УДК 004.934.2
Шифр специальности ВАК: 2.3.1

А.А. Шелупанов, Е.Ю. Костюченко, Г.Л. Погорелов, Н.Д. Никитин, А.В. Максимов

Оценка качества синтеза речи с применением попарного сравнения и сиамской нейронной сети

Актуальность. Развитие технологий систем синтеза речи по образцу обусловливает необходимость определения надежных методов оценки их реалистичности для обеспечения информационной безопасности. ***Цель исследования.*** Сравнительное исследование реалистичности моделей синтеза речи (RVC, Coqui TTS и Fish Speech) на основе комплексного подхода, сочетающего субъективное попарное сравнение аудиторами и объективную классификацию сиамской нейронной сети. ***Методы.*** Применены попарное сравнение аудиозаписей группой независимых аудиторов, а также сиамская нейронная сеть с обучаемым классификатором на основе MFCC-спектрограмм, обученная на авторском русскоязычном наборе данных. ***Научная новизна.*** Для задачи определения синтезированной речи предложена модифицированная архитектура сиамской нейронной сети, в которой вместо фиксированной метрики расстояния используется конкатенация эмбеддингов с последующей обработкой многослойным персептроном. ***Результаты.*** Результаты исследования показали, что модель Fish Speech обладает наивысшим акустическим сходством с оригиналом, вызывая максимальное число ложных идентификаций. Обученная сиамская сеть выявляет синтезированную речь с итоговой точностью 84%. ***Практическая значимость.*** Полученные выводы доказывают применимость предложенной архитектуры для независимой автоматизированной оценки качества синтеза и позволяют продолжить исследование в направлении защиты речевых систем от спуффинг-атак с возможностью сравнения с перцептивными способностями человека.

Ключевые слова: синтез речи, нейронные сети, попарное сравнение, экспертная оценка, Retrieval-based Voice Conversion, Text-to-Speech, Fish Speech, сиамская нейронная сеть.
DOI: 10.21293/1818-0442-2026-29-1-144-151

Введение

В рамках данного исследования оценка качества синтеза речи концептуально связывается с понятием перцептивной неотличимости генерированного сигнала от естественного голоса диктора. Несмотря на то, что формально в процессе экспертного аудирования и работы сиамской нейронной сети решается задача двухальтернативной классификации («реальное или синтезированное аудио»), результаты решения этой задачи напрямую транслируются в оценку качества речи. Способность генеративной модели вводить в заблуждение валидатора (человека или алгоритм), выражающаяся в высоком количестве ложных классификаций, служит наиболее строгим индикатором высокого акустического сходства. Таким образом, классификационный подход де-факто выступает инвертированной шкалой оценки: невозможность достоверного разделения выборки на классы верифицирует минимизацию машинных артефактов и подтверждает высокую реалистичность синтеза.

Для акустического анализа структуры речевого сигнала разрабатываются программные комплексы, способные оценивать базовые частотные-временные характеристики [1] и применяются государственные стандарты, например, ГОСТ Р 50840–95 «Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости» [2], требующие участия экспертов-аудиторов [3]. Метод попарного сравнения [4] может быть эффективен для поиска артефактов синтеза речи, однако данный процесс является субъективным [5] и трудоемким, что требует его автоматизации. Предложенные алгоритмы на базе математических метрик (DTW-расстояние [6] и анализ кепстральных коэффициентов [7]) в текущих условиях уступают нейросетевым архитектурам [8], которые сегодня становятся базовым стандартом в задачах обработки звука, включая автоматическое распознавание речевых команд [9].

Целью исследования является сравнительный анализ реалистичности синтеза речи по образцу с применением различных моделей: RVC [10], Coqui TTS [11] и Fish-Speech [12] путем оценки точности двухальтернативной классификации оригинальных и сгенерированных аудиозаписей.

В рамках данного исследования качество синтеза определяется через способность отличить сгенерированное аудио от реального. Решаемая задача классификации де-факто переходит в оценку качества: если аудиторы и нейронная сеть массово не могут разделить записи на классы (оригинал и синтез), это является прямым доказательством высокой реалистичности синтезированной речи.

Авторский набор данных

Такие наборы данных, как MLAAD (Multi-language Audio Anti-spoofing Dataset) или ASVspoof (отсутствуют русскоязычные записи), ориентированы на обобщенные задачи антиспуфинга. ASVspoof не содержат строго выровненных пар аудиозаписей (оригинальное и синтезированное произношение одной и той же фразы). MLAAD такие пары содержит, однако есть вопросы к речевому составу с точки зрения русского языка – необходимо дополнительное исследование по части фонетической взвешенности. Поскольку методология данного исследования требует проведения попарного сравнения (pairwise comparison) для детального анализа качества генера-

ции конкретными моделями (Fish Speech, RVC, Coqui TTS), использование готовых не размеченных по фразам баз оказалось нецелесообразным. В связи с этим был сформирован собственный, структурно сбалансированный набор данных, обеспечивающий прямое сопоставление оригинала и синтеза.

