УДК 004.934

ЭФФЕКТИВНАЯ АУГМЕНТАЦИЯ ДАННЫХ ДЛЯ ОБУЧЕНИЯ УСТОЙЧИВОЙ К ИСКАЖЕНИЯМ СИСТЕМЫ ГОЛОСОВОЙ ВЕРИФИКАЦИИ

Швец Никита Александрович, Лепендин Андрей Александрович, Карев Валентин Витальевич

Алтайский государственный университет, г. Барнаул
e-mail: lependin@phys.asu.ru

EFFICIENT DATA AUGMENTATION FOR TRAINING A VOICE VERIFICATION SYSTEM RESISTANT TO SPEECH DISTORTION

Shvets Nikita A., Lependin Andrey A., Karev Valentin V.

Altai State University, Barnaul

Аннотация: В данной работе представлена новая методика аугментации речевых данных для эффективного обучения систем голосовой верификации. Она основана на расширении набора преобразований аудиосигналов за счет добавления метода улучшения качества речи, применяемого к искаженным аудиосигналам. Тем самым обеспечивается учет всех основных способов применения современных систем верификации как с использованием предварительной обработки регистрируемого речевого сигнала, так и без таковых. Предложенная методика апробирована на голосовых записях набора VoxCeleb1, шумах и импульсных характеристиках из набора DNS Challenge 2023. В качестве нейронной сети для апробации предложенной методики использована архитектура FastResNet34. Показано, что обучение на расширенном аугментированном наборе данных с искусственно искаженными и очищенными от искажений речевыми образцами дало существенный прирост качества верификации во всех основных сценариях использования модельной системы верификации.

Ключевые слова: глубокие нейронные сети, голосовая биометрическая верификация, аугментация данных, очистка речи от шума, улучшение качества речи.

Abstract: In this paper a new method of augmentation of speech data for effective training of voice verification systems was presented. It was based on expanding the set of audio signal transformations by adding a speech quality improvement method applied to distorted audio signals. This ensures that all the main ways of using modern verification systems were taken into account, both with and without preprocessing of the recorded speech signal. The proposed technique had been tested on VoxCeleb1 voice recordings, noise and pulse characteristics from the DNS Challenge 2023 set. FastResNet34 architecture was used as a neural network for testing the proposed technique. It was shown that training on an expanded augmented data set with artificially distorted and distortion-free speech samples gave a significant increase in the quality of verification in all major scenarios of using the model verification system.

Keywords: deep neural networks, speech biometric verification, data augmentation, speech denoising, speech enhancement.

Для цитирования: Швец Н. А., Лепендин А. А., Карев В. В. Эффективная аугментация данных для обучения устойчивой к искажениям системы голосовой верификации // Проблемы правовой и технической защиты информации. 2024. №12. С.94-102.

For citation: Shvets N.A., Lependin A.A., Karev V.V. Efficient data augmentation for training a voice verification system resistant to speech distortion // Legal and Technical Problems of Information Security. 2024. No. 12. P.94-102.

Введение. В последнее время всё более популярным подходом к распознаванию личности пользователей становится голосовая биометрия, наиболее удобная из-за отсутствия необходимости непосредственного физического контакта с устройством сбора данных (микрофоном) [1]. Эта технология основана на анализе уникальных просодических, спектральных, интонационных и иных особенностей голоса человека. Широкое распространение голосовой биометрии и повышение требований к ее надежности привело к необходимости улучшении существующих и разработке новых методов. Одним из наиболее важных требований, которым должная удовлетворять современная голосовая биометрия, связана с возможностью надежного извлечения биометрических признаков при проверке подлинности субъекта в условиях улицы или помещения с различными источниками фонового шума.

Возможны две стратегии предварительной обработки искаженных голосовых сигналов. Первая связана с применением методов улучшения качества речи перед извлечением биометрических признаков. При этом требуется настройка или обучение метода предварительной обработки речевых данных для уменьшения искажений. Вторая предполагает обучение нейронной сети для извлечения признаков при верификации на искусственно расширенном (как говорят, аугментированном) наборе голосовых данных. Для этого каждый речевой образец подвергается искажению, происходит подмешивание фонового шума и воздействие искусственного эффекта реверберации. Фоновые шумы и импульсные характеристики для моделирования этих искажений обычно выбираются случайно из достаточно представительного наборов примеров. Нейронная сеть, обучаясь на искаженных примерах, учится нивелировать эффекты аугментации и вычислять устойчивые к искажениям признаки речевых сигналов.

