УДК 004.056.57 : 004.934

ФОНЕТИЧЕСКИЕ АПОСТЕРИОГРАММЫ ДЛЯ ОЦЕНКИ РАЗБОРЧИВОСТИ РЕЧИ ПРИ АВТОМАТИЧЕСКОЙ ВЕРИФИКАЦИИ ДИКТОРОВ

Лепендин Андрей Александрович, Зубков Павел Андреевич, Карев Валентин Витальевич

Алтайский государственный университет, Барнаул
lependin@phys.asu.ru, pav.zubkoff@mail.ru, krv.valentin@gmail.com

PHONETIC POSTERIORGRAMS FOR SPEECH INTELLIGIBILITY ASSESSMENT IN AUTOMATIC SPEAKER VERIFICATION

Lependin Andrey A., Zubkov Pavel A., Karev Valentin V.

Altai State University, Barnaul
lependin@phys.asu.ru, pav.zubkoff@mail.ru, krv.valentin@gmail.com

Аннотация. В работе предложен новый подход к оценке искажений, вносимых в речевой сигнал. Он основан на применении предварительно обученной нейросетевой модели для вычисления фонетических апостериограмм и оценки их отклонения от образцовых с помощью дивергенции Йенсена-Шеннона. Для вычисления апостериограмм применялась обученная на наборе данных Common Voice 21 модель High-Fidelity Neural Phonetic Posteriorgrams. На основе тестового подмножества набора данных VoxCeleb1 были сформированы три множества речевых записей с фоновым шумом контролируемой мощности, нелинейными искажениями и реверберацией. Вычислена дивергенция фонетических апостериограмм и проведена параллельная оценка качества верификации дикторов нейросетевой моделью с временными задержками TDNN. Показано, что дивергенция Йенсена-Шеннона обладает высокой чувствительностью к рассматриваемым искажениям речевых сигналов, хорошо коррелирует с частотой эквивалентных ошибок речевой верификации. Она может быть эффективно применена как для оценки качества речевых записей при биометрической

Abstract. In this paper a new approach to assessing distortions of speech signals was proposed. It was based on the use of a pretrained neural network model for calculating phonetic posteriorgrams and assessing their deviation from reference values using the Jensen-Shannon divergence. A High-Fidelity Neural Phonetic Posteriorgrams model trained on the Common Voice 21 dataset was used to calculate the posteriorgrams. Three sets of speech recordings with background noise of controlled power, nonlinear distortions, and reverberation were generated using a test subset of the VoxCeleb1 dataset. The divergence of the phonetic posteriorgrams was calculated, and a parallel assessment of the speaker verification quality was conducted using a TDNN model. The Jensen-Shannon divergence was shown to be highly sensitive to the considered speech signal distortions and correlates well with the equivalent error rate of speech verification. It can be effectively applied both to assess the quality of speech recordings during biometric verification of users and as a loss function in training new neural network methods for speech processing.

верификации пользователей, так и в качестве функции потерь при обучении новых нейросетевых методов обработки речи.

Ключевые слова: глубокие нейронные сети, голосовая биометрическая верификация, фонетическая апостериограмма, разборчивость речи

Keywords: deep neural networks, speech biometric verification, phonetic pos-teriograms, speech intelligibility

Для цитирования: *Лепендин А.А., Зубков П.А., Карев В.В.* Фонетические апостериограммы для оценки разборчивости речи при автоматической верификации дикторов // Проблемы правовой и технической защиты информации. 2025. № 13. С. 30–41.

For citation: *Lependin A.A., Zubkov P.A., Karev V.V.* Phonetic posteriorgrams for speech intelligibility assessment in automatic speaker verification. *Legal and Technical Problems of Information Security*. 2025. No. 13. P. 30–41.

Введение

В современных условиях голосовая биометрическая верификация становится все популярнее благодаря ее удобству и доступности. Однако ее широкое практическое применение сталкивается с серьезным препятствием, связанным с высокой чувствительностью к качеству речевого аудиосигнала. Фоновые шумы, реверберация из-за отражений звука от стен помещений, нелинейные искажения в аналоговом или цифровом компоненте аудиотракта могут резко снизить точность верификации. Существующие методы улучшения качества речи не решают проблему полностью, поскольку могут изменять те уникальные характеристики голоса, на которых строится верификация. В связи с этим одним из актуальных направлений исследований становится разработка новых подходов, целенаправленно повышающих помехоустойчивость биометрических систем. Одним из таких перспективных путей является интеграция метрик разборчивости речи непосредственно в процесс обучения моделей, что позволит создавать более надежные решения для эксплуатации в условиях реального мира.

В данной работе предложен новый подход к оценке искажений, вносимых в речевой аудиосигнал, который в перспективе может использоваться как для оценивания качества фонетической разборчивости речевых сигналов, так и для работы в качестве сложной перцептивной функции потерь при обучении систем обработки речи. Он основан на применении предварительно обученной нейросетевой модели для вычисления распределений вероятностей произносимых диктором фонем для каждого короткого временного отрезка (фрейма). Основная идея заключается в том, что чем чище речь, тем больше «уверенность» модели в конкретном звуке, произносимом в данный момент, что соответствует распределению вероятностей с одним выраженным пиком на одной предсказываемой фонеме. Если сигнал искажен, то распределение вероятностей становится более «размытым» и существенно отличается от образцового однопикового. Оценить степень отклонения распределений можно с помощью расстояния того или иного рода. В этой работе в качестве такового предложено использовать симметричную дивергенцию Йенсена-Шеннона.

Вычисление фонетических апостериограмм

Выделение фонетических признаков из речевого сигнала представляет собой ключевой этап во многих методах автоматической обработки речи, поскольку позволяет связать акустические особенности звукового сигнала с фонетической структурой языка. В отличие от символьных или графемных представлений, которые оперируют буквами или слогами, фонетические признаки фокусируются на акустически значимых свойствах звуков,

что особенно важно для обработки вариаций произношения, акцентов и шумовых условий.