В качестве основы для набора данных были использованы фразы из ГОСТ Р 50840–95, обеспечивающие фонетически сбалансированный набор фонем русского языка.

Авторский набор данных был сформирован из аудиозаписей реальной речи, записанной диктором, и синтезированной с использованием трех моделей.

Были выбраны три диктора, записавшие по 75 аудиозаписей, а именно 25 исходных фраз из ГОСТа были произнесены три раза с целью соответствия количества реальных и синтезированных. Синтезированные аудиозаписи были сформированы следующим образом: Fish-Speech и Coqui TTS использовали фразу из ГОСТа и образец голоса диктора, а RVC использовал аудиозаписи, синтезированные предобученным русскоязычным диктором мужского пола из Coqui TTS, и образец голоса диктора.

Таким образом, был сформирован набор данных, состоящий из 450 аудиозаписей реальной и синтезированной речи.

Предобработка данных

Аудиозаписи синтезированной речи обладают признаками, позволяющими аудиторам определить их происхождение – характерные паузы, а реальная речь обладает фоновым шумом, представляющим возможность выявить тип записи.

Таким образом, были применены следующие этапы обработки всех аудиозаписей: шумоочистка и удаление пауз.

На языке программирования Python версии 3.10 была разработана программа, где модуль шумоочистки был выполнен при помощи библиотеки librosa [13], и модуль удаления пауз реализован с использованием VAD (Voice activity detection) [14] библиотеки webrtcvad (Google WebRTC).

Процесс удаления шумов в аудиозаписях включал в себя следующее: спектральный анализ (STFT): входной сигнал $x(t)$ преобразуется в частотно-временную область с помощью оконного преобразования Фурье, где используются окно Ханна с 2 048 и шагом перекрытия 512 и оценка профиля шума, а алгоритм автоматически анализирует первые 0,5 с аудиозаписи (допущение о 0,5 с тишины основано на специфике используемых записей, имеющих стартовую паузу, при попадании речи в этот интервал возможно искажение полезного сигнала на этапе шумоочистки, поэтому для автоматического анализа произвольной речи данный алгоритм потребует доработки с использованием детекции голосовой активности. С другой стороны, предполагается применение оценки в рамках реальной системы, где запись ведется для отдельных фраз, выводимых на экран. В этом случае даже минимальное время реакции диктора на выведенную для зачитывания фразу превышает 0,5 с, что приводит к появлению необходимой паузы), предполагая отсутствие речи на этом участке, на основе этого фрагмента строится статистическая модель шума, расчет маски – для каждого частотно-временного бина рассчитывается отношение сигнал/шум, если амплитуда сигнала в конкретной частотной полосе не превышает порог, заданный профилем шума (с учетом коэффициента чувствительности), данный бин подавляется, обратное преобразование – модифицированная спектрограмма преобразуется обратно во временную область для получения очищенного сигнала.

Процесс удаления пауз состоит из данных шагов: ресемплинг, где сигнал приводится к формату, требуемому стандартом G.711/G.722, а частота дискретизации 16 кГц, моно, глубина 16 бит. Фрейминг с разбиением аудиопотока на кадры длительностью 30 мс. Классификация фреймов: вероятностная модель на основе гауссовых смесей (GMM) классифицирует каждый кадр как «речь» – «1» или «тишина» – «0». Использовался уровень агрессивности алгоритма, равный 2, для баланса между пропуском тихой речи и ложным срабатыванием на шум. Тримминг: на основе полученной карты активности производилась обрезка начальных и конечных сегментов тишины.

Проведена обработка имен аудиозаписей, чтобы облегчить их анализ после проведения попарного сравнения. Имя аудиозаписи после обработки состояло из имени диктора, происхождения (реальное или конкретная система синтеза) и порядкового номера записи.

Попарное сравнение

Разработана программа для проведения попарного сравнения на языке Python. Программа организована следующим образом: инициализация и идентификация – ввод имени аудитора и выбор целевого диктора из списка доступных, загрузка и подготовка данных – программа находит соответствующие директории и для выбранного диктора. Формирует две независимые очереди образцов аудиозаписей, рандомизация – сначала программа выполняет случайное перемешивание файлов.

В рамках каждой итерации с использованием генератора случайных чисел определяется порядок воспроизведения внутри пары: реальный или синтезированный будет проигрываться первым, цикл аудирования и управление воспроизведением – для каждой пары программа последовательно активирует аудиоканалы с возможностью дополнительного прослушивания пары аудиозаписей, сбор и фиксация результатов – после каждого выбора аудитора программа сопоставляет полученный результат с метаданными файла – его исходным типом и записываются в структуру DateFrame и дублируется в файл формата «.xlsx» с данными полями: название записи, ее происхождение и ответ пользователя.