В данной работе представлена новая методика аугментации речевых сигналов, предполагающая, что часть голосовых аудиозаписей поступает в нейросетевую модель для верификации в том виде как она была записана микрофоном, со всеми возможными искажениями, а часть подвергается предварительному воздействию алгоритма очистки от шума. Тем самым моделировалось применение разных устройств для регистрации и записи голосовых сигналов, работающих с одной единой системой голосовой верификации. Идея практической реализации предлагаемого подхода заключалась в расширении возможного множества методов аугментации речевых данных за счет применения к части искаженных примеров выбранного метода улучшения качества речи. Для контроля качества аугментированных сигналов и демонстрации того, что каждая из модификаций речевых записей вносит свой вклад при обучении нейросетевой верификационной модели использовалось вычисление распределений приближенных оценок качества речи DNSMOS P808 [2].

Аугментация речевых сигналов. Аугментации в терминологии биометрических методов – это создание на основе существующего набора данных его искаженных тем или иным способом версий. Наличие разного рода аугментаций в обучающей выборке делает обучаемую нейросетевую модель устойчивее и повышает ее обобщающую способность [1].

Первый вид аугментации аудиосигналов связан с добавлением различного рода шумов [3]. Шумом

считается любой нежелательный звуковой сигнал в окружении записывающего микрофона, включая различные естественные шумы, звуки техники, животных и голоса других людей. Подобные искажения моделируются аддитивной шумовой добавкой к основному речевому сигналу.

Второй вид аугментаций связан с мультипликативными искажениями, для которых амплитуда искажений зависит от амплитуды исходного сигнала с различными задержками в различных частотных диапазонах. Тем самым моделируется наличие реверберации или эхо в помещении, и возможное влияние расположения микрофонов при аудиозаписи [3]. Нивелирование мультипликативных искажений является более сложной задачей, чем удаление аддитивного шума по причине того, что первые связаны с исходным сигналом и их может приводить к существенным искажениям восстановленного аудиосигнала.

С математической точки зрения, внесение искусственного аддитивного шума и реверберации представляет собой следующее преобразование:

$$x(t) = y(t) + n(t) = s(t) * h(t) + n(t), \quad (1)$$

где x(t) – искаженный (аугментированный) сигнал, y(t) – сигнал с реверберацией, s(t) – чистый сигнал; n(t) – модельный аддитивный шум h(t) – импульсная характеристика модельного помещения, в котором «звучит» сигнал.

На рисунке 1 представлены примеры оригинального и аугментированных сигналов с аддитивным шумом и реверберацией. Видно, что добавление фонового шума приводит к существенному маскированию информативных участков спектра сигнала, находящихся в низких и средних частотах (рисунке 1б). Внесение мультипликативных искажений вызывает размытие «деталей» в спектре (рисунке 1в), что проявляется, в частности, в потере четко выраженных формантных полос гласных звуков.

