Построение апостериорных интерпретаций для моделей классификации аудиоданных

Ю. Р. Пак, И. Ю. Терёхина

***Аннотация*—В данной работе рассматривается задача построения интерпретаций для моделей машинного обучения, классифицирующих аудиоданные. Предлагается подход, позволяющий строить интерпретации в визуальной и прослушиваемой форме посредством маскирования спектрограмм на основе карт вкладов признаков и последующего восстановления сигнала. Для получения карт вкладов признаков использованы методы Saliency, Grad-CAM, LIME и SHAP. Эти методы универсальны и могут применяться к моделям различных архитектур. Эффективность подхода оценивалась с точки зрения согласованности интерпретаций с моделью и простоты их восприятия. Проводились эксперименты с различными типами масок, а также с добавлением фоновых шумов. Было показано, что для предложенного подхода наибольшее затруднение представляет компромисс между приближением поведения модели и простотой интерпретаций. Добавление шумов не нарушает общих тенденций, но тип шума влияет на поведение модели и характеристики соответствующих интерпретаций.**

***Ключевые слова*—апостериорная интерпретация, интерпретируемое машинное обучение, классификация аудиоданных, спектрограмма.**

I. Введение

Модели машинного обучения, работающие со звуком, успешно применяются для задач аутентификации по голосу [1], постановки медицинских диагнозов на основе респираторных звуков [2], акустического мониторинга окружающей среды [3], а также для обнаружения аномальных событий (крик, выстрел) [4]. Перечисленные области применения относятся к критическим сферам деятельности, требующим прозрачности и непредвзятости. Более того, в таких сферах ошибки в работе моделей могут привести к катастрофическим последствиям. В связи с этим важно обеспечить интерпретируемость моделей, принимающих на вход аудиоданные.

Под интерпретируемостью модели машинного обучения понимают способность объяснять принципы ее работы в терминах, понятных человеку [5]. Интерпретируемое машинное обучение изучает методы, обеспечивающие такую интерпретируемость [6]. Решение задачи построения интерпретируемых моделей, предназначенных для обработки и анализа аудиоданных, затруднено спецификой таких данных: они имеют большую размерность, а для понимания компактных признаковых представлений аудио требуются специальные знания в области обработки сигнала [7].

В последнее время можно выделить две тенденции в решении задачи построения интерпретируемых моделей для аудио [8]. Первый подход предполагает адаптацию популярных методов, разработанных для табличных данных и изображений, к аудиоданным. Многие методы ввиду своей универсальности переносятся на аудиоданные без изменений. Например, в [9] с помощью методов LIME [10] и SHAP [11] осуществлен подбор наиболее информативных признаков в частотной и временной областях для моделей машинного обучения, решающих задачу обнаружения неисправностей индустриальных машин на основе данных аудиодатчиков. В [12] метод Grad-CAM [13] применяется для анализа решений модели, обученной решать задачу классификации звуков внутри пчелиных ульев. Второй подход заключается в разработке методов, предназначенных специально для моделей, принимающих на вход данные, что позволяет лучше учитывать специфику такого типа данных. В работах [14]-[17] предложены модификации популярных методов интерпретации с учетом особенностей аудиоданных и их признаковых представлений. Например, метод SLIME [17] основан на методе LIME, но вместо суперпикселей использует сегменты во временной, частотной или частотно-временной области. Помимо адаптированных методов рассматриваются специфичные подходы, в основе которых лежит обучение декодера-интерпретатора [18, 19].

В настоящей работе исследуется подход, предложенный в публикации [19], посвященной методу LMAC. Данный подход строит карту вкладов признаков с помощью специально обученного декодера, повторяющего архитектуру модели в обратном порядке. На основе полученной с его помощью карты строится маска, которая применяется к спектрограмме. В результате получается интерпретация, содержащая только самые важные для предсказания модели компоненты спектрограммы. С помощью обратных преобразований из нее восстанавливается сигнал – интерпретация в прослушиваемой форме. Такой подход сочетает в себе преимущества использования

Статья получена 6 октября 2025
Пак Юлия Руслановна, МГУ имени М.В. Ломоносова (email: pjulie717119@gmail.com).
Терёхина Ирина Юрьевна, МГУ имени М.В. Ломоносова (email: iteryokhina@cs.msu.ru).

компактных признаковых представлений аудиосигнала, а также возможность анализа результатов интерпретации не только в визуальной, но и в прослушиваемой форме.