В субъективном эксперименте приняли участие пять аудиторов, заранее не знакомых с темой проводимой работы, что исключало влияние возможных знаний о качестве синтезаторов.

Протокол проведения попарного сравнения

1. Формирование группы: для проведения оценки была привлечена группа из 5 независимых респондентов, не знакомых с тематикой и целями проводимого исследования.

2. Подготовка условий: респонденты самостоятельно определяли место и время проведения эксперимента для обеспечения комфортных условий прослушивания и минимизации внешних шумов.

3. Предоставление материалов: после подтверждения готовности каждому участнику направлялся архив, содержащий исполняемые файлы, и набор аудиоданных для попарного сравнения.

4. Развертывание окружения: респонденты производили распаковку архива и запуск конфигурационного скрипта для установки необходимых программных зависимостей.

5. Регистрация: при запуске программы оценки участники проходили процедуру регистрации, вводя свои имена.

6. Выбор объекта оценки: респондент выбирал в интерфейсе программы конкретного диктора, аудиозаписи которого будут предъявляться для оценки.

7. Инициация сессии: запускалась процедура попарного сравнения для выбранного диктора.

8. Перемешивание набора аудиозаписей: программа в случайном порядке формировала пары аудиозаписей, состоящие из реальной и синтезированной речи. Записи в паре воспроизводились последовательно.

9. Процедура оценки: задача респондента заключалась в определении оригинальной записи из представленной пары (двухальтернативный выбор). В интерфейсе была предусмотрена функция повторного прослушивания записей.

10. Регистрация результатов: каждый выбор, сделанный респондентом, логировался и экспортировался в электронную таблицу формата .xlsx.

11. Итеративность: после завершения оценки всех пар аудиозаписей одного диктора респондент переходил к аналогичной процедуре для следующего диктора.

12. Сбор данных: по завершении всего эксперимента полученные файлы с результатами передавались исследователям для проведения последующего анализа.

Анализ попарного сравнения

Процедура постобработки и анализа собранных данных включала в себя следующие этапы:

Агрегация данных: результаты попарного сравнения, полученные от всех респондентов в формате электронных таблиц, были сведены в единую базу данных для последующего статистического анализа.

Расчет общей точности: на основе объединенной выборки была вычислена общая точность идентификации участниками оригинальных аудиозаписей дикторов.

Атрибуция ошибок: каждый случай неверной классификации (выбор синтезированного аудиофрагмента вместо оригинального) был промаркирован тегом соответствующей модели синтеза речи.

Формирование сводной статистики: на основе размеченных данных была составлена итоговая матрица ошибок. Данная матрица демонстрирует распределение количества ошибок двумерно: по каждому респонденту индивидуально и по каждой генеративной модели. Было рассчитано суммарное количество ложных срабатываний для каждой архитектуры, что выступает прямым индикатором качества и естественности генерируемого ею голоса (табл. 1).

Анализ сводной матрицы ошибок позволил ранжировать исследуемые модели по степени реалистичности синтезируемой речи. В контексте данного эксперимента критерием качества работы модели выступало количество ложных идентификаций – ситуаций, когда респонденты ошибочно принимали синтезированный аудиофрагмент за естественную речь диктора.

Для статистической проверки результатов субъективного прослушивания и оценки различий между алгоритмами применялся непараметрический критерий Фридмана. Выбор данного метода обусловлен самим экспериментом, при котором синтезированная речь всех трех моделей (RVC, Fish Speech, Coqui TTS) оценивалась одной и той же группой из 5 аудиторов (зависимые выборки). Анализ распределения количества ошибок выявил статистически значимые различия между архитектурами ($\chi^2 = 8{,}4$; $df = 2$; $p = 0{,}015$). Поскольку полученный уровень значимости $p < 0{,}05$, нулевая гипотеза об отсутствии разницы отвергается. Это позволяет утверждать, что наблюдаемые различия в качестве синтеза не являются следствием случайности в рамках проведения эксперимента.

Таблица 1

Количество ошибок, допущенных аудиторами для дикторов

Первый диктор			
	RVC	Fish	TTS
Аудитор 1	0	1	0
Аудитор 2	12	13	4
Аудитор 3	1	11	0
Аудитор 4	1	7	0
Аудитор 5	0	6	1
Итог	14	36	5
Второй диктор			
Аудитор 1	0	6	1
Аудитор 2	8	12	5
Аудитор 3	1	13	5
Аудитор 4	1	5	0
Аудитор 5	0	8	0
Итог	10	44	11
Третий диктор			
Аудитор 1	9	5	4
Аудитор 2	8	10	5
Аудитор 3	5	5	1
Аудитор 4	3	4	6
Аудитор 5	6	8	2
Итог	31	32	18

Модель Fish Speech статистически значимо превосходит остальные алгоритмы по способности генерировать естественную речь, вводящую аудиторов в заблуждение.