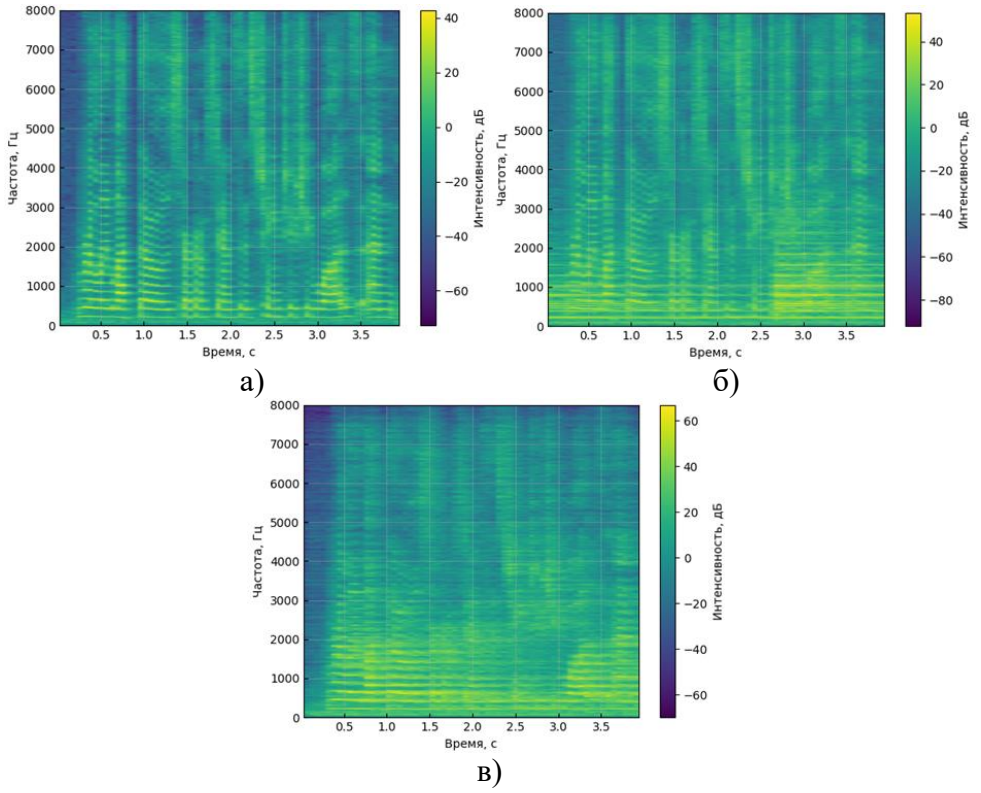

Рисунок 1. Пример спектрограмм чистой записи речевого сигнала (а), с аддитивным шумом с отношением сигнал-шум SNR = -5 дБ (б) и реверберацией (в)

Нейросетевая модель голосовой верификации. В качестве нейронной сети, применяемой для формирования признаков речевых аудиосигналов при верификации, применялась FastResNet34 (так называемая быстрая остаточная нейронная сеть с 34 слоями) [4]. Изначально эта сеть была разработана для классификации изображений, однако стала широко применяться и в других областях, в том числе при обработке спектрограмм речевых аудиозаписей [5]. Была проведена модификация данной нейронной сети, новая версия ее архитектуры представлена в табл. 1. Суть модификаций заключалась в изменении размеров выходного вектора, так как при обучении нейронной сети, это число определялось количеством различных дикторов в обучающем наборе данных. После обучения нейронной сети применение ее для верификации осуществлялось за счет вычисления вектора представления диктора, снимаемого с предпоследнего слоя сети. Данные вектора напрямую сравнивались друг с другом за счет вычисления косинусного расстояния [4] между ними, которое и отождествлялось со степенью сходства самих речевых аудиозаписей. По полученным степеням сходства для различных пороговых значений вычислялись несколько описанных ниже метрик качества верификации.

Таблица 1. Модифицированная нейронная сеть FastResNet34. Conv обозначены сверточные слои, linear – выходное линейное преобразование, применялось два вида пулинга выходов слоев (max – по максимальному значению, average – по среднему)

Название слоя	Выходной размер тензора слоя	Параметры и дополнительные преобразования в слое
conv1	112x112	7x7, 16
conv2_x	56x56	3x3 max pool
		$\begin{bmatrix}3x3, 16\\3x3, 16\end{bmatrix}$ x 3
conv3_x	28x28	$\begin{bmatrix}3x3, 32\\3x3, 32\end{bmatrix}$ x 4
conv4_x	14x14	$\begin{bmatrix}3x3, 64\\3x3, 64\end{bmatrix}$ x 6
conv5_x	7x7	$\begin{bmatrix}3x3, 128\\3x3, 128\end{bmatrix}$ x 3
linear	1x1	average pool

Метод шумоподавления. Для шумоподавления использовался алгоритм, основанный на методе «спектральных врат» [6, 7]. Он работал путём вычисления спектрограммы сигнала и оценки порога отсечки шума для каждой полосы частот этого сигнала. Затем этот порог используется для вычисления маски, которая, подавляет шум ниже порога, изменяющегося по частоте [7]. Можно выделить следующие версии алгоритма шумоподавления:

– стационарная, где значение отсечки постоянно на протяжении всего сигнала;

– нестационарная, где пороговое значение обновляется со временем.

