УДК 004.056.53
DOI: 10.18469/ikt.2025.23.1.09

# К ВОПРОСУ ОПРЕДЕЛЕНИЯ ПАРАМЕТРОВ ОБУЧАЮЩЕЙ ВЫБОРКИ ДЛЯ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ В ЗАДАЧАХ ОЦЕНКИ ЗАЩИЩЕННОСТИ РЕЧЕВОЙ АКУСТИЧЕСКОЙ ИНФОРМАЦИИ С ПРИМЕНЕНИЕМ РЕЧЕВОГО ХОРА

*Волков Н.А., Иванов А.В.*
*Самарский государственный технический университет, Самара, РФ*
*E-mail: volkovnikandr@gmail.com, andrej.ivanov@corp.nstu.ru*

Для оценки защищенности речевой акустической информации предлагается использовать сверточные нейронные сети. В данной работе рассматривается подбор наиболее подходящих параметров спектрограмм и мел-частотных кепстральных коэффициентов, сгенерированных на основе аудиозаписей речи с наложенной речеподобной помехой типа «речевой хор», для формирования обучающей выборки, используемой при обучении сверточной нейронной сети. Определены ключевые параметры архитектуры сверточной нейронной сети, а также сформулированы требования к набору данных, необходимому для ее обучения. В ходе исследования варьировался один из параметров обучающей выборки с целью выявления наиболее подходящих значений. По итогам анализа установлено, что наилучшие результаты в решении данной задачи достигаются при представлении данных в виде спектрограмм. В перспективе планируется расширение набора данных за счет увеличения количества дикторов.

***Ключевые слова:*** *глубокие нейронные сети, сверточные нейронные сети, отношение сигнал/шум, зашумленность аудиозаписи, распознавание речи, спектрограммы, мел-частотные кепстральные коэффициенты, оценка защищенности речевой акустической информации, речевой хор*

## Введение

Для предотвращения утечки речевой информации по техническим каналам широко применяются активные методы защиты, включающие использование генераторов акустического и виброакустического шума. В большинстве случаев такие устройства генерируют шум, характеризующийся нормальным распределением амплитудных значений и спектральными свойствами белого шума. Однако актуальной остается задача выбора помехи, которая обеспечивает требуемый уровень защиты, оцениваемый, как правило, через коэффициент словесной разборчивости речи W, при этом минимизируя общий уровень шума, снижающий акустический дискомфорт и способствующий маскировке переговоров [1].

Исследования показали, что наиболее эффективными шумовыми помехами являются сигналы, обладающие спектральной структурой, сходной с формантными характеристиками речи. В частности, доказано, что помехи с огибающей спектра, аналогичной формантам речевого сигнала, оказываются более действенными в сравнении с равномерным белым шумом. Кроме того, данные психоакустических исследований [2] свидетельствуют о высокой эффективности речеподобных помех, которые не только воспроизводят спектральную огибающую речи, но и обладают схожей временной структурой. Наибольшую эффективность имеет так называемая помеха «речевой хор», обладающая высокой степенью маскирующего эффекта и снижающая разборчивость речи до минимально возможных значений [3].

Для оценки уровня защищенности от утечки информации по техническим каналам в помещении, которое предназначено для проведения закрытых переговоров, чаще всего используется общепринятый инструментально-расчетный подход, предложенный Хоревым А.А., Желязняком В.К. и Макаровым Ю.К. [4], основанный на результатах экспериментальных исследований, проведенных Покровским Н.Б. [5], и называемый формантным методом. При использовании общепринятого инструментально-расчетного подхода маскировка речи учитывается только с помощью помех на основе белого шума (с изменением спектра путем применения фильтров), исключая возможность оценки защищенности с помощью помех типа «речевой хор» (помеха, состоящая из смеси звуков речи [6]).

В настоящем исследовании рассматривается применение методов машинного обучения для автоматизированной оценки защищенности речевой информации при использовании помехи типа «речевой хор». В частности, применение сверточных нейронных сетей позволяет анализировать разборчивость речи и объективно оценивать степень ее подавления под воздействием маскирующего сигнала. Такой подход обеспечивает возможность адаптивного подбора параме-

тров речеподобного шума с целью максимального снижения вероятности утечки информации при минимальном влиянии на акустическое восприятие. В рамках исследования предлагается обучение сверточной нейронной сети на примере задачи классификации разборчивости речи, что является актуальным в контексте применения активных методов защиты информации.

Для построения модели формируются обучающие выборки, включающие спектрограммы и графики мел-частотных кепстральных коэффициентов (Mel-Frequency Cepstral Coefficients, MFCC), полученные на основе зашумленных аудиозаписей речи. В работах [7; 8] обоснована целесообразность применения сверточных нейронных сетей для решения данной задачи, а также детально описан процесс подготовки обучающих данных, включающий формирование наборов аудиозаписей с различными значениями отношения сигнал/шум. Применение нейросетевых методов позволяет классифицировать спектрограммы и графики MFCC по уровням разборчивости речи, что способствует автоматизации процесса оценки защищенности речевой информации. Для этого требуется разработка и верификация обучающих выборок, а также обучение нейросетевой модели, что обеспечит объективную оценку эффективности помехи типа «речевой хор».

Цель данного исследования заключается в определении наиболее подходящих параметров спектрограмм и мел-частотных кепстральных коэффициентов, полученных из зашумленных речевых хором аудиозаписей речи, с целью формирования обучающей выборки для обучения сверточной нейронной сети. Также важно выявить те виды изображений, которые больше всего способствуют развитию подхода к оценке защищенности речевой акустической информации. Для достижения цели необходимо решить следующие задачи:

1. Определить параметры аудиозаписей и сгенерированных на их основе спектрограмм и графиков MFCC, используемых при формировании обучающей выборки.