Наивысшее качество синтеза продемонстрировала архитектура Fish Speech. При оценке аудиозаписей, сгенерированных данной моделью, слушатели допустили наибольшее количество ошибок (112 ложных срабатываний). Это свидетельствует о высокой степени акустической и перцептивной схожести синтезируемого сигнала с оригинальным голосом, что затрудняет их дифференциацию человеком.

Второе место по уровню естественности заняла модель RVC. Суммарно для данного алгоритма было зафиксировано 55 ошибок классификации. Этот показатель указывает на приемлемое, хотя и менее совершенное по сравнению с лидером качество генерации речи, при котором артефакты синтеза становятся более заметными для человеческого слуха.

Наименее убедительные результаты показала система Coqui TTS, для которой было допущено всего 34 ошибки. В подавляющем большинстве случаев при попарном сравнении респонденты успешно отличали естественную речь от сгенерированной. Столь низкий показатель ложных идентификаций прямо указывает на наличие выраженных машинных искажений и недостаточную реалистичность аудиозаписей данной модели.

Метод попарного сравнения позволил получить субъективную оценку качества синтеза речи. Однако для формирования независимой объективной оценки в работе будет использована нейросетевая модель.

Описание модели

В качестве модели была использована модифицированная архитектура сиамской нейронной сети (Siamese Neural Network). Предложенная модель предназначена для попарного сравнения аудиозаписей, представленных в виде двумерных частотно-временных признаков (MFCC-спектрограмм).

Архитектурно сеть можно разделить на два основных модуля: экстрактор признаков (сверточная часть) и блок классификации (многослойный персептрон).

1. Экстрактор признаков (Feature Extractor).

Извлечение признаков осуществляется параллельно для двух входных аудиозаписей через идентичные сверточные ветви с общими (разделяемыми) весами. Сверточный блок состоит из трех последовательных каскадов. Каждый каскад включает в себя:

– слой двумерной свертки (Conv2D) с ядром 3×3, параметром padding=1 и постепенным увеличением числа каналов (16, 32 и 64 соответственно);

– функцию нелинейной активации ReLU;

– слой максимального пулинга (MaxPooling2D) с окном 2×2.

В результате прохождения через сверточные каскады пространственная размерность входных спектрограмм (количество коэффициентов MFCC и длина аудиозаписи) уменьшается в 8 раз по каждой из осей. На выходе экстрактора формируются сглаженные одномерные векторы признаков (эмбеддинги) для каждой из двух аудиозаписей.

2. Блок классификации (Classifier).

Отличительной особенностью предложенной архитектуры является метод сравнения полученных представлений. В отличие от классических сиамских сетей, где для определения схожести эмбеддингов применяются заранее заданные метрики (например, вычисление евклидова расстояния или косинусного сходства), в данной работе используется обучаемый классификатор на базе многослойного персептрона (MLP).

Векторы признаков, полученные от обеих аудиозаписей, объединяются методом конкатенации в единый вектор размерности 2N (где N – размерность одного эмбеддинга). Объединенный вектор подается на вход полносвязной нейронной сети, состоящей из:

– первого скрытого слоя на 256 нейронов (активация ReLU);

– слоя прореживания (Dropout) с вероятностью 0,2, применяемого для предотвращения переобучения модели;

– второго скрытого слоя на 64 нейрона (активация ReLU);

– выходного слоя с двумя нейронами, формирующими логиты для двухальтернативной классификации.

Преимущество данного подхода – замена метрического расстояния (евклидова) на конкатенацию с последующей обработкой через персептрон – позволяет модели не просто оценивать геометрическую близость векторов в пространстве признаков, но и самостоятельно выучивать сложные нелинейные корреляции между акустическими характеристиками оригинальной и синтезированной речи, что повышает общую выразительную способность (expressive power) модели.

Набор данных для обучения модели

Для обучения сиамской нейронной сети был сформирован специализированный набор данных, базирующийся на том же массиве аудиозаписей, который ранее применялся в ходе попарного тестирования. На этапе тренировки модели использовались аудиозаписи двух дикторов. Общий объем подготовленного материала составил 150 сформированных пар. Данный массив был разделен на обучающую (Train) и валидационную (Validation) выборки в соотношении 80/20: в обучающее множество вошло 120 пар, а для промежуточной оценки и настройки нейронной сети (валидации) было выделено 30 пар.

Для дальнейшего обучения классификатора были выбраны мел-кепстральные коэффициенты (MFCC). Данный выбор был обусловлен возможностью извлечения значимых акустических характеристик, что оказывает позитивное влияние на обучение классификатора [15].