В данной работе использовалась нестационарная версия алгоритма со стандартными параметрами (см. детали в [7]). Временная константа для вычисления минимального уровня шума составляла 2 с, диапазон частот для вычисления порога равнялся 0,5 кГц, а диапазон времени для сглаживания пороговых значений составлял 50 мс.

Аугментированный набор данных. В табл. 2 представлены сводные данные размеров наборов данных. В данной работе в качестве основного применялся открытый

набор данных VoxCeleb1 [8]. Он был собран для обучения и оценивания качества моделей идентификации и верификации дикторов. Он содержал более 150 тысяч записей 1251 диктора, взятых из видеороликов, находящихся в открытом доступе. Он в свою очередь поделен на два непересекающихся по дикторам набора: обучающий и тестовый. Обучающий набор содержит 148642 записей высказываний от 1211 дикторов из 21819 видео, в то время как тестовый 4874 высказываний от 40 дикторов из 677 видео. Аудиозаписи монофонические, частота дискретизации 16 кГц.

На основе набора данных VoxCeleb1 были созданы его аугментированные версии для обучения модифицированной нейросетевой модели FastResNet34:

– набор оригинальных сигналов, состоящий из части обучающих образцов набора VoxCeleb1 (C);

– набор сигналов, зашумленных аддитивным шумом с тремя значениями отношения сигнал/шум, составлявшими - 5дБ, 15дБ, 35 дБ) (A);

– набор сигналов с реверберацией, зашумленных аддитивным шумом (AR);

– набор сигналов с реверберацией (R);

– набор сигналов с добавлением аддитивного шума и реверберации с последующей очисткой от искажений методом «спектральных врат» (D).

Таблица 2. Сводная таблица с размерами аугментированных обучающих и тестовых данных на основе набора VoxCeleb1

Тип искажения	Тип набора данных	Количество файлов, шт.	Время, ч.
Чистый (C)	Обучающий	87360	194
	Тестовый	4874	11
Аддитивный и реверберация (AR)	Обучающий	262080	582
	Тестовый	14622	33
Очищенный (D)	Обучающий	611520	1359
	Тестовый	34118	77
Аддитивный (A)	Тестовый	14622	33
Реверберация (R)	Тестовый	4874	11

Очистка после зашумления аудиосигнала рассматривалась как дополнительный вид модельных искажений, вносимых в сигнал при работе с устройства, которое улучшает запись голоса перед передачей в систему верификации. За счет ее добавления в обучающей выборке появлялись аудиосигналы промежуточного качества, не относящиеся ни к оригинальным чистым, ни к существенно искаженным. Это подтверждалось оценкой распределений метрики качества DNSMOS P808 [2], представленных на рисунке 2. Данная метрика вычислялась с помощью предварительно обученной нейронной сети, приближающей средние оценки качества речевых сигналов, которые получались в результате субъективной оценки мнений (MOS) людьми-экспертами [9]. Оценка вычислялась на основе искаженных аудиофайлов самих по себе, без необходимости сопоставления с неискаженным оригиналом. В данной работе применялась предсказанная нейронной сетью оценка, нормированная на единичный диапазон (0 – минимальное качество, 1 – максимальное).