2. Выбрать архитектурные параметры сверточной нейронной сети, включая число сверточных слоев, функцию активации и метод оптимизации.

3. Провести раздельное обучение модели на наборах данных, содержащих спектрограммы, и на наборах данных, включающих графики MFCC.

4. Провести анализ полученных результатов с целью определения наиболее информативного типа данных (спектрограммы или графики MFCC), а также уточнить параметры их генерации для повышения качества оценки защищенности речевой информации.

В дальнейшем это позволит модернизировать ранее предложенную [8] интеллектуальную систему оценки защищенности речевой акустической информации по определению процента разборчивости речи на основе распознавания зашумленных аудиозаписей, представленных в виде спектрограмм или графиков мел-частотных кепстральных коэффициентов.

## Определение параметров визуализации и методов обработки аудиозаписей для обучения нейронной сети

Опираясь на работу [8] основной акцент будет сделан на оценке защищенности речевой акустической информации с использованием средств активной защиты. Для этого необходимо сформировать аудиозаписи дикторов, содержащие речеподобную помеху типа «речевой хор». Для обеспечения полноты эксперимента требуется записать 6 различных типов голосов: женские голоса с низким, средним и высоким тембрами, а также мужские голоса с аналогичными тембрами [9]. Затем необходимо сгенерировать помехи типа «речевой хор», после чего провести наложение таких помех на речь дикторов, при этом отношения «сигнал/шум» должны соответствовать заранее установленным процентам разборчивости речи. В результате будет сформирован набор данных, классифицированный по уровням разборчивости речи, который определяется субъективным методом, путем прослушивания аудиозаписей несколькими аудиторами. В рамках первоначального этапа исследования для проведения эксперимента был использован голос одного диктора – мужчины со средним тембром, на который была наложена речеподобная помеха типа «речевой хор». Для разметки данных выбраны уровни разборчивости речи от 20% до 80% с шагом в 10%. На этом наборе данных планируется обучить нейронную сеть, которая станет основой интеллектуальной системы для оценки защищенности речевой информации через определение процента разборчивости речи.

Следует подчеркнуть, что настоящее исследование фокусируется на валидации подхода к оценке разборчивости речи при фиксированных параметрах речевого сигнала. Использование записей одного диктора (мужского голоса со средним тембром) позволяет исключить вариабельность, связанную с индивидуальными особенностями артикуляции, и сосредоточиться на анализе влияния параметров речевого хора. Такой подход

согласуется с методиками, применявшимися в [7; 8] для первоначальной отладки алгоритмов. Однако, как показано в [10], окончательная оценка качества модели требует тестирования на разнообразных голосах, что предусмотрено дальнейшими этапами исследования.

Как показано в исследованиях [7; 8], сверточные нейронные сети, работающие с изображениями, показывают высокую эффективность в различных задачах распознавания. В работе [10] рассматриваются различные архитектуры таких сетей, среди которых выделяется архитектура ResNet, созданная на языке программирования Python [11]. В данном исследовании будет использована именно такая архитектура.

Поскольку сверточные нейронные сети предназначены для работы с визуальными данными, а звуковая информация часто представлена в виде изображений при обработке, для обучения такой сети необходимо подготовить набор данных, включающий визуализации зашумленных аудиозаписей речи. Существует множество методов представления акустических сигналов. Одним из наиболее распространенных подходов является метод, который основывается на создании спектрограмм и вычислении мел-частотных кепстральных коэффициентов [7].

Спектрограмма представляет собой отображение изменения спектральной плотности мощности сигнала в зависимости от времени. Для ее формирования применяется оконное преобразование Фурье, которое декомпозирует сигнал на интервалы с определенными диапазонами частот и времени [12] (рисунок 1).

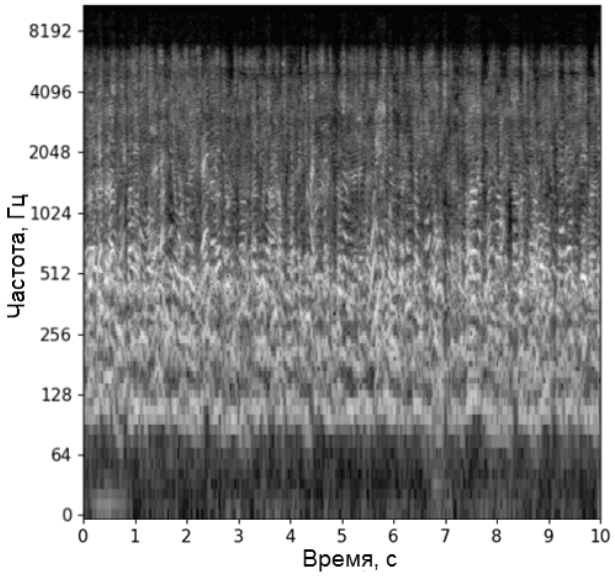


Рисунок 1. Спектрограмма аудиозаписи речи диктора длиной 10 секунд, зашумленной речевым хором с процентом разборчивости речи 20%

MFCC представляют собой набор относительно небольших значений, которые описывают частотные характеристики сигнала. Этот метод основывается на равномерном распределении частотных полос по логарифмической шкале, отражающей восприятие человеком звуков [13] (рисунок 2).

