фразы.

Значения критериев для сравниваемых программ, полученные в результате экспериментальной апробации, приведены в Таблице 3.2.1. Временные показатели взяты без учета времени необходимого на запись голоса.

Таблица 3.2.1 – Экспериментальные значения критериев для сравниваемых программ

Критерий сравнения	Разработанное программное средство	ALIZE	MSR Identity Toolbox
Количество ошибок 1-го рода (отказов)	12	18	26
Количество ошибок 2-го рода	42	74	131
Среднее время обучения модели диктора, с.	3,56	0,41	0,48
Среднее время проведения верификации для одной фразы, с.	0,17	0,08	0,09

Для того чтобы сравнить примененные программные средства на основе многокритериального подхода, был использован метод анализа иерархий [138, 139]. Этот метод основан на разбиении сложной задачи сравнения альтернатив на множество попарных сравнений, более простых.

В данном методе каждый критерий имеет свой приоритет, который задается весом. Для оценки весов используют шкалу (Таблица 3.2.2),

позволяющую оценить критерии по их значимости. Альтернативы же оцениваются по степени интенсивности проявления критерия.

Таблица 3.2.2 – Шкала относительной важности для оценочного сравнения критериев и альтернатив

Интенсивность относительной важности	Определение	Объяснение
0	Несравнимы	Сравнение невозможно
1	Равная важность	Равный вклад двух видов деятельности в цель
3	Умеренное превосходство одного над другим	Опыт и суждения дают легкое превосходство одному виду деятельности над другим
5	Существенное или сильное превосходство	Опыт и суждения дают сильное превосходство одному виду деятельности над другим
7	Значительное превосходство	Одному из видов деятельности дается настолько сильное превосходство, что оно становится практически значительным
9	Очень сильное превосходство	Очевидность превосходства одного вида деятельности над другим подтверждается наиболее сильно
2, 4, 6, 8	Промежуточные решения между двумя соседними суждениями	Применяются в компромиссном случае
Обратные величины приведенных выше чисел	Если при сравнении одного вида деятельности с другим получено одно из вышеуказанных чисел (например, 3), то при сравнении второго вида деятельности с первым получим обратную величину (т.е. 1/3)	

Самым важным критерием для систем верификации является ошибка второго рода, так как последствия вторжения в систему злоумышленника могут иметь значительные последствия для организации, в которой используется система верификации. Ошибка первого рода с этой точки зрения является менее значимым критерием, однако слишком большое количество отказов

санкционированным пользователям приведет к потере работоспособности системы и возмущению пользователей.

Временные критерии в данном случае играют гораздо меньшую роль из-за того, что они не оказывают влияние на работоспособность системы, но влияют на скорость и удобство работы с системой. Попарное сравнения критериев на основе данной шкалы приведено в Таблице 3.2.3.

Таблица 3.2.3 – Числовые оценки матрицы попарных сравнений для критериев

Критерии	Количество ошибок 1-го рода (отказов)	Количество ошибок 2-го рода	Среднее время обучения модели диктора	Среднее время проведения верификации для одной фразы
Количество ошибок 1-го рода (отказов)	1	1/2	9	7
Количество ошибок 2-го рода	2	1	9	7
Среднее время обучения модели диктора	1/9	1/9	1	1/3
Среднее время проведения верификации для одной фразы	1/7	1/7	3	1

Локальный приоритет *i*-го критерия x_i производится по формуле 3.2.1.

$$x_i = \frac{\sqrt[n]{\prod_{j=1}^{n} a_{ij}}}{\sum_{k=1}^{n} \sqrt[n]{\prod_{j=1}^{n} a_{kj}}}, \quad (3.2.1)$$

где a_{ij} – результат попарного сравнения *i*-го и *j*-го критерия, n – количество критериев, $i = 1..n$, $j = 1..n$, $k = 1..n$. Результат расчета локальных приоритетов критериев приведен в Таблице 3.2.4.

Таблица 3.2.4 – Числовые оценки локальных приоритетов для критериев

Критерии	Локальный приоритет критерия (вес критерия), x_i
Количество ошибок 1-го рода (отказов)	0,366
Количество ошибок 2-го рода	0,518
Среднее время обучения модели диктора	0,039
Среднее время проведения верификации для одной фразы	0,077

Согласно методу анализа иерархий необходимо провести проверку согласованности локальных приоритетов. Для этого требуется выполнение следующего условия: отношение согласованности (ОС) матрицы не должно превышать 10-15%. Данное отношение вычисляется согласно формулам (3.2.2-3.2.4):

$$\lambda_{max} = \sum_{i=1}^{n}\left(x_i \sum_{j=1}^{n} a_{ji}\right) \quad (3.2.2)$$

$$\text{ИС} = \frac{\lambda_{max} - n}{n - 1} \quad (3.2.3)$$

$$\text{ОС} = \frac{\text{ИС}}{\text{ПСС}}, \quad (3.2.4)$$

где λ_{max} – собственное число матрицы, ИС – индекс согласованности, ПСС – показатель случайной согласованности, для n = 4, составляет 0,9. После проведения расчетов были получена оценка согласованности матрицы попарных сравнений критериев, показанная в Таблице 3.2.5.

Таблица 3.2.5 – Численные результаты оценки согласованности для матрицы попарных сравнений критериев

λ_{max}	n	ИС	ПСС	ОС	Результат
4,14	4	0,047	0,9	5,18 %	Значение ОС < 10 % соответствует условию согласованности матрицы

Была произведена оценка альтернатив по каждому из критериев. Для проведения попарного сравнения были использованы экспериментальные значения критериев из Таблицы 3.2.1. Для каждого из критериев значения попарных сравнений и оценка согласованности рассчитаны в Таблицах 3.2.6–3.2.13:

Таблица 3.2.6 – Числовые оценки матрицы попарных сравнений для альтернатив по критерию “Количество ошибок 1-го рода (отказов)”

Альтернативы	Разработанное программное средство	ALIZE	MSR Identity Toolbox	Локальный приоритет альтернативы
Разработанное программное средство	1	4	6	0,6817
ALIZE	1/4	1	4	0,2363
MSR Identity Toolboxt	1/6	1/4	1	0,0819