Таблица 3. Дисперсия и средние значения DNSMOS для оригинальных и аугментированных наборов данных

Набор данных	Нормированная метрика DNSMOS P.808	
	среднее значение	среднеквадратичное отклонение
Чистые из обучающего набора	0,705	0,063
С аддитивным шумом	0,624	0,078
С реверберацией	0,570	0,078
С аддитивным шумом и реверберацией	0,544	0,070
Очищенные от искажений	0,589	0,090

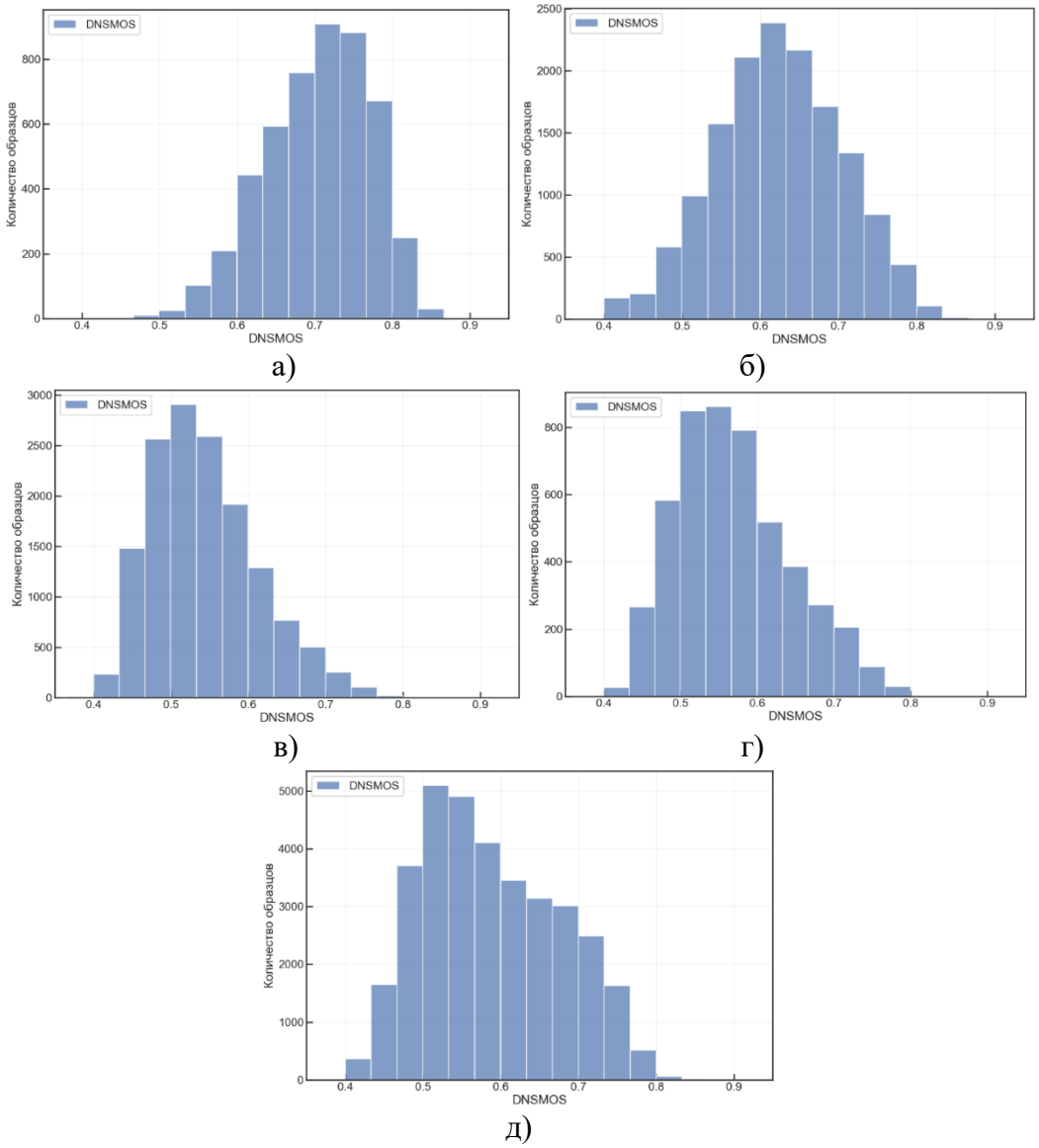

Рисунок 2. Распределение нормированных оценок всех типов аугментаций сигналов на основе набора данных VoxCeleb1: а) оригинальные сигналы набора VoxCeleb1; б) сигналы с аддитивным шумом; в) сигналы с аддитивным шумом и реверберацией; г) сигналы с реверберацией; д) очищенные от искажений сигналы

В табл. 3 для удобства приведены средние значения и среднеквадратичные отклонения нормированного качества сигналов для каждой из аугментаций и оригинального набора данных. Видно, что качество искаженных сигналов существенно смещалось относительно оригинальных, а восстановленные сигналы хоть и имели относительно низкую среднюю оценку, но отличались большим разбросом значений.