В [19] также приводится сравнение LMAC с другими методами, позволяющими получить интерпретации в виде карт вкладов признаков (атрибуций). Однако авторами не уточняется, по какой процедуре на основе этих атрибуций строится маска, и рассматривается только один тип маски. В связи с этим представляет интерес оценить эффективность популярных методов апостериорной интерпретации Saliency [20], Grad-CAM, LIME и SHAP в сочетании с различными способами получения масок (minmax, sigmoid, бинаризация, topK% важных компонентов). Перечисленные методы интерпретации хорошо изучены, универсальны и могут применяться к моделям различных архитектур. В связи с этим исследование их эффективности в контексте подхода, предполагающего маскирование и восстановление сигнала, может приблизить к нас к разработке универсального и удобного для пользователей инструмента интерпретации моделей, принимающих на вход аудиоданные.

Результаты исследования представлены в открытом репозитории [21].

II. Методология

Пусть $x \in [-1,1]^D$ – входной пример, представленный последовательностью отсчетов. С помощью преобразования $T(\cdot)\colon [-1,1]^D \to R^{F\times T}$ получаем спектрограмму $X = T(x)$, $X \in R^{F\times T}$, F – число частотных полос, T – число временных шагов. Обозначим $g(\cdot)\colon R^{F\times T} \to R^K$ – отображение, моделируемое целевым классификатором, $g(X)$ – логиты, $\hat{p}(X) = SoftMax\big(g(X)\big)$ – вероятности.

С помощью метода интерпретации $\mathcal{G}(\cdot,\cdot,\cdot)$ для модели g, спектрограммы X и класса $c \in \{1,2,\dots,K\}$ строится карта вкладов признаков:

$$L = \mathcal{G}(g, X, c),$$

из которой с помощью функции $Norm(\cdot)$ получается маска:

$$M = Norm(L).$$

Согласно описанной в [19] процедуре, маска M применяется к спектрограмме, после чего выполняется обратное преобразование $T_{inv}(\cdot)\colon R^{F\times T} \to [-1,1]^D$ и восстанавливается прослушиваемая интерпретация:

$$x_{int} = T_{inv}(X \otimes M).$$

A. Извлечение признаков и восстановление сигнала

В ходе исследования использовалась модель, принимающая на вход признаковые представления сигналов в частотно-временной области – энергетические спектрограммы и мел-спектрограммы в децибелах. Для восстановления прослушиваемых интерпретаций к спектрограммам применялись обратные преобразования.

На Рис. 1 приведена схема прямых и обратных преобразований. Стрелками со сплошными линиями обозначены точные преобразования. Приближенным преобразованиям соответствуют стрелки с пунктирными линиями. Энергетические спектрограммы и мел-спектрограммы в децибелах выделены, поскольку являются конечными признаковыми представлениями для подачи на вход модели.

Обращение мел-фильтров (InvMel) – приближенная операция, представляющая собой решение оптимизационной задачи методом наименьших квадратов. Обратное дискретное оконное преобразование Фурье (ISTFT) – точная операция, для выполнения которой используется заранее сохраненный фазовый спектр. Таким образом, для энергетической спектрограммы при отсутствии маскирования возможно восстановление сигнала без потерь, а в случае мел-спектрограммы возможно лишь приближенное восстановление.

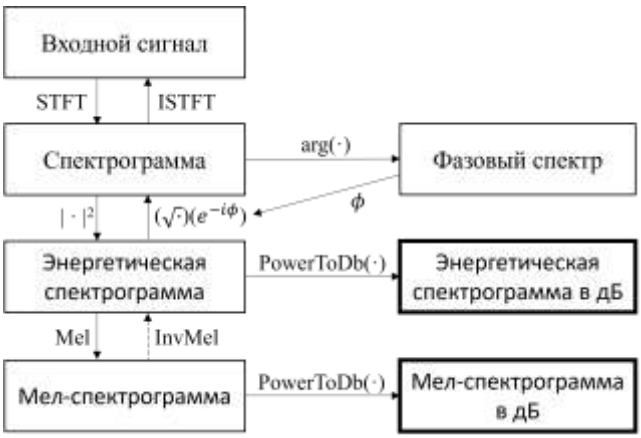

Рис.1. Схема прямых и обратных преобразований признаков

B. Построение апостериорных интерпретаций

Для получения карт признаков L в качестве функции $G(\cdot,\cdot,\cdot)$ были использованы следующие популярные методы апостериорной интерпретации:

Saliency [20]. Строит карту заметности признаков на основе градиента логита целевого класса по входному примеру.

Grad-CAM [13]. Строит карту заметности признаков на основе градиента логита целевого класса по тензору активаций последнего сверточного слоя модели.