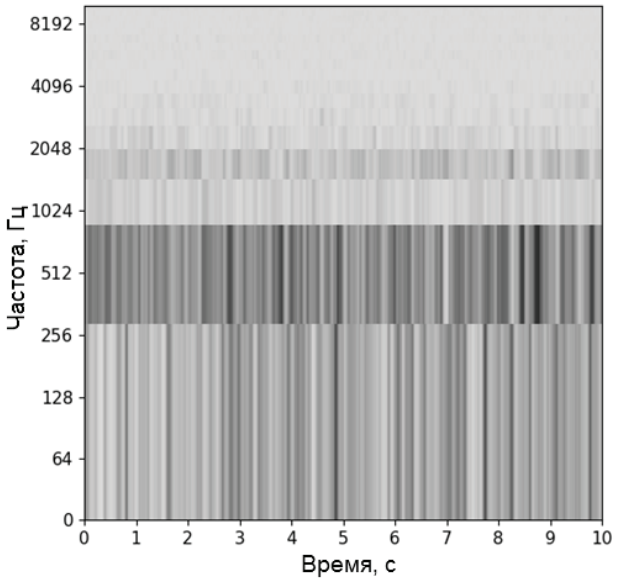


Рисунок 2. График мел-частотных кепстральных коэффициентов (MFCC) аудиозаписи речи диктора длиной 10 секунд, зашумленной речевым хором с процентом разборчивости речи 20%

Также, как и в исследовании [8], при определении наиболее подходящих параметров для обучения сверточной нейронной сети в каждом варианте обучающей выборки будет изменяться только один параметр данных в пределах заданного диапазона значений (таблица 1).

Исследования в области нейронных сетей для анализа аудиоданных [14–16] показывают, что наиболее эффективными являются аудиофайлы длительностью от 5 до 30 секунд. В данном исследовании предлагается экспериментально определить наиболее подходящую длительность аудиозаписей с шагом в 5 секунд и оценить ее влияние на обучение нейронной сети.

Не менее важным параметром является выбор окна Фурье, которое используется для анализа спектра сигнала. Различные окна, такие как прямоугольное, Хэмминга, Хеннинга и Блэкмана, обладают различными свойствами. Окно Хэмминга, обладающее хорошими частотными характеристиками, будет использоваться для генерации спектрограмм, так как оно минимизирует утечку спектра [17; 18].

Важным параметром является также длина быстрого преобразования Фурье (БПФ), опреде-

Таблица 1. Параметры обучающей выборки для обучения сверточной нейронной сети

| Параметр | Значения |
|---|---|
| Длительность аудиозаписи | от 5 до 30 секунд с шагом в 5 секунд |
| Параметры спектрального анализа | длина БПФ для спектрограмм – 1024, 2048, 4096, 8192, 16384; параметры спектральной огибающей (фреймы) для графиков MFCC – от 10 до 40 мс с шагом в 10 мс |
| Размер изображения в пикселях | от 100 х 100 пикселей до 1200 х 1200 пикселей с шагом в 100 х 100 пикселей |
| Тип масштабирования генерируемых спектрограмм и мел-частотных кепстральных коэффициентов | линейное или логарифмическое масштабирование по частоте |

ляющая частотное разрешение спектрограммы. Наиболее часто используются значения БПФ, такие как 1024, 2048, 4096, 8192 и 16384, в зависимости от желаемой точности и доступных вычислительных ресурсов [19–22]. Для мел-частотных кепстральных коэффициентов оптимальная длина фрейма варьируется от 10 до 40 мс с шагом в 10 мс, так как в этом интервале речевой сигнал остается стабильным [23].

Следующим важным параметром является размер изображения спектрограммы или графика мел-частотных кепстральных коэффициентов. Размер изображения должен учитывать баланс между детализацией и объемом данных. В исследовании предлагается использовать разрешения от 100 х 100 до 1200 х 1200 пикселей, увеличивая размер с шагом в 100 пикселей, что влияет на точность временно-частотного анализа [22].

Наконец, выбор типа масштабирования также имеет значение. Логарифмическое масштабирование по частоте используется для лучшего представления звуковых характеристик, соответствующих восприятию человеческим ухом [24]. Для MFCC применяется логарифмическая шкала на мел-частотах, что улучшает точность восприятия.

Проведенный анализ параметров данных подчеркивает их важность для формирования обучающих выборок, что обеспечит эффективное обучение сверточных нейронных сетей и повысит точность распознавания разборчивости речи в зашумленных аудиозаписях. В дальнейшем предстоит сформировать набор обучающих выборок для обучения модели.

## Формирование обучающих выборок для нейронной сети

Как было сказано ранее, обучающая выборка представляет собой записанную речь диктора с наложением речеподобной помехи типа «речевой хор» с различными отношениями сигнал/шум, соответствующими определенной градации разборчивости речи.

Формат аудиозаписи речи диктора – .WAV, так как он обеспечивает сохранение звука без потерь [12]. Также в аудиозаписи речи диктора имеются паузы между словами, присущие обычной речи.

Для контроля уровня сигналов и записи голосов без искажения частотных характеристик использовался измерительный микрофон с модулем аналого-цифрового преобразования для акустики LTR 24-2 и программное обеспечение Adobe Audition, обладающее функционалом обработки звука. В процессе записи речи дикторов частота дискретизации аудиозаписей составила 19,531 кГц [8; 25; 26].