Развитие современных методов нейросетевой обработки речевых сигналов привело к тому, что одним из наиболее перспективных методов извлечения фонетических признаков стало применение нейронных сетей для вычисления фонетических апостериограмм (англ. phonetic posteriorgram или PPG) — оценок распределений вероятности фонем во фреймах речевого сигнала [1]. Формально апостериограмму можно представить как матрицу $\hat{P} \in R^{T \times N}$, где T — количество фреймов, а N — число фонем в языке и/или фонетическом алфавите:

$$\hat{P}(i) = P(\text{фонема i во фрейме t}). \quad (1)$$

В данной работе применялась нейросетевая модель High-Fidelity Neural Phonetic Posteriorgrams [2]. В качестве набора фонем, распределение вероятностей на котором предсказывалось, выступало подмножество международного фонетического алфавита (IPA, International Phonetic Alphabet) [3] состоявшее из 94 звуков английского языка. Архитектура модели в упрощенной форме представлена на рисунке 1. Модель начинала обработку с преобразования мел-спектрограмм речевого аудиосигнала размером B×80×T (где B — размер батча обучающих примеров, T — количество фреймов в каждом примере) с помощью 1-D сверточного слоя (Conv1d) с ядром размера 5. Этот слой извлекал локальные спектральные паттерны, важные для акустической классификации, и преобразовывал входные данные в скрытое представление размера B×256×T. Далее предобработанные спектрограммы передавались в кодировщик, состоящий из пяти слоев трансформера. Каждый слой использовал механизм внимания с двумя головами и позиционные эмбеддинги, что позволяло модели учитывать долгосрочные зависимости в речевом сигнале при сохранении размерности B×256×T. На финальном этапе выходной сверточный слой с ядром размера 5 снижал размер представления сигнала до B×94×T, где 94 результирующих значения (логита распределения) каждого фрейма соответствовало числу фонем. Для вычисления распределения вероятностей фонем $\hat{P}_t$ применялась функция softmax:

$$\hat{P}_t = \text{softmax}(\hat{y}_t), \quad (2)$$

где $\hat{y}_t$ — не нормированный 94-мерный вектор логитов временного шага t.

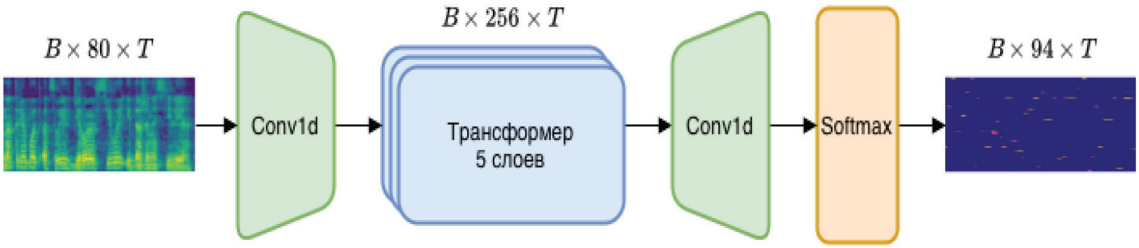


Рисунок 1. Архитектура для вычисления фонетических апостериограмм High-Fidelity Neural Phonetic Posteriorgrams [2]

Обучение модели проводилось с использованием комбинации двух функций потерь. В качестве первой выступало расстояние Кульбака-Лейблера [4], которое оценивало расхождение между истинным P_t и предсказанными $\hat{P}_t$ распределениями для каждого фрейма аудиосигнала:

$$D_{KL}(P_t \,||\, \hat{P}_t) = \sum_i P_t(i) \ln \frac{P_t(i)}{\hat{P}_t(i)}, \quad (3)$$

где индекс i нумеровал фонемы, на которых вычислялись распределения вероятностей.

Второй функцией потерь была бинарная кросс-энтропия с логитами, предсказываемыми моделью [2], которая сочетала sigmoid-активацию и бинарную кросс-энтропию [5]:

$$D_{BCE}(y_t, \hat{y}_t) = \frac{1}{N}\sum_i [-y_t(i)\ln(\sigma(\hat{y}_t(i))) - (1-y_t(i))\ln(1-\sigma(\hat{y}_t(i)))], \quad (4)$$

где $\hat{y}$ — логиты, y — истинные метки фонем (0 или 1), σ(x) — сигмоидная функция. Комбинация этих функций потерь позволила достичь устойчивой сходимости и высокой точности выделения фонем.

Обучение модели извлечения фонетических апостериограмм

В работе было проведено обучение описанной выше модели [2] на наборе данных Common Voice Corpus 21.0 [6]. Данный набор разработан и поддерживается компанией Mozilla Foundation в рамках инициативы по созданию свободно доступных ресурсов для обучения систем распознавания речи. Он охватывает широкий спектр сценариев использования благодаря своей гетерогенности и масштабу. Использованная в работе версия для английского языка содержит более 110 тысяч часов аудиозаписей, собранных от более чем 75 тысяч добровольцев по всему миру. Каждая запись представляла собой короткую фразу, прочитанную носителем языка в различных условиях, что обеспечивает естественную вариативность по таким параметрам, как акцент, тембр, фоновый шум, устройство записи и эмоциональная окраска высказывания. Участники проекта предоставили данные на условиях открытой лицензии Creative Commons Zero, что позволяет вторичным пользователям распространять, ремикшировать, адаптировать и дополнять материал на любом носителе или в любом формате без каких-либо условий.