LIME [10]. Строит интерпретацию с помощью обучения простой и интерпретируемой по своей природе модели, аппроксимирующей целевой классификатор в окрестности рассматриваемого примера.

SHAP [11]. Строит карту вкладов признаков, используя числа Шепли из теории игр. В данной работе используется метод Deep SHAP – модификация метода SHAP, в основе которой лежит метод DeepLIFT [22, 23].

C. Построение масок

Результатом применения функции $Norm(\cdot)$ к карте вкладов признаков может быть бинарная маска $M \in \{0,1\}^{F\times T}$, полностью сохраняющая или удаляющая компоненты спектрограммы, или мягкая маска $M \in [0,1]^{F\times T}$, допускающая частичное сохранение и подавление компонентов. Отметим, что маски строились как для полных карт атрибуций L, так и для карт $L_+ = max(0, L)$, содержащих только неотрицательные вклады. При необходимости особенности применения функции $Norm(\cdot)$ к L и L_+ описываются отдельно. В остальных случаях мы ограничиваемся приведением формулы для L. Обозначим $f \in \{1,2,\dots,F\}$ – номер частотной полосы, $t \in \{1,2,\dots,T\}$ – номер шага по времени.

Минимаксная нормализация:

$$M_{ft}^{minmax} = \frac{L_{ft} - \min_{f,t}(L)}{\max_{f,t}(L) - \min_{f,t}(L)}.$$

Сигмоида:

$$M_{ft}^{sigmoid} = \sigma(L_{ft}), \text{ где}$$
$$\sigma(u) = \frac{1}{1+exp^{-u}}.$$

Бинаризация:

$$M_{ft}^{binary} = \mathbb{1}\{L_{ft} > 0\}.$$

Для L_+ бинаризация выполнялась по порогу $\tau \in (0,1)$:

$$\left(M_{\tau}^{binary}\right)_{ft} = \mathbb{1}\{(L_+)_{ft} \geq \tau\}.$$

Бинаризация по topk%:

$$\left(M_{k}^{\text{top}K}\right)_{ft} = \mathbb{1}\{L_{ft} \geq \theta_k\}.$$

Здесь θ_k – порог, равный $(1-k)$-квантилю всех значений вкладов для заданной доли $k \in (0,1]$:

$$\theta_k = quantile_{\,1-k}((L)_{f,t}).$$

D. *Метрики качества*

Для оценки эффективности рассматриваемого подхода использовался набор метрик, подробно описанный в [19]. Данный набор метрик отражает два критерия качества, которым должны отвечать генерируемые интерпретации:

- *Согласованность с моделью*. Маски, построенные на основе карт вкладов признаков, должны сохранять наиболее важные для предсказания признаки и отбрасывать наименее важные;
- *Лаконичность*. Интерпретации, полученные в результате применения масок к спектрограммам, не должны содержать избыточной информации и должны быть прицельно сфокусированы на наиболее значимых участках спектрограмм.

Рядом с сокращенными названиями метрик указаны стрелки:

- ↑ – для метрики предпочтительнее более высокое значение;
- ↓ – для метрики предпочтительнее более низкое значение.

Следующие метрики оценивают степень согласованности метода с моделью для некоторого класса $c \in \{1, 2 \ldots, K\}$:

Faithfulness on Spectra (FF ↑). Измеряет, насколько маскированная часть важна для предсказания.

$$\text{FF} = \frac{1}{N}\sum_{n=1}^{N} \text{FF}_n, \text{ где}$$
$$\text{FF}_n = \hat{p}(X_n)_c - \hat{p}\left(X_n \otimes (1-M)\right)_c.$$

Average Increase (AI ↑). Измеряет степень положительного влияния маскирования на предсказание.

$$\text{AI} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}[\hat{p}(X_n \otimes M) > \hat{p}(X_n)_c] \times 100.$$

Average Drop (AD ↓). Оценивает потерю уверенности модели при маскировании.

$$\text{AD} = \frac{1}{N}\sum_{n=1}^{N} \frac{\max(0, \hat{p}(X_n)_c - \hat{p}(X_n \otimes M)_c)}{\hat{p}(X_n)_c} \times 100.$$

Average Gain (AG ↑). Оценивает прирост уверенности модели при маскировании.

$$\text{AG} = \frac{1}{N}\sum_{n=1}^{N} \frac{\max(0, \hat{p}(X_n \otimes M)_c - \hat{p}(X_n)_c)}{1 - \hat{p}(X_n)_c} \times 100.$$

Input Fidelity (Fid-In ↑). Оценивает долю интерпретаций, для которых предсказание сохраняется.