На следующем этапе на аудиозапись речи диктора накладываются различные вариации речевого хора. Для формирования такого типа помехи были записаны 24 голоса – 12 мужских голосов и 12 женских голосов при среднем уровне речи 70 дБ [27]. В течение 1 минуты дикторы читали разные связные тексты, которые не повторялись – отрывки из рассказов А.П. Чехова. При обработке записанной речи были вырезаны все паузы, таким образом аудиозапись сократилась до 50 секунд. Так как мы ограничиваемся максимальной длительностью аудиозаписи диктора, на которую будет наложен речевой хор, в 30 секунд, то и помеха будет иметь такую же длительность. После этого полученные голоса были наложены друг на друга с некоторой задержкой по времени. Сформированные помехи были сохранены в формате .WAV.

Опираясь на работы [28; 29] для генерации речеподобных помех типа «речевой хор» были введены следующие правила:

1. Речевой хор должен состоять из 3 мужских голосов и 3 женских голосов – итого в одном речевом хоре используется 6 голосов.
2. При создании нового варианта речевого хора обязательно должны меняться хотя бы 2 голоса – 1 мужской и 1 женский.

Таким образом из 24 голосов было сгенерировано 48400 вариантов речеподобной помехи типа

«речевой хор». В данном исследовании ограничимся 1435 вариантами речевого хора. Каждый из полученных вариантов такой помехи был наложен отдельно на каждую аудиозапись речи. Наложение речевого хора на аудиозаписи речи диктора проводилось с различными значениями отношения сигнал/шум, соответствующими определенным градациям разборчивости речи. Соответствие отношения сигнал/шум к градации разборчивости речи определялось экспериментально путем прослушивания зашумленных аудиозаписей 6 аудиторами [30]. В результате такие аудиозаписи были проградуированы в процентах разборчивости речи от 20% до 80% с шагом в 10%.

Итого набор данных содержит 10045 зашумленных аудиозаписей. Формируя набор данных для сверточной нейронной сети, необходимо на основе 10045 зашумленных аудиозаписей сгенерировать спектрограммы и графики мел-частотных кепстральных коэффициентов в формате .PNG, так как алгоритм компрессии PNG 24 сжимает изображения без потери качества, нежели если использовать формат .JPEG [31]. Разделим полученный набор данных на тренировочный и тестовый, где тренировочный набор данных будет иметь 8036 изображений (80% от всего набора данных), а тестовый – 2009 изображений (20% от всего набора данных).

Последовательно изменяя параметры обучающей выборки (длительность аудиозаписи, параметры спектрального анализа, размер изображения в пикселях и тип масштабирования изображения) были сформированы 43 варианта обучающего набора данных для обучения сверточной нейронной сети: 22 варианта с использованием спектрограмм и 21 вариант с графиками мел-частотных кепстральных коэффициентов.

## Обучение сверточной нейронной сети и анализ полученных результатов

В рамках данного исследования используется архитектура сверточной нейронной сети ResNet. Параметры этой архитектуры были выбраны в соответствии с рекомендациями разработчиков ResNet, которые предложили оптимальные конфигурации для улучшения процесса обучения глубоких нейронных сетей и предотвращения проблемы исчезновения градиента [8]. Эти подходы доказали свою эффективность в задачах классификации изображений, что подтверждается их успешным применением (таблица 2) [7; 8; 10].

Параметром, позволяющим оценить успешность процесса обучения, является процент точности распознавания, который должен быть равным или превышать 90% [32–34]. Дополнительно в качестве критериев оценки будут использоваться количество эпох, затраченных на обучение, и коэффициент потерь на протяжении этого процесса.

После проведения обучения на 43 вариантах обучающей выборки следует провести анализ полученных результатов, на основании которого будут выбраны наиболее подходящие параметры для дальнейшего обучения сверточной нейронной сети. На основе этих параметров будет проведено повторное обучение, в результате чего должны быть получены две модели сверточной нейронной сети, каждая из которых будет обучена на одном из двух типов данных:

1. Архитектура сверточной нейронной сети ResNet, обученная на наборе данных, состоящих из спектрограмм, полученных из зашумленных аудиозаписей речи одного диктора, с разделением на градации разборчивости речи от 20% до 80% с шагом в 10%.

2. Архитектура сверточной нейронной сети ResNet, обученная на наборе данных, содержащих графики мел-частотных кепстральных коэффициентов, полученных на основе зашумленных аудиозаписей речи одного диктора, с разделением на градации разборчивости речи от 20% до 80% с шагом в 10%.

Результаты обучения сверточных нейронных сетей с наиболее подходящими параметрами обучающих выборок, выявленных после обучения на 43 вариантах обучающих данных, представлены в таблице 3.

Если говорить о том, какой вид набора данных лучше подходит для развития подхода к оценке защищенности речевой акустической информации, то становится очевидно, что наилучшие результаты показала модель, в которой использовались спектрограммы. Для достижения высокого процента точности распознавания нейронной сетью потребовалось 75 эпох на обучение, а коэффициент потерь составил 0,1323. При использовании обучающей выборки, состоящей из графиков мел-частотных кепстральных коэффициентов зашумленных аудиозаписей речи диктора, модель сверточной нейронной сети не смогла обеспечить достижение высокого процента точности распознавания. В таблице 3 показано, что максимальный процент точности распознавания нейронной сетью при обучении на 21 варианте обучающих выборок был достигнут только у того варианта, в котором длительность аудиозаписи составляла 20 секунд – 53,45% при коэффициенте потерь равном 1,1985. Стоит заметить, что после прохожде-

ния отметки в 140 эпох такая модель сверточной нейронной сети начала переобучаться – процент точности распознавания стал понижаться, а коэффициент потерь при обучении стал расти.