Для непосредственно обучения была использована подвыборка «Validated» речевого набора Common Voice 21.0, включавшая в себя 1 838 943 аудиозаписи речевых высказываний, каждая из которых сопровождалась соответствующей текстовой транскрипцией, что позволило автоматизировать предварительную фонетическую разметку. Все оригинальные аудиофайлы были представлены в формате MP3 с частотой дискретизации 48 кГц, в данной работе они были автоматически конвертированы в монофонический формат WAV с частотой дискретизации 16 кГц. На этапе предварительной подготовки текстовые транскрипции подверглись нормализации: все символы были приведены к нижнему регистру, удалены знаки препинания и специальные символы. Такая обработка позволила устранить вариации в представлении лексических единиц, которые могут возникнуть из-за регистровых различий или наличия пунктуационных элементов, не имеющих акустического эквивалента. Это особенно важно для согласования текстовой и звуковой модальностей перед этапом выравнивания.

Далее был применен метод выравнивания Montreal Forced Aligner [7] для сопоставления фонем и звуковой дорожки. Обработка проводилась с применением предобученной акустической модели для английского языка и лексического словаря, включающего транскрипцию слов в международный фонетический алфавит. Акустическая модель MFA имеет высокую обобщающую способность при анализе артикуляционных паттернов, включая вариации произношения и коартикуляционные эффекты. В результате обработки был получен набор файлов в формате TextGrid [8], содержащих временные метки для каждой фонемы в соответствующих аудиозаписях. Для фонемной разметки на основе этих файлов применялось мягкое унитарное кодирование, учитывающее перекрытия интервалов фонем. Это позволило учесть переходные эффекты между фонемами в каждом фрейме за счет сохранения информация о частичном присутствии нескольких фонем в одном фрейме, избегая потери данных на их границах. Словарь фонем был дополнен специальными токенами: /unk/ для неизвестных фонем и /_/ для пустых интервалов.

Модель Neural PPG обучалась до сходимости на подготовленных размеченных данных в течение 102 эпох с размером батча 64 образ-

ца. Использовались реализации оптимизатора Adam и планировщика ReduceLROnPlateau из библиотеки PyTorch. Планировщик снижал скорость обучения в два раза при отсутствии улучшений ошибки валидации на протяжении трех эпох. Точность составляла 84% на выделенной тестовой подвыборке при предсказании по максимально вероятной фонеме. На рисунке 2 представлен пример предсказываемых фонетических апостериограмм в сопоставлении с мел-спектрограммой и истинной апостериограммой из разметки. Видно, что обученная модель достаточно точно предсказывала истинные фонемы, допуская небольшие ошибки, связанные с наличием шума в аудиофайлах.

Формирование наборов искаженных голосовых аудиозаписей для верификации

Для проведения дальнейших экспериментов по вычислению расстояний на основе тестового подмножества стандартного для задачи верификации набора данных VoxCeleb1 [9] были сформированы 3 множества искаженных аудиопримеров. Использовалось подмножество VoxCeleb1-Test, которое включало в себя 4874 записи, собранные из интервью и выступлений 40 знаменитостей различного возраста, пола и этнической принадлежности. Аудиодорожки были извлечены из роликов, загруженных на платформу YouTube и охватывали широкий спектр условий и сценариев записи: фоновые шумы, разнообразие акцентов, эмоциональные интонации, изменение тембра голоса в зависимости от контекста, а также использование различных типов микрофонов и устройств. Все аудиофайлы были оригинально представлены в формате WAV с частотой дискретизации 16 кГц.

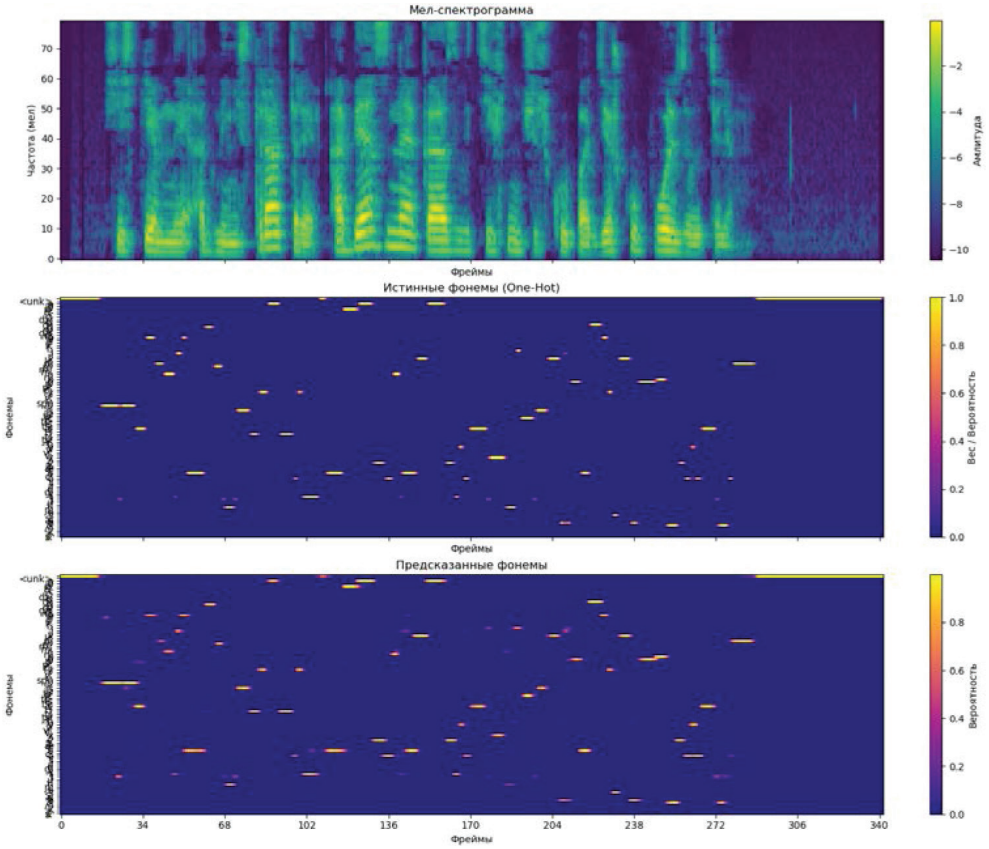


Рисунок 2. Пример фонетической апостериограммы (внизу) на основе мел-спектрограммы (вверху); для сравнения приведена истинная фонемная разметка (в середине)