Метрики качества верификации. Обучаемая модель верификации дикторов оценивалась на тестовых аудиозаписях, аугментированных аналогично обучающему набору. При этом вычислялись метрики EER и minDCF [10]. Первая, эквивалентная частота ошибок EER традиционно используется для оценки работы биометрических систем. EER соответствует порогу принятия решения t, при котором вероятность ложного срабатывания $P_{fa}(t)$ равна вероятности пропуска самозванца $P_{miss}(t)$. Минимальное значение функции стоимости ошибок minDCF – это метрика, которая учитывает не только частоты ошибок, но и «стоимости» событий ошибочного поведения биометрической системы [10]. Она представляет собой минимальное значение функции стоимости детектирования (DCF), зависящей от порогового значения степени схожести, выше которого соответствующие аудиофрагменты считаются схожими и соответствуют одному и тому же человеку:

$$DCF(t) = C_{miss} * P_{target} * P_{miss}(t) + C_{fa} * (1 - P_{target}) * P_{fa}(t), \quad (2)$$

где C_{miss} и C_{fa} – это стоимости ошибок пропуска самозванца и ложного срабатывания, P_{target} – априорная вероятность попытки получения несанкционированного доступа. minDCF отражала баланс между различными типами ошибок и их стоимостями [10].

Результаты и обсуждение. Для оценки влияния аугментаций обучающего набора данных на качество верификации, модель была обучена на наборах с аугментациями разного рода, а именно: модель была обучена на чистом наборе данных (C), без аугментаций, на наборе данных с аугментацией аддитивного шума + реверберации (AR), а также на наборе данных с аугментацией аддитивного шума + реверберации (AR) и очищенных от шума аудиозаписей (D).

Все расчеты проводились на устройстве с процессором с видеокартой Nvidia Geforce RTX 3070. Для реализации нейронной сети и скриптов для аугментаций и обучения использовался язык программирования Python. Использовался оптимизатор параметров сети ADAM и пошаговый планировщик, уменьшающий начальную скорость обучения от 10-3 на 5% каждые 4 эпохи. Для расчета minDCF использовались следующие значения стоимостей $C_{miss} = 1$, $C_{fa} = 1$, $P_{target} = 0{,}5$. Результаты расчета EER и minDCF для моделей, обученных на разных наборах данных, и различных тестовых наборов данных представлены в табл. 4.

Из представленных результатов в табл. 4 видно, что расширение набора данных за счет аугментации положительно сказывается на метриках качества верификации при тестировании на чистых образцах. На модели, обученной на данных с добавлением образцов с аддитивным шумом и реверберацией (C+AR), на оригинальных образцах была получена оценка EER = 5,9%. При этом, когда в аугментированный набор данных включались сигналы, очищенные от шума (C+AR+D), EER уменьшалась до 5,5%. В целом можно отметить, что добавление к обучающему набору очищенных от шума сигналов привело к уменьшению ошибки на тестовых образцах со всеми возможными видами аугментации. При этом разрыв по EER между моделями, обучавшимися на наборах C+AR и C+AR+D составил от относительно малых 0,3% (для тестовых образцов с реверберацией) до значительных в сравнении с базовым уровнем ошибок 5% (для тестовых примеров, включавших в себя очищенные от шума примеры). Таким образом, можно сделать вывод о том, что при обучении моделей верификации, использующих методы улучшения качества речи (шумоочистку), критичным является наличие в обучающей выборке

соответствующих аугментированных сигналов. Это, вероятно, можно связать с тем, что образцы из набора D обладали наиболее широким разбросом оценок DNSMOS P.808 и обеспечивали улучшение обобщающей способности обучаемой модели верификации.