### Анализ ошибок, допущенных моделью сверточной нейронной сети при классификации

Для детального анализа работы модели была построена нормализованная матрица ошибок (рисунок 3), отражающая распределение ошибочных классификаций между всеми рассматриваемыми классами разборчивости речи (20 – 80% с шагом 10%). В данном разделе рассматривается матрица ошибок для обученной модели сверточной нейронной сети, где используются спектрограммы, так как она успешно обучилась. Как видно из матрицы ошибок, модель демонстрирует различную эффективность для разных уровней разборчивости речи. Наибольшее количество ошибок наблюдается при значениях разборчивости «20%» и «70%». Это можно объяснить плавным характером изменения спектральных характеристик речевого сигнала при постепенном увеличении значения отношения сигнал/шум. Для значения разборчивости «80%» модель демонстрирует высокую точность классификации, что обусловлено выраженными различиями в спектральной плотности мощности сигнала. Однако наблюдаются ошибки классификации, связанные с наличием артефактов в исходных аудиозаписях, неравномерным распределением энергии помехи по частотным полосам и индивидуальными особенностями диктора (тембр голоса, темп речи). Для улучшения качества классификации предлагается заострить внимание на сложных случаях (классификации разборчивости речи «20%» и «70%»), добавив им больший «вес» при обучении. Также предлагается расширить набор данных. Предложенные улучшения могут быть реализованы без принципиального изменения архитектуры модели, что сохраняет ее вычислительную эффективность.


Рисунок 3. Нормализованная матрица ошибок, отражающая распределение ошибочных классификаций между всеми рассматриваемыми классами разборчивости речи

### Заключение

В рамках данного исследования были изучены различные параметры, применяемые для формирования обучающих выборок на основе аудиозаписей речи дикторов, зашумленных речеподобной помехой типа «речевой хор». Результаты показали, что использование графиков мел-частотных кепстральных коэффициентов в качестве входных данных для сверточной нейронной сети не позволяет достичь более быстрого и точного

Таблица 2. Параметры используемой архитектуры сверточной нейронной сети ResNet

| Элемент архитектуры | Параметры |
|---|---|
| Используемые слои | 15 сверточных слоев, 2 полносвязных слоя, выходной слой |
| Сверточные слои | 64 фильтра размером 7 x 7 с шагом 2 пикселя для уменьшения размерности изображения |
| Операция максимальной подвыборки | Пул размером 3 x 3 с шагом 2 пикселя для снижения размерности |
| Остаточные блоки | 2 сверточных слоя с 128 фильтрами размером 3 x 3 с шагом 2 пикселя и с 256 фильтрами размером 3 x 3 с шагом 2 пикселя с последующей объединенной связью |
| Операция средней подвыборки | Пул размером 7 x 7 для уменьшения пространственных размерностей |
| Полносвязные слои | Каждый слой содержит 64 нейрона |
| Функция активации | Нелинейная |
| Оптимизационный алгоритм | Стохастический градиентный спуск |

Таблица 3. Результаты обучения сверточных нейронных сетей с наиболее подходящими параметрами обучающих выборок

| Обучающая выборка, состоящая из спектрограмм зашумленных аудиозаписей | | | | | | |
|---|---|---|---|---|---|---|
| Длительность аудиозаписи, с | Длина БПФ, точек | Размер изображения, пикселей | Тип масштабирования изображения | Количество эпох на обучение | Процент точности распознавания нейронной сетью, % | Коэффициент потерь при обучении |
| 15 | 2048 | 600 x 600 | Логарифмическое масштабирование по частоте | 75 | 96,13 | 0,1323 |
| Обучающая выборка, состоящая из графиков MFCC зашумленных аудиозаписей | | | | | | |
| Длительность аудиозаписи, с | Длина фрейма, мс | Размер изображения, пикселей | Тип масштабирования изображения | Количество эпох на обучение | Процент точности распознавания нейронной сетью, % | Коэффициент потерь при обучении |
| 20 | 10 | 600 x 600 | Логарифмическое масштабирование по частоте | 140 | 53,45 | 1,1985 |

обучения по сравнению со спектрограммами. В частности, модель, обученная на мел-частотных кепстральных коэффициентах, продемонстрировала низкую точность распознавания (53,45%) за такое количество эпох, чтобы не произошло переобучение модели (140). Минимальный коэффициент потерь был достигнут в 1,1985, после чего происходит переобучение. Обучение на спектрограммах требовало значительно меньшего числа эпох (75) и имело более высокий коэффициент потерь (0,1323). Процент точности распознавания нейронной сетью составил 96,13%. Результаты подтверждают, что правильный выбор параметров обучающих выборок существенно влияет на успешность обучения модели, а предложенный подход, основанный на анализе зашумленных речевым хором аудиозаписей, может быть использован для оценки защищенности речевой акустической информации.

В дальнейшем предлагается увеличить количество дикторов – записать мужские голоса с низким и высоким тембрами, а также записать женские голоса с низким, средним и высоким тембрами и пройти все этапы, указанные в работе. Таким образом, появляется возможность модернизации созданной интеллектуальной системы оценки защищенности речевой акустической информации, основанной на обученной сверточной нейронной сети на спектрограммах зашумленных речевым хором аудиозаписях речи.