Таблица 3.2.7 – Численные результаты оценки согласованности для матрицы попарных сравнений по критерию “Количество ошибок 1-го рода (отказов)”

λ_{max}	n	ИС	ПСС	ОС	Результат
3,1	3	0,054	0,58	9,3 %	Значение ОС < 10 % соответствует условию согласованности матрицы

Таблица 3.2.8 – Числовые оценки матрицы попарных сравнений для альтернатив по критерию “Количество ошибок 2-го рода”

Альтернативы	Разработанное программное средство	ALIZE	MSR Identity Toolbox	Локальный приоритет альтернативы
Разработанное программное средство	1	5	8	0,7334
ALIZE	1/5	1	4	0,1991
MSR Identity Toolboxt	1/8	1/4	1	0,0675

Таблица 3.2.9 – Численные результаты оценки согласованности для матрицы попарных сравнений по критерию “Количество ошибок 2-го рода”

λ_{max}	n	ИС	ПСС	ОС	Результат
3,09	3	0,047	0,58	8,1 %	Значение ОС < 10 % соответствует условию согласованности матрицы

Таблица 3.2.10 – Числовые оценки матрицы попарных сравнений для альтернатив по критерию “Среднее время обучения модели диктора”

Альтернативы	Разработанное программное средство	ALIZE	MSR Identity Toolbox	Локальный приоритет альтернативы
Разработанное программное средство	1	1/9	1/9	0,0513
ALIZE	9	1	2	0,582
MSR Identity Toolboxt	9	1/2	1	0,3667

Таблица 3.2.11 – Численные результаты оценки согласованности для матрицы попарных сравнений по критерию “Среднее время обучения модели диктора”

λ_{max}	n	ИС	ПСС	ОС	Результат
3,053	3	0,027	0,58	4,62 %	Значение ОС < 10 % соответствует условию согласованности матрицы

Таблица 3.2.12 – Числовые оценки матрицы попарных сравнений для альтернатив по критерию “Среднее время проведения верификации для одной фразы”

Альтернативы	Разработанное программное средство	ALIZE	MSR Identity Toolbox	Локальный приоритет альтернативы
Разработанное программное средство	1	1/5	1/5	0,0887
ALIZE	5	1	2	0,5591
MSR Identity Toolboxt	5	1/2	1	0,3522

Таблица 3.2.13 – Численные результаты оценки согласованности для матрицы попарных по критерию “Среднее время проведения верификации для одной фразы”

λ_{max}	n	ИС	ПСС	ОС	Результат
3,053	3	0,027	0,58	4,62 %	Значение ОС < 10 % соответствует условию согласованности матрицы

Для проведения сравнения альтернатив по предложенным критериям, необходимо рассчитать глобальный приоритет s_i i-ой альтернативы согласно формуле (3.2.5):

$$s_i = \sum_{j=1}^{n} x(A_i, K_j)x(K_j), \quad (3.2.5)$$

где n – количество критериев (n = 4); s_i – глобальный приоритет i-ой альтернативы, i = 1, 2, 3; $x(A_i, K_j)$ – локальный приоритет i-ой альтернативы в рамках j-го критерия; $x(K_j)$ – локальный приоритет j-го критерия. Рассчитанные глобальные приоритеты альтернатив представлены в Таблице 3.2.14.

Таблица 3.2.14 – Сводная таблица рассчитанных локальных приоритетов и рассчитанные значения глобальных приоритетов альтернатив

Альтернативы	Количество ошибок 1-го рода (отказов), K_1	Количество ошибок 2-го рода, K_2	Среднее время обучения модели диктора, K_3	Среднее время проведения верификации для одной фразы, K_4	Глобальный приоритет альтернативы A_i (s_i)
	Локальный приоритет, $x(K_i)$				
	0,366	0,518	0,039	0,077	
Разработанное программное средство (A_1)	0,6817	0,7334	0,0513	0,0887	0,6382
ALIZE (A_2)	0,2363	0,1991	0,582	0,5591	0,2554
MSR Identity Toolboxt (A_3)	0,0819	0,0675	0,3667	0,3522	0,1064

На основе полученных результатов можно сделать вывод, что среди рассмотренных альтернатив наилучшей по представленным критериям является разработанное программное средство с глобальным приоритетом альтернативы s_1 = 0,6382. Соответственно, можно говорить о превосходстве разработанного программного средства по точности верификации диктора, при условии, что время работы системы остается адекватным по требуемым параметрам.

Разработанное программное средство верификации диктора по произвольной фразе внедрено в деятельность АО «ОЭЗ ТВТ «Томск» (Приложение А). Программное средство верификации диктора по произвольной фразе позволило достичь точности верификации пользователей автоматизированных систем, составляющей 99,5%. По сравнению с аналогами, использовавшимися при апробации, общая ошибка верификации уменьшена на 28,5%.

Результаты данной работы также используются в учебном процессе на факультете безопасности ТУСУР при чтении курса лекций и проведении лабораторных работ по дисциплине “Программно-аппаратные средства

обеспечения информационной безопасности” для подготовки студентов, обучающихся по специальностям “10.05.02 – Информационная безопасность телекоммуникационных систем” и “10.05.03 – Информационная безопасность автоматизированных систем” (Приложение Б). Алгоритмы верификации диктора по голосу включены в лекционный материал на тему “Методы биометрической аутентификации”. Лабораторная работа, посвященная верификации диктора по голосу с применением экспериментальных данных, проводится с использованием программного средства, реализующего разработанные алгоритмы верификации диктора.

Разработанные алгоритмы и программное средство использованы в рамках мероприятия 1.3 ФЦП “Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014—2020 годы” (соглашение о предоставлении субсидии № 14.577.21.0172 от 27 октября 2015 г.; уникальный идентификатор RFMEFI57715X0172).

Проведенные исследования и разработанные алгоритмы были получены в рамках выполнения базовой части государственного задания Минобрнауки России, проект 8.9628.2017/8.9 на базе лаборатории медико-биологических исследований (ЛМБИ) ТУСУР. Таким образом, полученные от внедрения результаты подтверждают защищаемые положения данной научной работы.