Первый вид искажений аудиосигналов заключался в добавлении случайно выбранного аддитивного шума с контролируемым уровнем отношения сигнал-шум. Для обеспечения корректного наложения шума был реализован механизм динамической адаптации длины шумового сегмента к длительности целевого сигнала: при недостаточной продолжительности шума выполняется циклическая репликация, а при избыточной — случайная сегментация с равномерным распределением начальных позиций. Для каждой оригинальной аудиозаписи генерировались шесть версий с различными значениями SNR от -5 дБ до 20 дБ с шагом 5 дБ. Источниками шума служили аудиозаписи из набора данных DEMAND [10] включавшие в себя разнообразные типы фоновых помех (дорожные, бытовые, электронные и другие), что обеспечивало достаточную вариативность набора.

Моделирование различных уровней сигнал-шум осуществлялось через вычисление энергетических характеристик сигналов. Мощность чистого речевого сигнала определялась как среднее квадратов амплитудных отсчетов, аналогично для шумовой компоненты:

$$P_{\text{сигнал}} = \frac{1}{T}\sum_{t=1}^{T} x_t^2,\ P_{\text{шум}} = \frac{1}{T}\sum_{t=1}^{T} n_t^2. \qquad (5)$$

Целевая мощность аддитивного шума вычислялась по формуле

$$P_{\text{цель}} = P_{\text{сигнал}} 10^{\frac{-\text{SNR}}{10}}. \qquad (6)$$

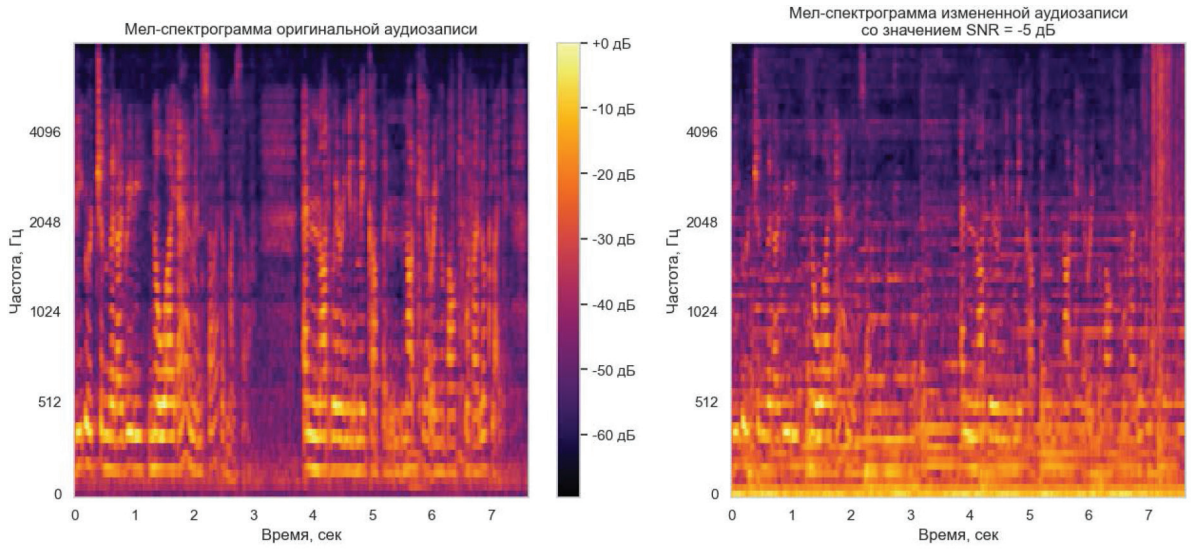
Коэффициент α, обеспечивающий необходимое соотношение мощностей, вычислялся как корень из отношения мощностей

$$\alpha = \sqrt{\frac{P_{\text{ссеп}}}{P_{\text{ссеп}}}}. \qquad (7)$$

Итоговый сигнал с наложенным шумом формировался согласно формуле

$$y_t = x_t + \alpha n_t, \qquad (8)$$

где y — зашумленный сигнал, x — оригинальный сигнал, n — шумовая добавка. Пример мел-спектрограммы оригинального сигнала и искаженного при помощи наложения шума представлен на рисунке 3.



а) б)

Рисунок 3. Спектрограммы аудиосигналов: а) оригинального чистого; б) искаженного за счет добавления шума с SNR = -5 дБ

Второй вид искажений представлял собой контролируемые нелинейные преобразования, заключающиеся в применении функции мягкого ограничения на амплитудных отсчетах исходных речевых сигналов [11]. Процесс формирования искаженных сигналов включал несколько этапов. Сначала к исходному сигналу, нормированному по амплитуде (в диапазоне [-1; 1]), применялась функция гиперболического тангенса:

$$y_t = \tanh(Lx_t), \quad (9)$$

где L — заданный уровень искажений. Увеличение уровня приводило к сжатию амплитудных пиков, вызывало генерацию гармонических искажений и изменение тембра речи. За этим преобразованием следовала нормализация за счет деления на максимальную амплитуду сигнала. Это обеспечивало сохранение энергетического баланса и предотвращало возникновение новых дополнительных артефактов, связанных с выходом за пределы допустимого динамического диапазона.