## Литература


1. Трушин В.А., Иванов А.В. Возможности снижения интегрального уровня помехи в средствах активной защиты информации речевой информации (состояние и перспективы) // Доклады ТУСУР. 2018. Т. 21. № 2. С. 38–42.
2. Алдошина И., Притц Р. Музыкальная акустика. СПб.: Композитор, 2006. 720 с.
3. Авдеев В.Б., Трушин В.А., Кунгуров М.А. Унифицированная речеподобная помеха для средств активной защиты речевой информации // Труды СПИИРАН. 2020. Т. 19, № 5. С. 991–1017.
4. Железняк В.К., Макаров Ю.К., Хорев А.А. Некоторые методические подходы к оценке эффективности защиты речевой информации // Специальная техника. 2000. № 4. С. 39–45.
5. Покровский Н.Б. Расчет и измерение разборчивости речи. М.: Связьиздат, 1962. 391 с.
6. Макаров Ю.К., Хорев А.А. К оценке эффективности защиты акустической (речевой) информации // Специальная техника. 2000. № 5. С. 46–56.
7. Волков Н.А., Иванов А.В. Подход к оценке защищенности речевой акустической информации с применением нейронных сетей // Вестник СибГУТИ. 2024. Т. 18, № 2. С. 43–56. DOI: 10.55648/1998-6920-2024-18-2-43-56
8. Волков Н.А., Иванов А.В. Определение параметров обучающей выборки для нейронной сети, применяемой при определении процента разборчивости речи в задачах оценки защищенности речевой акустической информации // Вопросы защиты информации. 2024. № 4 (147). С. 44–51. DOI: 10.52190/2073-2600_2024_4_44
9. Жабыко Е.И., Рублевская Н.И. Акустическое

проектирование залов многоцелевого назначения: учебное пособие. Владивосток: Изд-во ДВГТУ, 2008. 89 с.
10. Volkov N.A., Ivanov A.V. On the use of convolutional neural networks in the tasks of assessing the security of speech acoustic information // AISMA-2024: International Workshop on Advanced in Information Security Management and Applications. Stavropol, 2024. P. 320–327.
11. Deep residual learning for image recognition. URL: https://arxiv.org/pdf/1512.03385.pdf/ (дата обращения: 08.01.2025).
12. Умняшкин С.В. Основы теории цифровой обработки сигналов: учебное пособие. М.: Техносфера, 2016. 528 с.
13. Герасимов С.М., Жаринов О.О. Исследование методов анализа речевых сигналов // Семьдесят третья международная студенческая научная конференция ГУАП: материалы конференции. СПб.: ГУАП, 2020. Ч. 3. С. 36–41.
14. Муромцев Д.И., Шилин И.А., Исаев И.В. Структурирование, разметка и обогащение данных. СПб.: Университет ИТМО, 2024. 72 с.
15. Novoselov S., Volokhov V., Lavrentyeva G. Universal speaker recognition encoders for different speech segments duration // ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Rhodes Island, Greece, 2023. DOI: 10.1109/ICASSP49357.2023.10096081
16. Анализ аудиоданных с помощью глубокого обучения и Python (часть 1). URL: https://nuancesprog.ru/p/6713/ (дата обращения: 08.01.2025).
17. Христофоров А.В. Методы анализа спектра сигнала: учебно-методическое пособие. Казань, 2004. 21 с.
18. Функции оконного сглаживания // DSPLIB.org. URL: https://ru.dsplib.org/content/windows/windows.html/ (дата обращения: 08.01.2025).
19. Лукин А. Введение в цифровую обработку сигналов (математические основы): учебно-методическое пособие. М.: МГУ, Лаборатория компьютерной графики и мультимедиа, 2007. 54 с.
20. Adavanne S., Virtanen T. A report on sound event detection with different binaural features. URL: https://www.researchgate.net/publication/320297754_A_report_on_sound_event_detection_with_different_binaural_features/ (дата обращения: 09.01.2025).
21. Guide to audio classification using deep learning – analytics vidhya. URL: https://www.analyticsvidhya.com/blog/2022/04/guide-to-audio-classification-using-deep-learning/ (дата обращения: 08.02.2025).
22. Осипов О.В. Итерационные алгоритмы БПФ с высоким частотным разрешением // Вычислительные методы и программирование. 2021. Т. 22, № 2. С. 121–134.
23. Первичный анализ речевых сигналов. URL: https://alphacephei.com/ru/lecture1.pdf/ (дата обращения: 08.01.2025).
24. Ахмад Х.М., Жирков В.Ф. Введение в цифровую обработку речевых сигналов: учебное пособие. Владимир: Владимирский государственный университет, 2007.192 с.
25. Многоканальные системы сбора данных. LTR24. Руководство программиста. URL: https://www.lcard.ru/download/ltr24api.pdf/ (дата обращения: 09.01.2025).
26. Тюрин А. Схема простой звуковой карты (ЦАП) // Радио Лоцман. URL: https://www.rlocman.ru/shem/schematics.html?di=660639/ (дата обращения: 09.01.2025).
27. Исследование воздействия речеподобной помехи на психоэмоциональное состояние человека / В.А. Трушин [и др.] // Динамика систем, механизмов и машин. 2020. Т. 8, № 2. С. 138–144. DOI: 10.25206/2310-9793-8-2-138-144
28. Зельманский О.Б. Методика синтеза речеподобных сигналов на разных языках для систем защиты информации // Информационные системы и технологии. 2012. № 4. С. 122–131.
29. Воробьев В.И., Давыдов А.Г., Давыдов Г.В. Речеподобные сигналы: разновидности, основные параметры, способы формирования, области применения // Доклады Белорусского государственного университета информатики и радиоэлектроники. 2009. № 3. С. 9–16.
30. ГОСТ Р 50840-95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. М.: Госстандарт России, 1995. 249 с
31. Волков Н.А., Иванов А.В. К вопросу оценки защищенности речевой акустической информации с применением сверточной нейронной сети // Перспектива-2023: материалы X Всероссийской молодежной школы семинара по проблемам информационной безопасности. Красноярск, 2023. М.: Издательский дом Академии Естествознания, 2023. С. 40–45. DOI: 10.17513/np.579
32. Паршин С.Е. Исследование параметров алгоритмов распознавания лиц // Сборник научных трудов Новосибирского государственного технического университета. 2019. № 1 (94). С. 55–70. DOI: 10.17212/2307-6879-2019-1-55-70

33. Simple MNIST convnet // KERAS. URL: https://keras.io/examples/vision/mnist_convnet/ (дата обращения: 10.01.2025).
34. Гафаров Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения: учебное пособие. Казань: Изд-во Казанского университета, 2018. 121 с.