3.3 Выводы

1. Разработано программное средство для верификации диктора по произвольной фразе. Данное средство включает в себя все необходимые модули для извлечения речевых признаков, обучения моделей дикторов и УФМ, а также проведения верификационных испытаний. Средство позволяет произвести отбор речевых признаков с помощью алгоритма жадного добавления-удаления и генетического алгоритма.

2. Были произведены эксперименты по оценке времени работы модуля обучения УФМ в различных реализациях – на центральном процессоре, на процессоре видеокарты и в комбинированном варианте. Реализованный модуль обучения УФМ с комбинированными вычислениями на центральном и графическом процессоре по сравнению с обучением УФМ на центральном процессоре позволяет уменьшить время работы на 36,95%, по сравнению с обучением на процессоре видеокарты позволяет уменьшить время работы на 10%.

3. Разработанное программное средство верификации диктора по произвольной фразе внедрено в деятельность АО «ОЭЗ ТВТ «Томск», где применяется с целью осуществления контроля доступа к рабочим компьютерам сотрудников предприятия. Программное средство верификации диктора по произвольной фразе позволило достичь точности верификации сотрудников, составляющей 99,5 %. По сравнению с аналогами, использовавшимися при апробации, общая ошибка верификации уменьшена на 28,5 %.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена важная задача повышения точности методов верификации диктора по произвольной фразе, имеющая существенное значение для развития теории и практики обработки речи и машинного обучения.

Основные результаты диссертационной работы:

1. Произведен обзор существующих методов и алгоритмов верификации диктора по произвольной фразе. К основным методам верификации диктора можно отнести методы, использующие Гауссовы смеси; методы, основанные на факторном анализе; методы с применением глубоких нейронных сетей. В качестве речевых признаков, используемых для верификации диктора, применяют мел-кепстральные коэффициенты; формантные частоты; признаки, извлеченные из глубоких нейронных сетей; частоту основного тона; энергию сигнала и другие.

2. Разработан оригинальный алгоритм верификации диктора, отличающийся от существующих применением речевых признаков, полученных с помощью жадного алгоритма отбора признаков. В алгоритме используется следующий набор признаков: 13 мел-кепстральных коэффициентов, 10 дельта и 2 двойных дельта мел-кепстральных коэффициента, вероятность вокализации, коэффициент линейного предсказания и пара линейного спектра.

3. Разработан алгоритм генерации признаков, основанный на применении сверточной глубокой сети доверия. Данный алгоритм отличается от существующих расширенной архитектурой нейронной сети, позволяющей выделять более высокоуровневые признаки и уменьшить их количество. Применение признаков, полученных с помощью созданного алгоритма, возможно для повышения точности систем верификации диктора, распознавания речи или идентификации пола говорящего.

4. Разработан гибридный алгоритм верификации диктора по произвольной фразе на основе ансамбля классификаторов. Отличительной особенностью алгоритма является применение в ансамбле классификаторов, использующих признаки, извлеченные из аудиозаписей с помощью сверточной глубокой сети доверия.

5. Создано программное средство верификации диктора по произвольной фразе, отличающееся от существующих применением алгоритмов обучения универсальной фоновой модели (УФМ) на центральном и графическом процессорах.

6. Алгоритм верификации, использующий речевые признаки, полученные с помощью жадного алгоритма отбора признаков, и гибридный алгоритм на основе ансамбля классификаторов показали лучшую точность верификации диктора по сравнению с аналогами. Созданное программное средство верификации диктора по произвольной фразе внедрено в деятельность АО «ОЭЗ ТВТ «Томск», где применяется с целью контроля доступа к рабочим компьютерам сотрудников предприятия.

СПИСОК СОКРАЩЕНИЙ

ГА – генетический алгоритм.

ГНС – глубокая нейронная сеть.

ГП – графический процессор.

ГС – Гауссова смесь.

ДПФ – дискретное преобразование Фурье.

ЖА – жадный алгоритм.

ЛДА – линейный дискриминантный анализ, Linear Discriminant Analysis.

МКК – мел-кепстральные коэффициенты.

СГСД – сверточная глубокая сеть доверия, Convolutional Deep Belief Network.

СОМБ – сверточная ограниченная машина Больцмана, Convolutional Restricted Boltzmann Machine.

УФМ – универсальная фоновая модель.

ЦП – центральный процессор.

ША – шумоподавляющие автоэнкодеры, Denoising Auto-Encoders.

BNF – Bottleneck Features.

DET – Detection Error Trade-off, компромиссное определение ошибки.

EER – равная ошибка первого и второго рода.

EM – Expectation-Maximization.

JFA – Joint Factor Analysis, комбинированный факторный анализ.

LPC - Linear Prediction Coefficients, коэффициенты линейного предсказания.

LSP – Line Spectral Pair, пары линейного спектра.

MAP – Maximum a posteriori, апостериорный максимум.

MFCC - Mel frequency cepstral coefficients, мел-кепстральные коэффициенты.

minDCF – минимальная функция стоимости обнаружения, Minimum Detection Cost Function.

SVM – Support Vector Machine, машина опорных векторов.

СПИСОК ЛИТЕРАТУРЫ

1. Reynolds D. A., Quatieri T. F., Dunn R. B. Speaker verification using adapted Gaussian mixture models // Digital signal processing. – 2000. – Т. 10. – №. 1. – С. 19-41.

2. Reynolds D. A., Rose R. C. Robust text-independent speaker identification using Gaussian mixture speaker models // IEEE transactions on Speech and Audio Processing. – 1995. – Т. 3. – №. 1. – С. 72-83.

3. Rosenberg A. E. The use of cohort normalized scores for speaker verification // Proc. ICSLP-92. – 1992. – С. 599-602.

4. Rosenberg A. E., Parthasarathy S. Speaker background models for connected digit password speaker verification // Acoustics, Speech, and Signal Processing (ICASSP-96), IEEE International Conference on. – 1996. – Т. 1. – С. 81-84.

5. Matsui T., Furui S. Likelihood normalization for speaker verification using a phoneme-and speaker-independent model // Speech communication. – 1995. – Т. 17. – №. 1. – С. 109-116.