Для каждой оригинальной аудиозаписи из подмножества VoxCeleb1-Test генерировались шесть версий с заданными уровнями искажения от 1 до 21. Низкие значения соответствовали слабой компрессии сигнала, а высокие — выраженному ограничению амплитуды, имитирующему критическое насыщение аналоговых усилителей или перегрузку микрофонов. Пример мел-спектрограммы оригинального сигнала и искаженного при помощи изменения уровня Distortion представлен на рисунке 4.

Третий вид искажений заключался в добавлении реверберационных эффектов в аудиозаписи в виртуальном замкнутом пространстве. Применялся метод геометрической акустики, который позволял учесть многократные отражения звуковых волн от поверхностей помещения с заданной степенью звукопоглощения. Алгоритм моделирования [12] включал создание виртуальной комнаты размером 6×5×3 м^3, где источник звука и микрофон размещались в фиксированных положениях. Распространение звука рассчитывалось с учетом прямоугольной геометрии помещения и изотропных свойств материалов, применялось до 20 порядков отражений для воспроизведения сложных траекторий волн. Регистрируемый сигнал ограничивался по длине исходной аудиозаписи. Для предотвращения возникновения артефактов, связанных с выходом за пределы динамического диапазона результирующий сигнал нормализовался по амплитуде.

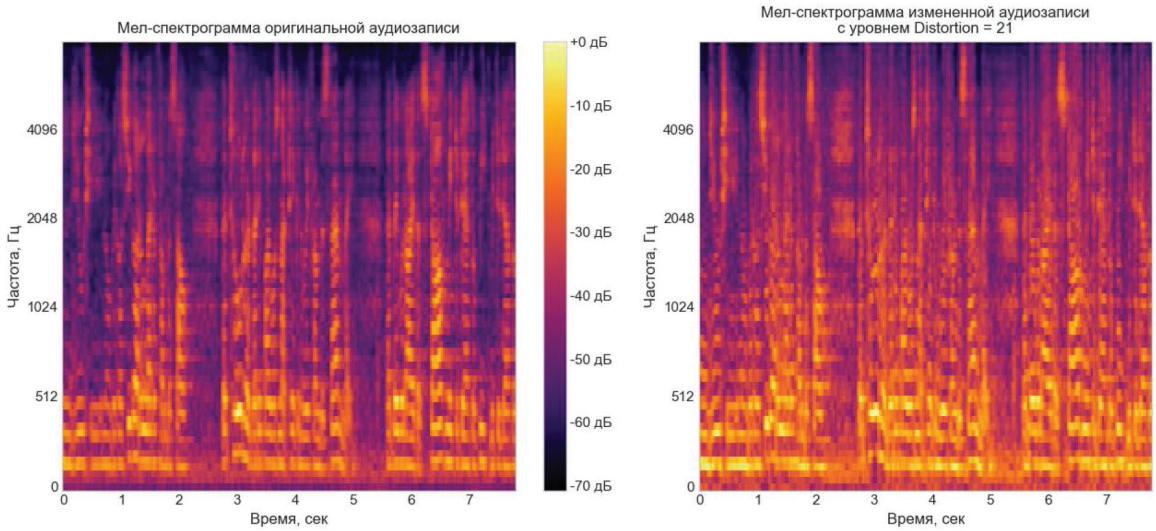
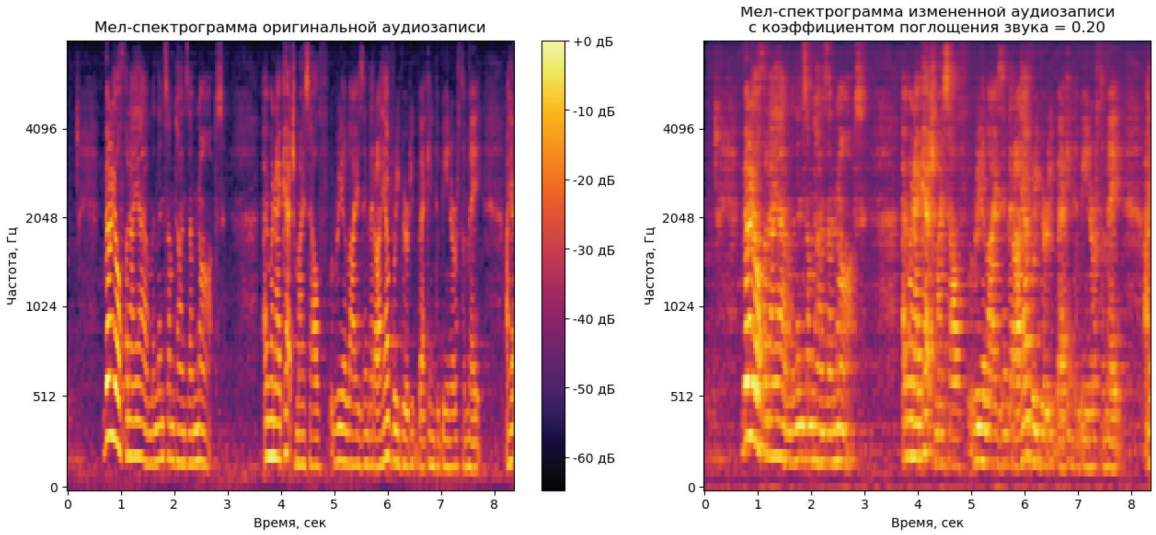


Рисунок 4. Спектрограммы аудиосигналов: а) оригинального чистого; б) нелинейно искаженного с параметром L = 21.