*Дата поступления 27.01.2025*
*Дата принятия 27.02.2025*

**Волков Никита Андреевич,** аспирант кафедры электронных систем и информационной безопасности (ЭСИБ) Самарского государственного технического университета (СамГТУ). 443100, Российская Федерация, г. Самара, ул. Молодогвардейская, 244. Тел. +7 846 337-31-96. E-mail: volkovnikandr@gmail.com

**Иванов Андрей Валерьевич,** к.т.н., доцент кафедры ЭСИБ СамГТУ. 443100, Российская Федерация, г. Самара, ул. Молодогвардейская, 244. Тел. +7 846 337-31-96. E-mail: andrej.ivanov@corp.nstu.ru

UDC 004.056.53
DOI: 10.18469/ikt.2025.23.1.09

# THE ISSUE OF DETERMINING THE PARAMETERS OF A TRAINING SAMPLE FOR A NEURAL NETWORK TRAINING TO SOLVE TASKS OF ASSESSING THE SECURITY OF SPEECH ACOUSTIC INFORMATION USING A SPEECH CHOIR

*Volkov N.A., Ivanov A.V.*
*Samara State Technical University, Samara, Russian Federation*
*E-mail: volkovnikandr@gmail.com*

In order to assess the speech acoustic information security, it is proposed to use convolutional neural networks. This article considers the selection of the most appropriate parameters of spectrograms and mel-frequency cepstral coefficients generated on the basis of audio recordings of speech with superimposed speech-like interference of the «speech chorus» type to get a training sample used in the convolutional neural network training process. Key parameters of the convolutional neural network architecture, as well as the requirements for the data set necessary for its training, are defined. During the study, one of the parameters of the training sample was varied in order to identify the most appropriate values. Based on the results of the analysis, it was found that the best results in solving this problem are achieved when data is presented in the form of spectrograms. In future, it is planned to expand the data set by increasing the number of speakers.

***Keywords****: deep neural networks, convolutional neural networks, signal-to-noise ratio, audio recording noise, speech recognition, spectrograms, mel-frequency cepstral coefficients, assessment of the security of speech acoustic information, speech chorus*

**Volkov Nikita Andreevich,** Samara State Technical University, 244, Molodogvardeyskaya Street, Samara, 443100, Russian Federation; PhD Student of Electronic Systems and Information Security Department. Tel. +7 846 337-31-96. E-mail: volkovnikandr@gmail.com

**Ivanov Andrey Valerievich,** Samara State Technical University, 244, Molodogvardeyskaya Street, Samara, 443100, Russian Federation; Associated Professor of Electronic Systems and Information Security Department, PhD in Technical Science. Tel. +7 846 337-31-96. E-mail: andrej.ivanov@corp.nstu.ru

**References**

1. Trushin V.A., Ivanov A.V. Possibilities and outlooks of integral noise level decrease in voice information active protection means. *Doklady TUSUR*, 2018, vol. 21, no. 2. pp. 38–42. (In Russ.)
2. Aldoshina I., Pritts R. *Musical Acoustics.* Saint Petersburg: Kompozitor. 2006, 720 p. (In Russ.)