6. Сорокин В. Н., Вьюгин В. В., Тананыкин А. А. Распознавание личности по голосу: аналитический обзор // Информационные процессы. – 2012. – Т. 12. – №. 1. – С. 1-30.

7. Murty K. S. R., Yegnanarayana B. Combining evidence from residual phase and MFCC features for speaker recognition // IEEE signal processing letters. – 2006. – Т. 13. – №. 1. – С. 52-55.

8. Campbell W. M., Sturim D. E., Reynolds D. A. Support vector machines using GMM supervectors for speaker verification // IEEE signal processing letters. – 2006. – Т. 13. – №. 5. – С. 308-311.

9. Капустин А.И., Симончик К.К. Система верификации дикторов по голосу на основе использования CGP-SVM подхода // Труды 12-й международной конференции «Цифровая обработка сигналов и ее применение» (DSPA-2010). – Т. 2. – 2010. – С. 207-210.

10. Liu C. S. et al. Study of line spectrum pair frequencies for speaker recognition // Acoustics, Speech, and Signal Processing (ICASSP-90), International Conference on. – 1990. – С. 277-280.

11. Hermansky H. Perceptual linear predictive (PLP) analysis of speech // The Journal of the Acoustical Society of America. – 1990. – Т. 87. – №. 4. – С. 1738-1752.

12. Adami A. G. et al. Modeling prosodic dynamics for speaker recognition // Acoustics, Speech, and Signal Processing (ICASSP'03), IEEE International Conference on. – 2003. – Т. 4. – С. 788-791.

13. Farrús M. Jitter and shimmer measurements for speaker recognition // 8th Annual Conference of the International Speech Communication Association. – 2007. – С. 78-81.

14. Davis S., Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences // IEEE transactions on acoustics, speech, and signal processing. – 1980. – Т. 28. – №. 4. – С. 357-366.

15. Atal B.S. Automatic recognition of speakers from their voices // Proceedings of the IEEE. – 1976. – Т. 64. – № 4. – С. 460-475.

16. Jurafsky D., Martin J.H. Speech and Language Processing, second ed. – New Jersey: Pearson Education, 2009. – 1027 с.

17. Lavner Y., Gath I., Rosenhouse J. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels // Speech Communication. – 2000. – Т. 30. – №. 1. – С. 9-26.

18. Takemoto H. et al. Acoustic roles of the laryngeal cavity in vocal tract resonance // The Journal of the Acoustical Society of America. – 2006. – Т. 120. – №. 4. – С. 2228-2238.

19. Сапожков М.А. Речевой сигнал в кибернетике и связи. – М.: Государственное издательство по вопросам связи и литературы, 1963. – 453 с.

20. Rabiner L. R., Schafer R. W. Digital processing of speech signals. – NJ.: Prentice Hall, 1978 – 509 с.

21. Snell R. C., Milinazzo F. Formant location from LPC analysis data // IEEE Transactions on Speech and Audio Processing. – 1993. – Т. 1. – №. 2. – С. 129-134.

22. Welling L., Ney H. Formant estimation for speech recognition // IEEE Transactions on Speech and Audio Processing. – 1998. – Т. 6. – №. 1. – С. 36-48.

23. Kim C., Sung W. Vowel pronunciation accuracy checking system based on phoneme segmentation and formants extraction // Proceedings of International Conference on Speech Processing. – 2001. – С. 447-452.

24. Atal B. S., Hanauer S. L. Speech analysis and synthesis by linear prediction of the speech wave // The journal of the acoustical society of America. – 1971. – Т. 50. – №. 2-2. – С. 637-655.

25. Rabiner L. et al. A comparative performance study of several pitch detection algorithms // IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1976. – Т. 24. – №. 5. – С. 399-418.

26. Deller Jr J. R., Proakis J. G., Hansen J. H. Discrete time processing of speech signals. – Wiley-IEEE Press, 1999 – 936 с.

27. Bell C. G. et al. Reduction of Speech Spectra by Analysis-by-Synthesis Techniques // The Journal of the Acoustical Society of America. – 1961. – Т. 33. – №. 12. – С. 1725-1736.

28. Daqrouq K. et al. Wavelet formants speaker identification based system via neural network // International Journal of Recent Trends in Engineering. – 2009. – Т. 2. – №. 5. – С. 140-144.

29. Kim C., Seo K., Sung W. A robust formant extraction algorithm combining spectral peak picking and root polishing // EURASIP Journal on Applied Signal Processing. – 2006. – Т. 2006. – С. 1-16.

30. Kaneko T., Shimamura T. Noise-Reduced Complex LPC Analysis for Formant Estimation of Noisy Speech // International Journal of Electronics and Electrical Engineering. – 2014. – Т. 2. – № 2. – С. 90-94.

31. Iwai Y., Shimamura T. Formant frequency estimation with windowless autocorrelation in the presence of noise // Circuits and Systems (APCCAS), 2014 IEEE Asia Pacific Conference on. – 2014. – С. 81-84.

32. Сорокин В. Н., Леонов А. С., Макаров И. С. Устойчивость оценок формантных частот // Речевые технологии. – 2009. – № 1. – С. 3-21.

33. Сорокин В. Н., Ромашкин Ю. Н., Тананыкин А. А. Распознавание пола по параметрам голосового источника // Речевые технологии. – 2012. – № 4. – С. 49-67.

34. Dissen Y., Keshet J. Formant estimation and tracking using deep learning // The 17th Annual Conference of the International Speech Communication Association. – 2016.

35. Sambur M. Selection of acoustic features for speaker identification // IEEE Transactions on Acoustics, Speech, and Signal Processing. – 1975. – Т. 23. – №. 2. – С. 176-182.

36. Osanai P. R. T., Kinoshita Y. Strength of forensic speaker identification evidence: multispeaker formant and cepstrum-based segmental discrimination with a bayesian likelihood ratio as threshold // Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne. – 2002. – С. 303-308.

37. Goldstein U. G. Speaker-identifying features based on formant tracks // The Journal of the Acoustical Society of America. – 1976. – Т. 59. – №. 1. – С. 176-182.