Для каждой оригинальной аудиозаписи формировались шесть версий с коэффициентами энергетического поглощения α в диапазоне от 0,2 до 0,95. Высокие значения α соответствовали материалам с сильным поглощением (ковры, акустические панели), а низкие — с существенным отражением (бетон или стекло). Таким образом, обеспечивается диапазон виртуальных условий от почти «заглушенных» комнат до эховых залов с длительным временем реверберации. Пример мел-спектрограммы оригинального сигнала и искаженного при помощи изменения коэффициента поглощения представлен на рисунке 5.



а) б)

Рисунок 5. Спектрограммы аудиосигналов: а) оригинального чистого; б) версия с реверберацией с коэффициентов поглощения от поверхностей 0,2

Метрики для сравнения аудиосигналов и оценки качества верификации

Сравнение фонетических апостериограмм искаженных речевых аудиосигналов осуществлялось за счет вычисления усредненного расстояния между распределениями вероятностей фонем, полученных на соответствующих фреймах сигналов за счет применения обученной нейросетевой модели извлечения фонетических апостериограмм. Модельные искажения аудиосигналов описанные выше (аддитивный шум, нелинейные искажения и реверберация), не приводили к изменению длительности аудиозаписей, поэтому не возникало проблем, связанных с сопоставлением фреймов друг с другом. Мерой расстояния между распределениями фонем была выбрана дивергенция Йенсена-Шеннона [4]:

$$\mathrm{JSD}(P \,\|\, Q) = \frac{1}{2} D_{KL}(P \,\|\, M) + \frac{1}{2} D_{KL}(Q \,\|\, M), \qquad (10)$$

где $\underline{M \quad Q}$ — среднее распределение, D_{KL} — дивергенция Кульбака–Лейблера для сопоставления распределений вероятностей. Для каждой пары аудиосигналов (оригинальной и искаженной) проводилось вычисление по формуле (10) всех пар фреймов с последующим усреднением по числу фреймов.

Для сравнительной оценки эффективности верификации дикторов при различных искажениях аудиосигналов использовался

уровень эквивалентных ошибок (equal error rate, EER) [13]. Процесс расчета EER включал построение зависимости частот ошибок ложного допуска $FAR(t)$ и ложного отказа доступа $FRR(t)$ от порога t принятия решения (рабочей точки). Затем с помощью линейной интерполяции определялось пороговое значение t_{EER}, при котором выполнялось условие равенства частот ошибок:

$$EER = FAR(t_{EER}) = FRR(t_{EER}). \quad (11)$$

Данная величина вычислялась с использованием сторонней предварительно обученной модели голосовой верификации TDNN [14]. Каждая группа аудиофайлов с заданным видом или уровнем искажения характеризовалась соответствующим значением EER.

Обсуждение результатов

На рисунках 6–8 представлены графики зависимостей среднего значения JSD между оригинальными и искаженными аудиофайлами и соответствующего уровня эквивалентных ошибок EER. Среднее значение JSD характеризовало степень расхождения фонетических апостериограмм. Каждый из графиков этой величины демонстрировал рост дивергенции при росте искажений в аудиосигнале. Чем эта величина становилась больше, тем сильнее отличались друг от друга вычисленные по искаженным сигналам и образцовые периодограммы, что было ожидаемым эффектом и может быть связать с тем, что ухудшение качества приводило к «размытию» распределений вероятностей фонем для индивидуальных фреймов. Таким образом, средняя JSD выступала как мера разборчивости речи.

Рост мощности аддитивной шумовой добавки (рис. 6) приводил к закономерному увеличению частоты ошибок при верификации. При этом изменение EER и JSD проходило практически синхронно с коэффициентом корреляции Пирсона близким к единице. Разброс EER составлял от 9,4% до 29,1%, а средняя дивергенция Йенсена-Шеннона менялась еще существеннее — в шесть раз, от 0,1 до 0,6 при росте шумовой добавки. Аналогичное поведение наблюдалось и при внесении нелинейных искажений (рис. 7). Разброс частот ошибок составил от 9,6% до 23,6%, JSD менялось от 0,66 до 0,16. Однако поведение метрик при добавлении реверберации в сигналы (рис. 8) несколько изменилось. Частота эквивалентных ошибок была практически стабильной, меняясь в диапазоне от 9 до 12%, в то время как JSD находилось в диапазоне от 0,23 до 0,61 и было существенно более чувствительно к этим искажениям.

Таким образом, из сопоставления метрики верификации EER и метрики фонетической разборчивости JSD можно сделать следующие выводы. Во-первых, оценка последней может играть роль независимой оценки качества голосовых аудиозаписей, напрямую связанную с возможным качеством работы систем верификации. Во-вторых, JSD, вероятно, более чувствительна к искажениям, и может достаточно точно работать как независимая оценка разборчивости речи. В-третьих, полученные в ходе экспериментов результаты и само построение этой метрики дают основания предположить, что JSD может служить индикатором надежности признакового пространства, формируемого моделью извлечения фонетических апостериограмм. Она может использоваться для регуляризации обучения систем речевой верификации, для повышения устойчивости последней к фонетическим искажениям.