Таблица 4. Оценки качества работы моделей верификации, обученных и протестированных на различных наборах аугментированных данных

Обучающие данные	Оценки EER / minDCF на тестовых данных						
	C	C+AR	C+AR +D	C+AR +R+D +A	C+A	C+D	C+R
C	0,075 / 0,15	0,240 / 0,48	0,282 / 0,56	0,266 / 0,53	0,152 / 0,30	0,266 / 0,53	0,160 / 0,32
C+AR	0,059 / 0,12	0,122 / 0,24	0,170 / 0,34	0,159 / 0,32	0,110 / 0,22	0,166 / 0,33	0,076 / 0,15
C+AR+D	**0,055 / 0,11**	**0,112 / 0,22**	**0,116 / 0,23**	**0,112 / 0,22**	**0,095 / 0,19**	**0,112 / 0,19**	**0,073 / 0,15**

Заключение. Предложенное расширение набора возможных аугментаций речевых сигналов для обучения нейросетевой модели показало свою эффективность при обучении модельной системы верификации, построенной на основе модифицированной архитектуры FastResNet34. Добавление речевых образцов, очищенных от искусственно внесенных искажений, улучшило работу не только на аналогичных тестовых данных, но и уменьшило частоту появления ошибок верификации на других видах образцов, как с искажениями, так и без таковых. Методика расширения обучающей выборки, рассмотренная в данной работе, может иметь существенное практическое значение при обучении современных голосовых биометрических систем.

Исследование выполнено по Программе стратегического академического лидерства «Приоритет - 2030», проект «Разработка экспертной автоматической системы по выявлению неправомерного воздействия на аудиофайлы».

Библиографический список

1. Rituerto-Gonzalez E., Minguez-Sanchez A., Gallardo-Antolin A., Pelaez-Moreno C. Data Augmentation for Speaker Identification under Stress Conditions to Combat Gender-Based Violence // Applied sciences. 2019. Т. 9. № 11. С. 2298.

2. Reddy C., Gopal V., Cutler R. DNSMOS: A Non-Intrusive Perceptual Objective Speech Quality metric to evaluate Noise Suppressors // ICASSP 2020. Proc. IEEE 2020, Barcelona, Spain, 4-8 мая 2020.

3. Abayomi-Alli O. O., Damasevicius R., Qazi A., Adedoyin-Olowe M. Data Augmentation and Deep Learning Methods in Sound Classification: A Systematic Review // Electronics. 2022. Т. 22. № 11. С. 3795.

4. Chung J. S., Jaesung H., Seongkyu M., Lee M., Heo H. S., Choe S., Ham C., Jung S., Lee B. J., Han I. In defence of metric learning for speaker recognition // Proc. Interspeech. 2020. С. 2977-2981.

5. Kaiming H., Xiangyu Z., Shaoqing R., Jian S. Deep Residual Learning for Image Recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Proc. IEEE 2016, Las Vegas, 26 июня - 1 июля. С. 770-778.

6. Sainburg T., Thielk M., Gentner T. G. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires // PLOS Computational Biology. 2020. Т. 10. № 16.

7. timsainb/noisereduce // Github.com: сайт. URL: https://github.com/timsainb/noisereduce (дата обращения: 15.10.2024).

8. Nagrani, A., Chung, J.S., Zisserman, A. VoxCeleb: A Large-Scale Speaker Identification Dataset // Proc. Interspeech. 2017. С. 2616-2620.

9. BS.562: Subjective assessment of sound quality // itu.int: сайт. URL: https://www.itu.int/rec/ R-REC-BS.562/en (дата обращения: 15.10.2024).

10. Муртазин Р. А., Кузнецов А. Ю., Фёдоров Е. А., Гарипов И. М., Холоденина А. В., Балданова Ю. Ю., Воробьева А. А. Алгоритм выявления синтезированного голоса на основе кепстральных коэффициентов и сверточной нейронной сети // Научно-технический вестник информационных технологий, механики и оптики. 2021. Т. 21. № 4. С. 545–552.