3. Avdeev V.B., Trushin V.A., Kungurov M.A. Unified speech-like interference for active protection of speech information. *Trudy SPIIRAN*, 2020, vol. 19, no. 5, pp. 991–1017. (In Russ.)
4. Zheleznyak V.K., Makarov Yu.K., Khorev A.A. Some methodological approaches to assessing the effectiveness of speech information protection. *Special'naya tehnika*, 2000, no. 4, pp. 39–45. (In Russ.)
5. Pokrovsky N.B. *Calculation and Measurement of Speech Intelligibility.* Moscow: Svyazizdat, 1962, 391 p. (In Russ.)
6. Makarov Yu.K., Khorev A.A. Assessment of the effectiveness of acoustic (speech) information protection. *Special'naya tehnika*, 2000, no. 5, pp. 46–56. (In Russ.)
7. Volkov N.A., Ivanov A.V. An approach to assessing the security of speech acoustic information using neural networks. *Vestnik SibGUTI*, 2024, vol. 18, no. 2. pp. 43–56. (In Russ.)
8. Volkov N.A., Ivanov A.V. Determination of the parameters of a training sample for a neural network used to determine the percentage of speech intelligibility in tasks of assessing the security of speech acoustic information. *Voprosy zaschity informacii*, 2024, no. 4 (147), pp. 44–51. DOI: 10.52190/2073-2600_2024_4_44 (In Russ.)
9. Zhabyko E.I., Rublevskaya N.I. *Acoustic* B*esign of Multipurpose* P*alls:* Textbook. Vladivostok: Izd-vo DVGTU, 2008, 89 p. (In Russ.)
10. Volkov N.A., Ivanov A.V. On the use of convolutional neural networks in the tasks of assessing the security of speech acoustic information. *AISMA-2024: International Workshop on Advanced in Information Security Management and Applications*. Stavropol, 2024, pp. 320–327.
11. Deep residual learning for image recognition. URL: https://arxiv.org/pdf/1512.03385.pdf/ (accessed: 08.01.2025).
12. Umnyashkin S.V. *Fundamentals of the Theory of Digital Signal Processing:* Textbook. Moscow: Technosfera, 2016, 528 p. (In Russ.)
13. Gerasimov S.M., Zharinov O.O. Research of methods of speech signal analysis. *Sem'desyat tret'ya mezhdunarodnaya studencheskaya nauchnaya konferenciya GUAP: materialy konferencii*. Saint Petersburg: GUAP, 2020, vol. 3, pp. 36–41. (In Russ.)
14. Muromtsev D.I., Shilin I.A., Isaev I.V. *Structuring, Markup, and Data Enrichment.* Saint Petersburg: Universitet ITMO, 2024, 72 p. (In Russ.)
15. Novoselov S., Volokhov V., Lavrentyeva G. Universal speaker recognition encoders for different speech segments duration. *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*). Rhodes Island, Greece, 2023. DOI: 10.1109/ICASSP49357.2023.10096081
16. Audio data analysis using deep learning and Python (part 1). URL: https://https://nuancesprog.ru/p/6713/ (accessed: 08.01.2025). (In Russ.)
17. Khristoforov A.V. *Methods of Signal Spectrum Analysis:* Textbook. Kazan, 2004, 21 p. (In Russ.)
18. Window smoothing functions. *DSPLIB.org*. URL: https://ru.dsplib.org/content/windows/windows.html/ (accessed: 08.01.2025). (In Russ.)
19. Lukin A. *Introduction to Digital Signal Processing (mathematical foundations):* Textbook. Moscow: MGU, Laboratoriya komp'yuternoj grafiki i mul'timedia, 2007, 54 p. (In Russ.)
20. Adavanne S., Virtanen T. A report on sound event detection with different binaural features. URL: https://www.researchgate.net/publication/320297754_A_report_on_sound_event_detection_with_different_binaural_features/ (accessed: 09.01.2025).
21. Guide to audio classification using deep learning – analytics vidhya. URL: https://www.analyticsvidhya.com/blog/2022/04/guide-to-audio-classification-using-deep-learning/ (accessed: 08.02.2025).
22. Osipov O.V. Iterative FFT-algorithms with high frequency resolution. *Vychislitel'nye metody i programmirovanie*, 2021, vol. 22, no. 2, pp. 121–134. (In Russ.)
23. Primary analysis of speech signals. URL: https://alphacephei.com/ru/lecture1.pdf/ (accessed: 08.01.2025).

24. Akhmad Kh.M., Zhirkov V.F. *Introduction to Digital Speech Signal Processing:* Textbook. Vladimir: Vladimirskij gosudarstvennyj universitet, 2007, 192 p. (In Russ.)
25. Multi-channel data collection systems. LTR24. Programmer's guide. URL: https://www.lcard.ru/download/ltr24api.pdf/ (accessed: 09.01.2025) (In Russ.)
26. Tyurin A. The scheme of a simple sound card (DAC). *Radio Locman.* URL: https://www.rlocman.ru/shem/schematics.html?di=660639/ (accessed: 09.02.2025). (In Russ.)
27. Trushin V.A. et al. Method of classes assessment the «loyal employee» and «insider». *Dinamika sistem, mekhanizmov i mashin,* 2020, vol. 8, no. 2, pp. 138–144. DOI: 10.25206/2310-9793-8-2-138-144 (In Russ.)
28. Zelmansky O.B. Method of speechlike signals synthesis in different languages for information protection systems. *Informacionnye sistemy i tekhnologii*, 2012, no. 4, pp. 122–131. (In Russ.)
29. Vorobyov V.I., Davydov A.G., Davydov G.V. Speech-like signals: varieties, basic parameters, methods of formation, areas of application. *Doklady Belorusskogo gosudarstvennogo universiteta informatiki i radioelektroniki*, 2009, no. 3, pp. 9–16. (In Russ.)
30. GOST R 50840 – 95. Speech transmission over communication paths. Methods for assessing quality, legibility, and recognition. Moscow: Gosstandart Rossii, 1995, 249 p. (In Russ.)
31. Volkov N.A., Ivanov A.V. On the issue of assessing the security of speech acoustic information using a convolutional neural network. *Perspektiva-2023: materialy X Vserossijskoj molodezhnoj shkoly seminara po problemam informacionnoj bezopasnosti.* Krasnoyarsk, 2023. Moscow: Izdatelskiy dom Akademii Estestvoznaniya, 2023, pp. 40–45. DOI: 10.17513/np.579 (In Russ.)
32. Parshin S.E. Investigation of the parameters of face recognition algorithms. *Sbornik nauchnyh trudov Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta*, 2019, no. 1 (94), pp. 55–70. DOI: 10.17212/2307-6879-2019-1-55-70 (In Russ.)
33. Simple MNIST convnet. *KERAS*. URL: https://keras.io/examples/vision/mnist_convnet/ (accessed: 10.01.2025). (In Russ.)
34. Gafarov F.M., Galimyanov A.F. *Artificial Neural Networks and Applications: Textbook.* Kazan: Izd-vo Kazanskogo universiteta, 2018, 121 p. (In Russ.)

*Received 27.01.2025*
*Accepted 27.02.2025*