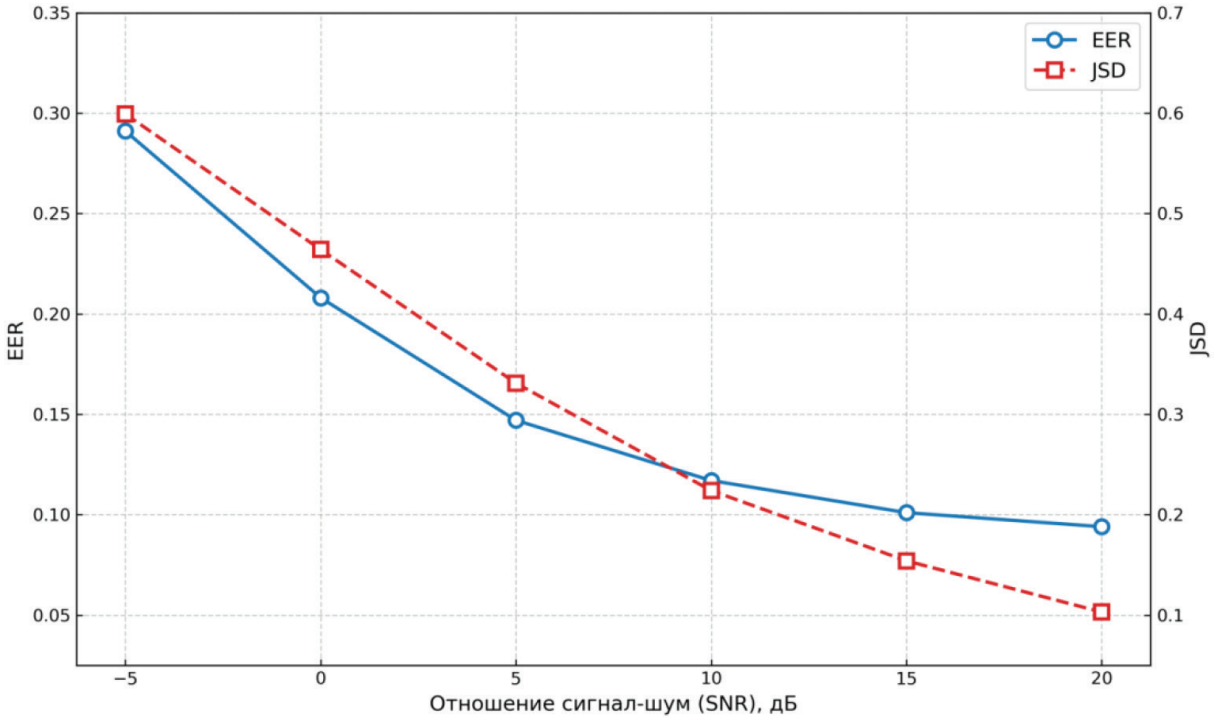


Рисунок 6. Среднее значение дивергенции Йенсена-Шеннона
и эквивалентных ошибок EER при добавлении аддитивного шума

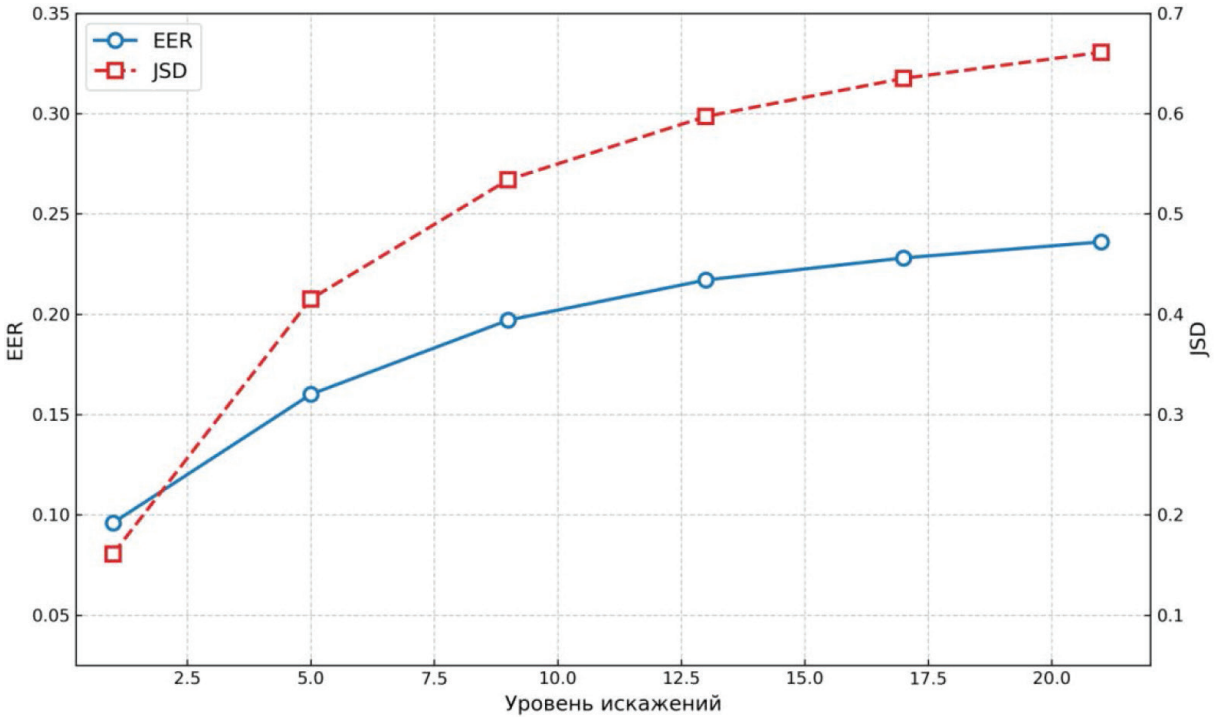


Рисунок 7. Среднее значение дивергенции Йенсена-Шеннона
и эквивалентных ошибок EER при различных значениях уровня искажений