Анализ результата работы классификатора

Обучение проводилось при помощи трехэтапной кроссвалидации: всего три диктора, обучение проводилось на двух, а на третьем (независимом) проводи-

лось практическое тестирование. Таким образом было проведено три обучения (первый и второй с третьим, первый и третий со вторым, второй и третий тестировался с первым). На рис. 1 показаны результаты обучения на наборе данных первого и второго диктора).

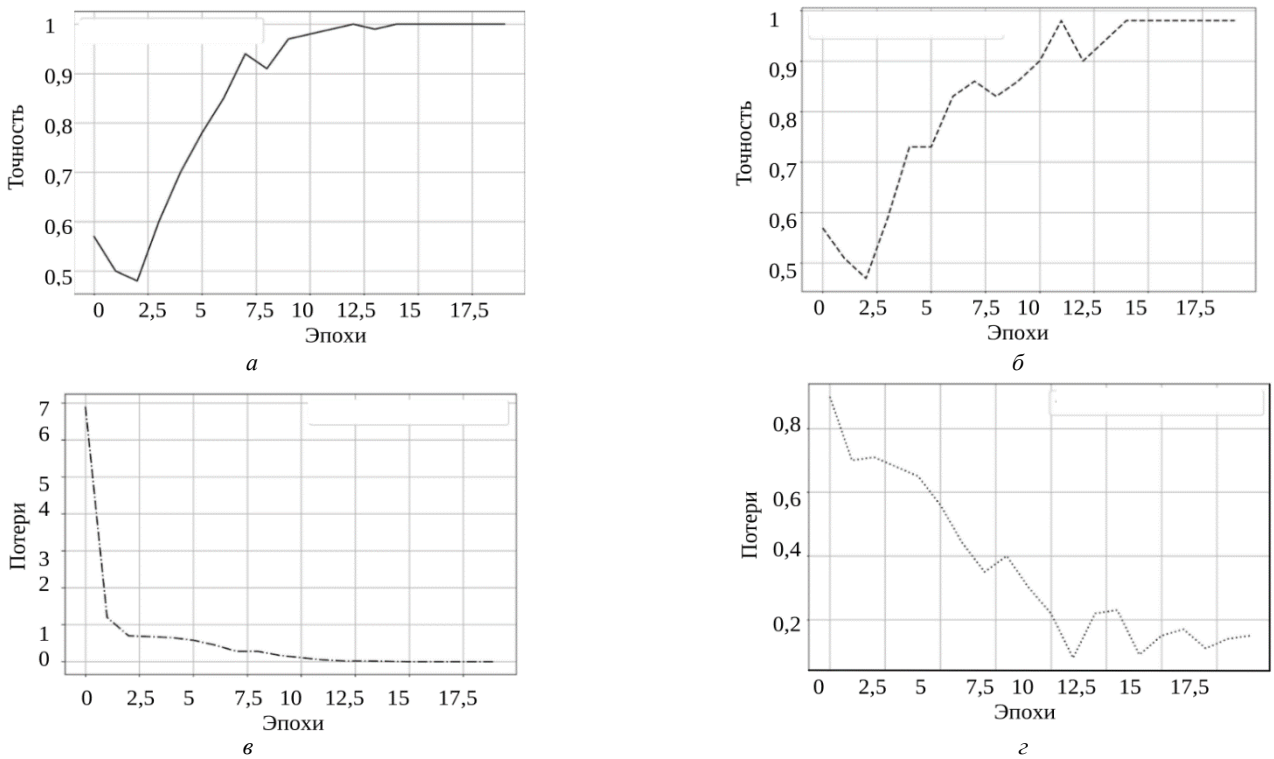
По результатам практического тестирования предложенная архитектура продемонстрировала высокую точность: нейронная сеть корректно классифицировала 189 из 225 предъявленных пар, допустив 36 ошибок (ложных срабатываний). Итоговая точность (Accuracy) на тестовой выборке составила 84,00%, что подтверждает способность модели выявлять паттерны синтезированной речи на ранее не встречавшихся аудиозаписях.

Для оценки сходимости алгоритма и контроля процесса обучения анализировалась динамика изменения функции потерь (loss-функция – кроссэнтропия), представленная на рис. 1, *в*, *г*. В контексте данной задачи метрика потерь отражает общую величину ошибки модели: насколько сильно предсказанные вероятности отклоняются от истинных меток классов (естественная или синтезированная речь). Постепенное снижение тренировочных потерь (рис. 1, *в*) свидетельствует о том, что нейронная сеть успешно выявляет релевантные акустические признаки. В свою очередь, синхронное снижение валидационных потерь (рис. 1, *г*) подтверждает хорошую обобщающую способность классификатора и отсутствие эффекта переобучения (overfitting) на тренировочных данных.

Распределение 36 ложных классификаций на тестовой выборке показало следующую динамику: наибольшую сложность представили аудиозаписи, синтезированные при помощи модели Fish Speech (31 ошибка классификации). Аудиозаписи, синтезированные посредством Coqui TTS, спровоцировали 4 ложных срабатывания системы. Наиболее легко распознаваемыми для сиамской сети оказались аудиозаписи системы RVC – в данном сегменте зафиксирована лишь 1 ошибка (табл. 2).