38. Kinnunen T., Hautamäki V., Fränti P. Fusion of spectral feature sets for accurate speaker identification // In Proc. 9th Int. Conf. Speech and Computer (SPECOM 2004). – 2004. – С. 361-365.

39. Ручай А. Н. Формантный метод текстозависимой верификации диктора // Вестник Челябинского государственного университета. – 2010. – №. 23. – С. 121-131.

40. Lu X., Dang J. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification // Speech communication. – 2008. – Т. 50. – №. 4. – С. 312-322.

41. Becker T., Jessen M., Grigoras C. Forensic speaker verification using formant features and Gaussian mixture models // Interspeech. – 2008. – С. 1505-1508.

42. Rose R. C., Reynolds D. A. Text independent speaker identification using automatic acoustic segmentation // Acoustics, Speech, and Signal Processing (ICASSP-90), International Conference on. – 1990. – С. 293-296.

43. Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm // Journal of the royal statistical society. Series B (methodological). – 1977. – С. 1-38.

44. Baum L. E. et al. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains // The annals of mathematical statistics. – 1970. – Т. 41. – №. 1. – С. 164-171.

45. Campbell W. M. et al. Support vector machines for speaker and language recognition // Computer Speech & Language. – 2006. – Т. 20. – №. 2. – С. 210-229.

46. Fine S., Navratil J., Gopinath R. A. A hybrid GMM/SVM approach to speaker identification // Acoustics, Speech, and Signal Processing (ICASSP'01), IEEE International Conference on. – 2001. – Т. 1. – С. 417-420.

47. Solomonoff A., Campbell W. M., Boardman I. Advances in channel compensation for SVM speaker recognition // Acoustics, Speech, and Signal Processing (ICASSP'05), IEEE International Conference on. – 2005. – Т. 1. – С. 629-632.

48. You C. H., Lee K. A., Li H. GMM-SVM kernel with a Bhattacharyya-based distance for speaker recognition // IEEE Transactions on Audio, Speech, and Language Processing. – 2010. – Т. 18. – №. 6. – С. 1300-1312.

49. Meuwly D., Drygajlo A. Forensic speaker recognition based on a Bayesian framework and Gaussian mixture modelling (GMM) // 2001: A Speaker Odyssey-The Speaker Recognition Workshop. – 2001. – C.145-150.

50. Ding J., Yen C. T. Enhancing GMM speaker identification by incorporating SVM speaker verification for intelligent web-based speech applications // Multimedia Tools and Applications. – 2015. – Т. 74. – №. 14. – С. 5131-5140.

51. Campbell W. M. A SVM/HMM system for speaker recognition // Acoustics, Speech, and Signal Processing (ICASSP'03), IEEE International Conference on. – 2003. – Т. 2. – С. 209-212.

52. Nakagawa S., Zhang W., Takahashi M. Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM // Acoustics, Speech, and Signal Processing (ICASSP'04), IEEE International Conference on. – 2004. – Т. 1. – С. 81-84.

53. Wang L., Kitaoka N., Nakagawa S. Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM // Speech communication. – 2007. – Т. 49. – №. 6. – С. 501-513.

54. Rodríguez E. et al. Speech/speaker recognition using a HMM/GMM hybrid model // Proceedings of the First International Conference on Audio- and Video-Based Biometric Person Authentication. – 1997. – С. 227-234.

55. Seo C., Lee K. Y., Lee J. GMM based on local PCA for speaker identification // Electronics Letters. – 2001. – Т. 37. – №. 24. – С. 1486-1488.

56. Thyes O. et al. Speaker identification and verification using eigenvoices // INTERSPEECH. – 2000. – С. 242-245.

57. Lee K. Y. Local fuzzy PCA based GMM with dimension reduction on speaker identification // Pattern recognition letters. – 2004. – Т. 25. – №. 16. – С. 1811-1817.

58. Kinnunen T., Karpov E., Franti P. Real-time speaker identification and verification // IEEE Transactions on Audio, Speech, and Language Processing. – 2006. – Т. 14. – №. 1. – С. 277-288.

59. Pelecanos J. et al. Vector quantization based Gaussian modeling for speaker verification // Pattern Recognition, Proceedings, 15th International Conference on. – 2000. – Т. 3. – С. 294-297.

60. Chen K., Wang L., Chi H. Methods of combining multiple classifiers with different features and their applications to text-independent speaker identification // International Journal of Pattern Recognition and Artificial Intelligence. – 1997. – Т. 11. – №. 03. – С. 417-445.

61. AboElenein N. M. et al. Improved text-independent speaker identification system for real time applications // Electronics, Communications and Computers (JEC-ECC), Fourth International Japan-Egypt Conference on. – 2016. – С. 58-62.

62. Desai D., Joshi M. Speaker recognition using MFCC and hybrid model of VQ and GMM // Recent Advances in Intelligent Informatics. – 2014. – С. 53-63.

63. Motlicek P. et al. Employment of subspace gaussian mixture models in speaker recognition // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. – 2015. – С. 4445-4449.

64. Bao L., Shen X. Improved Gaussian mixture model and application in speaker recognition // Control, Automation and Robotics (ICCAR), 2nd International Conference on. – 2016. – С. 387-390.

65. Yang Y., Deng L. Score regulation based on GMM Token Ratio Similarity for speaker recognition // Chinese Spoken Language Processing (ISCSLP), 9th International Symposium on. – 2014. – С. 424-424.

66. Nakagawa S., Asakawa K., Wang L. Speaker recognition by combining MFCC and phase information // Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH). – 2007. С.2005-2008.

67. Hosseinzadeh D., Krishnan S. Combining vocal source and MFCC features for enhanced speaker recognition performance using GMMs // Multimedia Signal Processing (MMSP), IEEE 9th Workshop on. – 2007. – С. 365-368.

68. Miyajima C. et al. Speaker identification using Gaussian mixture models based on multi-space probability distribution // Acoustics, Speech, and Signal Processing (ICASSP'01), IEEE International Conference on. – 2001. – Т. 1. – С. 433-436.