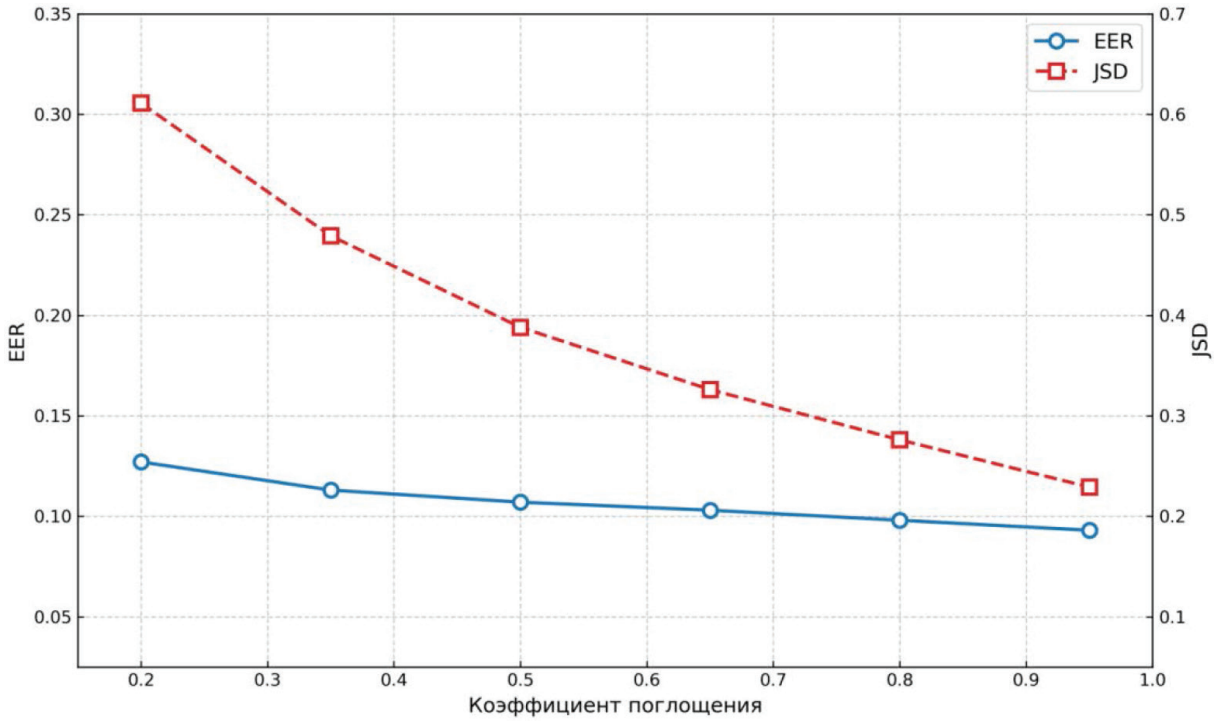


Рисунок 8. Среднее значение дивергенции Йенсена-Шеннона
и эквивалентных ошибок EER для значений коэффициента поглощения

Заключение

В работе предложен новый метод оценки фонетической разборчивости речи, основанный на применении глубокой нейросетевой модели вычисления фонетических апостериограмм и вычислении «расстояния» между ними с помощью дивергенции Йенсена-Шеннона. Апробация этого метода на сгенерированных примерах речевых аудиозаписей продемонстрировал высокую чувствительность вычисляемого расстояния от степени искажения сигнала. Сопоставление с оценками качества верификации продемонстрировало, что средняя дивергенция может быть эффективно применена для оценки качества речевых записей при биометрической верификации пользователей.

Библиографический список

1. Hazen T.J. Shen W., White C. Query-by-example spoken term detection using phonetic posteriorgram templates // 2009 IEEE Workshop on Automatic Speech Recognition & Understanding. Moreno. Italy. 2009. P. 421–426.

2. Cameron C., Churchwell C., Morrison M., Pardo B. High-Fidelity Neural Phonetic Posteriorgrams // 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). Seoul. Korea. 2024. P. 823–827.

3. Handbook of the International Phonetic Association: a guide to the use of the International Phonetic Alphabet. Cambridge: Cambridge University Press, 1999. ix + 204 p.

4. Cover T., Thomas J.A. Elements of Information Theory. 2nd ed. Wiley. New Jersey, 2006. 748 p.

5. Binary Cross-Entropy Loss // PyTorch 2.9 documentation: сайт. URL: https://docs.pytorch.org/docs/stable/generated/torch.nn.BCELoss.html (дата обращения: 10.10.2025).

6. Ardila R., Branson M., Davis K., Kohler M., Meyer J., Henretty M., Morais R., Saunders L., Tyers F., Weber G. Common Voice: A Massively-Multilingual Speech Corpus // Proceedings of the Twelfth Language Resources and Evaluation Conference. Marseille, France. European Language Resources Association, 2020. P. 4218–4222.

7. McAuliffe M., Socolof M., Mihuc S., Wagner M., Sonderegger M. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi // Proc. Interspeech. 2017. P. 498–502.

8. Neuroth H., Lohmeier F., Smith K.M. TextGrid — Virtual Research Environment for the Humanities // The International Journal of Digital Curation. Issue 2, Volume 6.| 2011. P. 222–231.

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

10. Thiemann J., Ito N., Vincent E. The Diverse Environments Multi-Channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings // The Journal of the Acoustical Society of America, 2013.

11. Schuck Jr. A., Bodmann B. Audio non-linear modeling through hyperbolic tangent functionals // Proceedings of the 19th International Conference on Digital Audio Effects (DAFx-16). 2016. P. 103–108.

12. Scheibler R., Bezzam E., Dokmanic I. Pyroomacoustics: A Python package for audio room simulations // IEEE Signal Processing Letters, 2020 – vol. 27, P. 133–137.

13. Болл Р.М., Коннел Дж.Х., Панканти Ш., Ратха Н.К., Сеньор Э.У. Руководство по биометрии. М. : Техносфера, 2007. 368 с.

14. Snyder D., Garcia-Romero D., Sell G., Povey D., Khudanpur S. X-Vectors: Robust DNN Embeddings for Speaker Recognition // IEEE International Conference on Acoustics, Speech and Signal Processing, 2018. P. 5329–5333.