Сравнение с алгоритмом DTW (усреднение через метрику RMS по 225 парам MFCC-спектрограмм) показало следующие результаты: Fish Speech – 167,16; Coqui TTS – 189,33; RVC – 194,21. DTW подтверждает акустическое лидерство Fish Speech, модель Coqui TTS получила оценку выше RVC. Это прямо противоречит результатам субъективного аудирования.

Данное расхождение эмпирически показывает, что математические метрики не улавливают перцептивные машинные артефакты, с выявлением которых успешно справляется разработанная сиамская нейросеть. Кроме того, приведенный в работе подход на основе классификации не учитывает степень схожести образцов. Для нас если есть один удачный экземпляр синтеза и один крайне неудачный – это критерий качества 1/2. С точки зрения же расстояния это более половины большей метрики, которая может быть огромной для крайне неудачного экземпляра, первый результат, по сути, будет не учтен. Это говорит об ограниченной применимости метрики DTW в чистом виде – она показывает среднее расстояние между экземплярами, но при этом усредняет и выбросы, что может качественно влиять на итоговую метрику.



а *б* *в* *г*

Рис. 1. Динамика процесса обучения сиамской нейронной сети на выборке: *а* – обучающей; *б* – валидационной. Значение функции потерь на выборке: *в* – обучающей; *г* – валидационной

Таблица 2

Количество ошибок, допущенных сиамской нейронной сетью

Количество определений	RVC	Fish	TTS
Правильных	74	44	71
Ложных	1	31	4

Заключение

Субъективная оценка, реализованная методом попарного сравнения с участием группы респондентов, показала, что наибольшую сложность для человеческого восприятия представляют аудиозаписи, сгенерированные моделью Fish Speech (112 ошибок классификации). Модели RVC и Coqui TTS распознавались слушателями значительно легче: при их прослушивании было допущено 55 и 34 ошибки соответственно. Данные показатели свидетельствуют о том, что синтез Fish Speech обладает наибольшим акустическим сходством с естественной речью.

Для проведения объективной оценки была разработана сиамская нейронная сеть, классифицирующая аудиозаписи на основе извлечения и конкатенации MFCC-спектрограмм. Тестирование алгоритма на независимой выборке (записи третьего диктора, не использовавшегося при обучении) показало итоговую точность распознавания 84,00%. Из 225 предъявленных тестовых пар модель верно классифицировала 189.

Сравнение результатов субъективного и объективного подходов выявило прямое совпадение распределения ошибок. Анализ 36 ложных срабатываний нейронной сети показал, что большая их часть (31 ошибок) также пришлась на аудиозаписи модели Fish Speech. Для моделей Coqui TTS и RVC алгоритм допустил 4 и 1 ошибку соответственно.

Таким образом, результаты машинной классификации полностью согласуются с результатами субъективного аудирования. Установлено, что частотно-временные характеристики записей, сгенерированных архитектурой Fish Speech, представляют наибольшую сложность для дифференциации от оригинального голоса как для человеческого слуха, так и для обученной нейронной сети.

Работа выполнена при поддержке Министерства науки и высшего образования России: FEWM-2026-0009 (ТУСУР).

Литература

1. Репьюк Н.С. Программный комплекс для исследования речевых сигналов / Н.С. Репьюк, А.А. Конев // Доклады ТУСУР. – 2025. – Т. 28, № 1. – С. 93–99.

2. ГОСТ Р 50840–95. Передача речи по трактам связи. Методы оценки качества, разборчивости и узнаваемости. – М.: Изд-во стандартов, 1996. – 18 с.

3. Kostuchenko E. The evaluation process automation of phrase and word intelligibility using speech recognition systems / E. Kostuchenko et al. // International Conference on Speech and Computer. – Istanbul, Turkey, Cham: Springer, 2019. – P. 237–246.

4. Семенчук К.В. Сущность метода попарного сравнения в функционально-стоимостном анализе / К.В. Семенчук, П.Е. Минкевич, А.А. Пузыревская // Science Time. – 2016. – № 12. – С. 107–109.

5. Оценка качества синтезированной речи: проблемы и решения / А.И. Соломенник, А.О. Таланов, М.В. Соломенник и др. // Известия вузов. Приборостроение. – 2013. – Т. 56, № 2. – С. 38–41.

6. Новохрестова Д.И. Алгоритм и методика количественной оценки схожести речевых сигналов / Д.И. Новохрестова, Е.Ю. Костюченко, И.А. Ходашинский // Доклады ТУСУР. – 2022. – Т. 25, № 3. – С. 45–51.