69. Reynolds D. A. Comparison of background normalization methods for text-independent speaker verification // Fifth European Conference on Speech Communication and Technology. – 1997. С. 963-966.

70. Hermansky H., Malayath N. Speaker verification using speaker-specific mappings // Proc. of Speaker Recognition and its Commercial and Forensic Applications. – 1998. – С.111-114.

71. Quatieri T. F. et al. Speaker and language recognition using speech codec parameters // Proc. Eurospeech'99. – 1999. – Т. 2. – №. 1. – С. 787-790.

72. Isobe T., Takahashi J. Text-independent speaker verification using virtual speaker based cohort normalization // Sixth European Conference on Speech Communication and Technology. – 1999. – С. 987-990.

73. Duda R. O. et al. Pattern classification and scene analysis. – New York : Wiley, 1973. – 512 с.

74. Gauvain J. L., Lee C. H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains // Speech and audio processing, IEEE transactions on. – 1994. – Т. 2. – №. 2. – С. 291-298.

75. Kenny P. et al. A study of interspeaker variability in speaker verification // Audio, Speech, and Language Processing, IEEE Transactions on. – 2008. – Т. 16. – №. 5. – С. 980-988.

76. Kenny P. et al. Joint factor analysis versus eigenchannels in speaker recognition // Audio, Speech, and Language Processing, IEEE Transactions on. – 2007. – Т. 15. – №. 4. – С. 1435-1447.

77. Kenny P. et al. Speaker and session variability in GMM-based speaker verification // Audio, Speech, and Language Processing, IEEE Transactions on. – 2007. – Т. 15. – №. 4. – С. 1448-1460.

78. Dehak N. et al. Front-end factor analysis for speaker verification // Audio, Speech, and Language Processing, IEEE Transactions on. – 2011. – Т. 19. – №. 4. – С. 788-798.

79. Kenny P., Boulianne G., Dumouchel P. Eigenvoice modeling with sparse training data // Speech and Audio Processing, IEEE Transactions on. – 2005. – Т. 13. – №. 3. – С. 345-354.

80. Vapnik V. The nature of statistical learning theory. – Springer Science & Business Media, 2013. – 340 с.

81. Hatch A. O., Kajarekar S. S., Stolcke A. Within-class covariance normalization for SVM-based speaker recognition // Proc. Interspeech (ICSLP). – 2006. – С. 1471–1474.

82. Campbell W. M. et al. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. – 2006. – Т. 1. – С. 97-100.

83. Kenny P. Bayesian speaker verification with heavy-tailed priors // Odyssey: The Speaker and Language Recognition Workshop. – 2010. – С. 1-10.

84. Garcia-Romero D., Espy-Wilson C. Y. Analysis of i-vector Length Normalization in Speaker Recognition Systems // Proc. Interspeech (ICSLP). – 2011. – С. 249-252.

85. Matějka P. et al. Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. – 2011. – С. 4828-4831.

86. Garcia-Romero D., McCree A. Supervised domain adaptation for i-vector based speaker recognition // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. – 2014. – С. 4047-4051.

87. Richardson F., Nemsick B., Reynolds D. Channel compensation for speaker recognition using MAP adapted PLDA and denoising DNNs //Proc. Speaker Lang. Recognit. Workshop. – 2016. – С. 225-230.

88. Mak M. W., Pang X., Chien J. T. Mixture of PLDA for noise robust i-vector speaker verification // IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP). – 2016. – Т. 24. – №. 1. – С. 130-142.

89. Cumani S., Laface P. I–vector transformation and scaling for PLDA based speaker recognition // Proc. Odyssey. – 2016. – С. 39-46.

90. NIST Speaker Recognition Evaluation // Speaker Recognition: [сайт]. [2017]. URL: https://www.nist.gov/itl/iad/mig/speaker-recognition (дата обращения: 29.09.2017)

91. Rouvier M. et al. LIA system description for NIST SRE 2016 // arXiv preprint: [сайт]. [2016]. URL: https://arxiv.org/pdf/1612.05168.pdf (дата обращения: 29.09.2017)

92. Madikeri S. et al. IDIAP submission to the NIST SRE 2016 speaker recognition evaluation // Idiap: [сайт]. [2016]. URL: https://infoscience.epfl.ch/record/223757/files/Madikeri_Idiap-RR-32-2016.pdf (дата обращения: 29.09.2017)

93. Zeinali H., Sameti H., Maghsoodi N. SUT System Description for NIST SRE 2016 // arXiv preprint: [сайт]. [2017]. URL: https://arxiv.org/pdf/1706.05077.pdf (дата обращения: 29.09.2017)

94. Stafylakis T. et al. Compensation for phonetic nuisance variability in speaker recognition using DNNs // Odyssey: The Speaker and Language Recognition Workshop. – 2016. – С. 340-345.

95. Kenny P. et al. Deep neural networks for extracting baum-welch statistics for speaker recognition // Proc. Odyssey. – 2014. – С. 293-298.

96. Variani E. et al. Deep neural networks for small footprint text-dependent speaker verification // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. – 2014. – С. 4052-4056.

97. Ahmad K. S. et al. A unique approach in text independent speaker recognition using MFCC feature sets and probabilistic neural network // Advances in Pattern Recognition (ICAPR), Eighth International Conference on. – 2015. – С. 1-6.

98. Lei Y. et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. – 2014. – С. 1695-1699.

99. McLaren M., Ferrer L., Lawson A. Exploring the role of phonetic bottleneck features for speaker and language recognition // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. – 2016. – С. 5575-5579.

100. Richardson F., Reynolds D., Dehak N. Deep neural network approaches to speaker and language recognition // IEEE Signal Processing Letters. – 2015. – Т. 22. – №. 10. – С. 1671-1675.

101. Новосёлов С. А. и др. Противодействие спуфинг атакам на голосовые биометрические системы // Речевые технологии. – 2016. – № 2. – С. 22-31.

102. McLaren M., Lei Y., Ferrer L. Advances in deep neural network approaches to speaker recognition // Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. – 2015. – С. 4814-4818.

103. Vincent P. et al. Extracting and composing robust features with denoising autoencoders // Proceedings of the 25th international conference on Machine learning. – 2008. – С. 1096-1103.