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

8. A Survey of Audio Classification Using Deep Learning / K. Zaman, M. Sah, C. Direkoglu, M. Unoki // IEEE Access. – 2023. – Vol. 11. – P. 106620–106649.

9. Катаев М.Ю. Методика распознавания речевых команд в школьных информационных системах // Речевые технологии. – 2024. – № 1. – С. 18–34.

10. Оганесян О. Сравнение алгоритмов клонирования голоса в условиях нулевого и малого количества примеров / О. Оганесян, Д. Саргсян, А. Маладжян // Труды ИСП РАН. – 2024. – Т. 36, № 4. – С. 7–16.

11. Coqui TTS Documentation: XTTS [Электронный ресурс]. – URL: https://coqui-tts.readthedocs.io/en/latest/models/xtts.html (дата обращения: 15.01.2025).

12. Fish-Speech: Leveraging Large Language Models for Advanced Multilingual Text-to-Speech Synthesis / S. Liao, Y. Wang, T. Li et al. // arXiv e-prints. – 2024. – Art. No. arXiv:2411.01156.

13. Librosa: Audio and music signal analysis in Python / B. McFee, C. Raffel, D. Liang et al. // Proc. of the 14th Python in Science Conf. (SciPy). – Austin, TX, USA. – 2015. – Vol. 8. – P. 18–25.

14. Voice activity detection algorithm based on long-term pitch information / X.K. Yang, L. He, Q. Dan, W.Q. Zhang // EURASIP Journal on Audio, Speech, and Music Processing. – 2016. – No. 1. – P. 1–9.

15. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques / M.K. Gourisaria et al. // Discover Internet of Things. – 2024. – Vol. 4, No. 1. – P. 1–23.

Шелупанов Александр Александрович
Д-р техн. наук, член-кор. Российской академии наук,
президент Томского государственного ун-та
систем управления и радиоэлектроники (ТУСУР)
Ленина пр-т, 40, г. Томск, Россия, 634050
ORCID: 0000-0003-2393-6701
Тел.: +7 (382-2) 90-71-55
Эл. почта: president@tusur.ru

Костюченко Евгений Юрьевич
Канд. техн. наук, доцент каф. комплексной
информационной безопасности электронно-вычислительных систем (КИБЭВС) ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
ORCID: 0000-0001-8000-2716
Тел.: +7 (382-2) 70-15-29
Эл. почта: key@fb.tusur.ru

Погорелов Григорий Леонидович
Студент каф. КИБЭВС ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: +7-996-636-60-88
Эл. почта: grenki70@gmail.com

Никитин Николай Дмитриевич
Студент каф. КИБЭВС ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: +7-902-763-76-06
Эл. почта: kolya.nikitin.2005@inbox.ru

Максимов Артём Владимирович
Студент каф. КИБЭВС ТУСУРа
Ленина пр-т, 40, г. Томск, Россия, 634050
Тел.: +7-903-993-81-13
Эл. почта: amaximov027@gmail.com

Поступила в редакцию: 19.03.2026.
Принята к публикации: 21.05.2026.

Shelupanov A.A., Kostyuchenko E.Y., Pogorelov G.L.,
Nikitin N.D., Maksimov A.V.

Speech Synthesis Quality Assessment Using Pairwise Comparison and Siamese Neural Networks

Relevance. The development of speech synthesis technologies necessitates the need to establish reliable methods for assessing their realism to ensure information security. ***Purpose.*** A comparative study of the realism of speech synthesis models (RVC, Coqui TTS, and Fish Speech) based on an integrated approach combining subjective pairwise comparison by auditors and objective classification of a Siamese neural network. ***Methods.*** Pairwise comparison of audio recordings by a group of independent auditors was used, as well as a Siamese neural network featuring a trainable classifier based on MFCC spectrograms, trained on authors' original Russian-language dataset. ***Novelty.*** For the problem of identifying synthesized speech, a modified Siamese neural network architecture is proposed, in which a concatenation of embeddings followed by processing via a multilayer perceptron is used instead of a fixed distance metric. ***Results.*** The results of the study showed that the Fish Speech model exhibits the highest acoustic similarity to the original, causing the maximum number of false identifications. The trained Siamese network detects synthesized speech with an overall accuracy of 84%. ***Practical significance.*** The findings demonstrate the applicability of the proposed architecture for independent automated evaluation of synthesis quality and enable further research into protecting speech systems from spoofing attacks, with the possibility of comparison with human perceptual abilities.

Keywords: speech synthesis, neural networks, pairwise comparison, expert evaluation, Retrieval-based Voice Conversion, Text-to-Speech, Fish Speech, Siamese neural network.
Funding. This work was supported by the Ministry of Science and Higher Education of Russia: FEWM-2026-0009 (TUSUR).
DOI: 10.21293/1818-0442-2026-29-1-144-151

References

1. Repyuk N.S., Konev A.A. [Software package for the study of speech signals]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radioelektroniki*, 2025, vol. 28, no. 1, pp. 93–99 (in Russ.).