104. Greenberg C. S. et al. The 2012 NIST speaker recognition evaluation // INTERSPEECH. – 2013. – С. 1971-1975.

105. Kudashev O. et al. Usage of DNN in speaker recognition: advantages and problems // International Symposium on Neural Networks. – 2016. – С. 82-91.

106. Eyben F. et al. Recent developments in opensmile, the munich open-source multimedia feature extractor // Proceedings of the 21st ACM international conference on Multimedia. – 2013. – С. 835–838.

107. Sadjadi S.O., Slaney M., Heck L. MSR identity toolbox v1.0: A MATLAB toolbox for speaker-recognition research // Speech and Language Processing Technical Committee Newsletter. – 2013. – Т. 1. – № 4. – С. 1–32.

108. Martin A. et al. The DET curve in assessment of detection task performance. – National Institute of Standards and Technology (NIST), Gaithersburg. – 1997. – С. 1-5.

109. Rakhmanenko I., Meshcheryakov R. Speech Features Evaluation for Small Set Automatic Speaker Verification Using GMM-UBM System // Speech and Computer (SPECOM 2016), Lecture Notes in Computer Science. – 2016. – № 9811. – С. 645-650.

110. Chandrashekar G., Sahin F. A survey on feature selection methods // Computers & Electrical Engineering. – 2014. – Т. 40. – №. 1. – С. 16-28.

111. Kohavi R., John G. H. Wrappers for feature subset selection // Artificial intelligence. – 1997. – Т. 97. – №. 1-2. – С. 273-324.

112. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. – Новосибирск: Изд-во Института математики, 1999. – 270 с.

113. Лбов Г. С. Выбор эффективной системы признаков // Вычислительные системы. – 1965. – № 19. – С. 21–34.

114. Емельянов В. В., Курейчик В. В., Курейчик В. М. Теория и практика эволюционного моделирования. – М.: Физматлит, 2003. – 432 с.

115. Рахманенко И.А., Мещеряков Р.В. Анализ идентификационных признаков в речевых данных с помощью GMM-UBM системы верификации диктора // Труды СПИИРАН. – 2017. – Т. 52. – № 3. – С.22-50.

116. Кормен Т. и др. Алгоритмы. Построение и анализ. Глава 16. Жадные алгоритмы:[пер. с англ.]. – Издательский дом Вильямс. – 2012. – 1296 с.

117. Holland J.H. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence // MIT press. – 1992. – 232 с.

118. Lipowski A., Lipowska D. Roulette-wheel selection via stochastic acceptance // Physica A: Statistical Mechanics and its Applications. – 2012. – Т. 391. – №. 6. – С. 2193-2196.

119. Grimaldi M., Cummins F. Speech style and speaker recognition: a case study // INTERSPEECH. – 2009. – С. 920-923.

120. Cummins F. et al. The chains corpus: Characterizing individual speakers // Proc. of SPECOM. – 2006. – Т. 6. – С. 431-435.

121. Grimaldi M., Cummins F. Speaker identification using instantaneous frequencies // IEEE Transactions on Audio, Speech, and Language Processing. – 2008. – Т. 16. – №. 6. – С. 1097-1111.

122. Lee H. et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations // Proceedings of the 26th annual international conference on machine learning. – 2009. – С. 609-616.

123. Lee H. et al. Unsupervised feature learning for audio classification using convolutional deep belief networks // Advances in neural information processing systems. – 2009. – С. 1096-1104.

124. Hinton G. E., Osindero S., Teh Y. W. A fast learning algorithm for deep belief nets // Neural computation. – 2006. – Т. 18. – №. 7. – С. 1527-1554.

125. Desjardins G., Bengio Y. Empirical evaluation of convolutional RBMs for vision // Technical Report 1327, Dept. IRO, Université de Montréal. – 2008. – С. 1-13.

126. Hinton G. E. Training products of experts by minimizing contrastive divergence // Neural Computation. – 2002. – Т. 14. – №. 8. – С. 1771-1800.

127. Jolliffe I. T. Principal Component Analysis and Factor Analysis // Principal component analysis. – Springer New York, 1986. – С. 115-128.

128. Cortes C., Vapnik V. Support-vector networks // Machine learning. – 1995. – Т. 20. – №. 3. – С. 273-297.

129. Fisher R. A. The use of multiple measurements in taxonomic problems // Annals of human genetics. – 1936. – Т. 7. – №. 2. – С. 179-188.

130. Freund Y., Schapire R. E. A desicion-theoretic generalization of on-line learning and an application to boosting // European conference on computational learning theory. – 1995. – С. 23-37.

131. Freund Y. et al. Experiments with a new boosting algorithm // Machine Learning, Proceedings of the Thirteenth International Conference. – 1996. – Т. 96. – С. 148-156.

132. Eibl G., Pfeiffer K. P. How to make AdaBoost.M1 work for weak base classifiers by changing only one line of the code // European Conference on Machine Learning. – 2002. – С. 72-83.

133. Мокшин В. В. и др. Определение транспортных средств на участках дорог классификатором Хаара и оператором LBP с применением AdaBoost и отсечением по дорожной разметке // Вестник Казанского технологического университета. – 2016. – Т. 19. – №. 18.

134. Рахманенко И.А. Программный комплекс для идентификации диктора по голосу с применением параллельных вычислений на центральном и графическом процессорах // Доклады ТУСУР. – 2017. – Т. 20. – № 1. – С. 70–74.

135. Bonastre J.F., Wils F., Meignier S. ALIZE, a free toolkit for speaker recognition // Acoustics, Speech, and Signal Processing. – 2005. – Т. 1. – С. 737–740.

136. Габдуллин В.В., Капустин А.И., Королев А.И. Применение технологии CUDA для задач голосовой биометрии на примере построения универсальной фоновой модели диктора // Параллельные вычислительные технологии (ПаВТ’2011). – 2011. – С. 107-116.

137. Larcher A. et al. ALIZE 3.0-open source toolkit for state-of-the-art speaker recognition // Interspeech. – 2013. – С. 2768-2772.

138. Саати Т. Принятие решений: метод анализа иерархий. – М. : Радио и связь, 1993. – 278 с.

139. Коробов В.Б., Тутыгин А.Г. Преимущества и недостатки метода анализа иерархий // Известия РГПУ им. А.И. Герцена. – 2010. – № 122. – С. 108-115.

ПРИЛОЖЕНИЕ А

Акт о внедрении в деятельность АО «ОЭЗ ТВТ «Томск»

Экз.№1

АО «ОСОБАЯ ЭКОНОМИЧЕСКАЯ ЗОНА
ТЕХНИКО-ВНЕДРЕНЧЕСКОГО ТИПА «ТОМСК»
Академический проспект, 8/8,
г. Томск, 634055
тел. (3822) 488-650, факс (3822) 488-665
office@oez.tomsk.ru
ОКПО 95124992 , ОГРН 1067017162420
ИНН/КПП 7017153992/701701001

www.oez.tomsk.ru

_____ № _____

На № _____ от _____

«УТВЕРЖДАЮ»
Советник генерального директора
по безопасности
_____ М.Г. Клименко
«12» сентября 2017 г.

А К Т

О внедрении результатов кандидатской диссертационной работы
Рахманенко Ивана Андреевича

Комиссия в составе:
Начальник службы безопасности – Сорокин Евгений Викторович
Заместитель начальника службы безопасности – Исхаков Сергей Юнусович

составила настоящий акт о нижеследующем.

Особая экономическая зона технико-внедренческого типа «Томск» предоставляет современную инфраструктуру для компаний, занимающихся инновационным бизнесом. В задачи ОЭЗ ТВТ «Томск» входит обеспечение безопасности зданий и промышленных объектов, расположенных на Южной и Северной площадках.

В рамках совместной деятельности ТУСУРа и ОЭЗ ТВТ «Томск» результаты диссертационной работы Рахманенко И.А. используются для осуществления контроля доступа к рабочим компьютерам сотрудников предприятия. Разработанное Рахманенко И.А. программное средство (ПС) позволяет осуществлять верификацию сотрудников предприятия, используя произвольные фразы. Благодаря этому упрощается процесс верификации сотрудников, так как отсутствует необходимость в запоминании сложных паролей. Применение произвольных парольных фраз позволяет осуществить защиту от повторного использования скрытно записанного пароля.

Экз.№1

Результаты работы Рахманенко И.А. внедрены в деятельность ОЭЗ ТВТ «Томск», благодаря чему были достигнуты следующие показатели:

1. Апробация программного средства, использующего предложенные Рахманенко И.А. алгоритмы верификации диктора по произвольной фразе, показала пригодность для использования с целью верификации пользователей предприятия. Представленное программное средство и алгоритмы отличается от аналогов применением речевых признаков, полученных с помощью жадного алгоритма отбора признаков и сверточной глубокой сети доверия.

2. Разработанное программное средство верификации диктора по произвольной фразе позволило достичь точности верификации пользователей автоматизированных систем, составляющей 99,5%. По сравнению с аналогами, использовавшимися при апробации, общая ошибка верификации уменьшена на 28,5%.

Настоящий акт составлен в 3 (трех) экземплярах.

Члены комиссии:

Начальник службы безопасности — Е.В. Сорокин

Заместитель начальника службы безопасности — С.Ю. Исхаков

ПРИЛОЖЕНИЕ Б

Акт о внедрении в учебный процесс

ТУСУР

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное
образовательное учреждение высшего образования
**«ТОМСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И
РАДИОЭЛЕКТРОНИКИ»**

ОКПО 02069326, ОГРН 1027000867068,
ИНН 7021000043, КПП 701701001

пр. Ленина, д. 40, г. Томск, 634050

тел: (382 2) 510-530
факс: (382 2) 513-262, 526-365
e-mail: office@tusur.ru
http:// www.tusur.ru

_____№_____

Утверждаю
Проректор по учебной работе

_____ П..Е. Троян

А К Т

Об использовании результатов диссертационной работы на соискание ученой степени кандидата технических наук Рахманенко Ивана Андреевича

Комиссия в составе:

Председателя:

Давыдова Е.М., декан факультета безопасности, канд. техн. наук.

Членов комиссии:

Костюченко Е.Ю., доцент каф. КИБЭВС ТУСУР, канд. техн. наук;

Евсютин О.О., доцент каф. БИС ТУСУР, канд. техн. наук.

составили настоящий акт о нижеследующем:

Результаты диссертационной работы И.А. Рахманенко на тему “Алгоритмы и программные средства верификации диктора по произвольной фразе” используются в учебном процессе на факультете безопасности ТУСУР при чтении курса лекций и проведении лабораторных работ по дисциплине “Программно-аппаратные средства обеспечения информационной безопасности” для подготовки студентов, обучающихся по специальностям “10.05.02 – Информационная безопасность телекоммуникационных систем” и “10.05.03 – Информационная безопасность автоматизированных систем ”.

Алгоритмы верификации диктора по голосу, предлагаемые Рахманенко И.А. включены в лекционный материал на тему “Методы биометрической аутентификации”. Лабораторная работа, посвященная верификации диктора по голосу с применением экспериментальных данных, проводится с использованием программного комплекса, реализующего разработанные Рахманенко И.А. алгоритмы верификации диктора. В данном программном комплексе используется алгоритм генерации признаков, основанный на применении глубокой нейронной сети доверия. Используется гибридный алгоритм верификации диктора по произвольной фразе на основе модели Гауссовых смесей и машины опорных векторов.

Результаты диссертационной работы Рахманенко И.А. включены в одну лекцию и одну лабораторную работу по дисциплине “Программно-аппаратные средства обеспечения информационной безопасности” на кафедрах комплексной информационной безопасности электронно-вычислительных систем и безопасности информационных систем Томского государственного университета систем управления и радиоэлектроники.

Декан факультета безопасности,
канд. техн. наук. _____ Е.М. Давыдова

Доцент каф. КИБЭВС ТУСУР,
канд. техн. наук. _____ Е.Ю. Костюченко

Доцент каф. БИС ТУСУР,
канд. техн. наук. _____ О.О. Евсютин