2. GOST R 50840–95. *Peredacha rechi po traktam svyazi. Metody otsenki kachestva, razborchivosti i uznavayemosti* [Speech transmission over varies communication channels. Techniques for measurements of speech quality, intelligibility and voice identification]. Moscow, Standartinform Publ., 1996, 18 p. (in Russ.).

3. Kostuchenko E. et al. The evaluation process automation of phrase and word intelligibility using speech recognition systems. International Conference on Speech and Computer, Istanbul, Turkey, Cham, Springer Publ., 2019, pp. 237–246.

4. Semenchuk K.V., Minkevich P.E., Puzyrevskaya A.A. [The essence of the pairwise comparison method in functional-cost analysis]. *Science Time*, 2016, no. 12 (36), pp. 107–109 (in Russ.).

5. Solomennik A.I., Talanov A.O., Solomennik M.V., Khomitsevich O.G., Chistikov P.G. [Evaluating the quality of synthesized speech: problems and solutions]. *Journal of Instrument Engineering*, 2013, vol. 56, no. 2, pp. 38–41 (in Russ.).

6. Novokhrestova D.I., Kostyuchenko E.Y., Khoda-shinsky I.A. [Algorithm and methodology for quantitative assessment of speech signal similarity]. *Doklady Tomskogo gosudarstvennogo universiteta sistem upravleniya i radio-elektroniki*, 2022, vol. 25, no. 3, pp. 45–51 (in Russ.).

7. Sorokin V.N., Vyugin V.V., Tananykin A.A. [Personal voice recognition: an analytical review]. *Information Processes*, 2012, vol. 12, no. 1, pp. 1–30 (in Russ.).

8. Zaman K., Sah M., Direkoglu C., Unoki M. A Survey of Audio Classification Using Deep Learning. *IEEE Access*, 2023, vol. 11, pp. 106620–106649.

9. Kataev M.Yu. [Methods of speech command recognition in school information systems]. *Speech Technologies*, 2024, no. 1, pp. 18–34 (in Russ.).

10. Oganesyan O., Sargsyan D., Malajyan A. [Comparison of voice cloning algorithms in zero-shot and few-shot scenarios]. *Proceedings of ISP RAS*, 2024, vol. 36, no. 4, pp. 7–16 (in Russ.).

11. Coqui TTS Documentation: XTTS. Available at: https://coqui-tts.readthedocs.io/en/latest/models/xtts.html (accessed: 15 January 2025).

12. Liao S., Wang Y., Li T. et al. Fish-Speech: Leveraging Large Language Models for Advanced Multilingual Text-to-Speech Synthesis. *arXiv e-prints*, 2024, Art. no. arXiv: 2411.01156.

13. McFee B., Raffel C., Liang D. et al. Librosa: Audio and music signal analysis in Python. Proc. of the 14th Python in Science Conf. (SciPy), Austin, TX, USA, 2015, vol. 8, pp. 18–25.

14. Yang X.K., He L., Dan Q., Zhang W.Q. Voice activity detection algorithm based on long-term pitch information. *EURASIP Journal on Audio, Speech, and Music Processing*, 2016, no. 1, pp. 1–9.

15. Gourisaria M.K. et al. Comparative analysis of audio classification with MFCC and STFT features using machine learning techniques. *Discover Internet of Things*, 2024, vol. 4, no. 1, pp. 1–23.

Alexander A. Shelupanov
Doctor of Engineering Sciences, Corresponding Member of the Russian Academy of Sciences,
President of Tomsk State University of Control Systems and Radioelectronics (TUSUR)
40 Lenin Ave., Tomsk, 634050, Russia
ORCID: 0000-0003-2393-6701
Phone: +7 (382-2) 90-71-55
Email: president@tusur.ru

Evgeny Y. Kostyuchenko
Candidate of Engineering Sciences, Associate Professor,
Department of Complex Information Security, TUSUR
40, Lenin Ave., Tomsk, Russia, 634050
ORCID: 0000-0001-8000-2716
Phone: +7 (382-2) 70-15-29
Email: key@fb.tusur.ru

Grigorii L. Pogorelov
Student, Department of Complex Information Security, TUSUR
40, Lenin Ave., Tomsk, Russia, 634050
Phone: +7-996-636-60-88
Email: grenki70@gmail.com

Artem V. Maximov
Student, Department of Complex Information Security, TUSUR
40, Lenin Ave., Tomsk, Russia, 634050
Phone: +7-903-993-81-13
Email: amaximov027@gmail.com

Nikolay D. Nikitin
Student, Department of Complex Information Security, TUSUR
40, Lenin Ave., Tomsk, Russia, 634050
Phone: +7-902-763-76-06
Email: kolya.nikitin.2005@inbox.ru

Received: 19.03.2026.
Accepted: 21.05.2026.