$$\text{Fid-In} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left[\operatorname*{argmax}_c g(X_n)_c = \operatorname*{argmax}_c g(X_n \otimes M)_c\right].$$

Следующие метрики предназначены для оценки лаконичности интерпретаций:

Sparseness (SPS ↑). Измеряет степень разреженности (лаконичности) интерпретации. Основана на индексе Джини.

$$\text{SPS} = 1 - 2\sum_{i=1}^{d} \frac{v_{(i)}}{|v|_1}\left(\frac{d-i+0.5}{d}\right).$$

Здесь $d = F \cdot T$ – число признаков, $v = \left(v_{(1)}, \ldots, v_{(d)}\right)$ – вектор упорядоченных по неубыванию значений вектора $v = (v_1, \ldots, v_d)$, который в свою очередь содержит абсолютные значения L.

Complexity (COMP ↓). Соответствует энтропии распределения вкладов признаков в интерпретацию.

$$\text{COMP} = E_i\left[-\ln\left(P_{\mathcal{G}}\right)\right] = -\sum_{i=1}^{d} P_{\mathcal{G}}(i) \ln\left(P_{\mathcal{G}}(i)\right), \text{ где}$$
$$P_{\mathcal{G}}(i) = \frac{|v_i|}{\sum_{j\in[d]}|v_j|}, \qquad P_{\mathcal{G}} = \left\{P_{\mathcal{G}}(1), \ldots, P_{\mathcal{G}}(d)\right\}$$

III. ЭКСПЕРИМЕНТЫ

A. *Описание экспериментального стенда*

Набор данных. Для проведения исследования был выбран набор данных ESC-50 [24], содержащий 2000 аудиозаписей звуков окружающей среды, распределенных равномерно в 50 классов. Набор данных разделен на 5 блоков для перекрестной проверки. В данной работе блоки 1, 2 и 3 были объединены в обучающую выборку. Блоки 4 и 5 использовались в качестве валидационной и тестовой выборки соответственно.

Модель. Для проведения исследования на признаковых представлениях были обучены модели с одинаковой архитектурой Cnn14, представленной в [25] Модель предобучена на наборе данных AudioSet [26]. В таблице I приведены результаты обучения моделей на спектрограммах двух типов. В сопровождающем репозитории [21] приведены параметры спектрограмм и подробная процедура обучения моделей.

Таблица I
Результаты обучения моделей

Признаки	Acc, % (обуч.)	Acc, % (вал.)	Acc , % (тест)	Эпоха
STFT-dB	87.58	79.50	78.00	20
Mel-dB	97.00	93.25	92.25	20

Применение методологии. В ходе экспериментов использовались методы интерпретации, способы построения масок и метрики, описанные в разделе II. Для метода LIME был выбран размер окрестности 1000. Deep SHAP применялся с опорным фоном из 100 примеров. Для бинарных масок на основе неотрицательных атрибуций использовались значения τ – 0.25, 0.5, 0.75. Для масок topK% использовались значения k – 5, 30, 50.

Реализация экспериментов основана на открытых библиотеках и исходном коде авторов методов интерпретаций и соответствующих публикаций. В репозитории [21] приведен полный список использованных библиотек, а также ссылки на первичные источники кода.

B. *Описание экспериментов*

Эксперимент 1 (чистая выборка). Данный эксперимент проводился для обеих моделей на тестовой выборке без каких-либо модификаций.

Эксперимент 2 (выборка с фоновыми шумами). Данный эксперимент проводился для модели Mel-dB,

поскольку она показала более высокую точность на тестовой выборке. В ходе данного эксперимента к примерам из тестовой выборки добавлялись примеси.

В качестве примесей были выбраны следующие звуки:

- Белый шум (white noise)[1];
- Гудение техники в офисе (room ambience)[2];
- Стук копыт и ржание лошади (horse)[3].

Такой эксперимент позволяет оценить, насколько эффективны методы интерпретации при работе с реалистичными аудиоданными, подверженными искажениям.

IV. Результаты

Основные результаты экспериментов представлены в виде Таблиц I–VI. Полный набор таблиц по проведенным экспериментам доступен в сопровождающем репозитории [21], где приведены результаты для каждой комбинации сценария (чистые данные, white noise, room, horse) и соответствующего метода интерпретации.

A. Эксперимент 1.

В таблицах II и III представлены сводные результаты для лучших сочетаний “метод+маска”. Лучшее сочетание выбиралось посредством усреднения показателей метрик: в соответствующей таблице для данного сценария и метода: значения в столбцах приводились к диапазону $[0, 1]$, после чего для каждой строки вычислялось геометрическое среднее (значения метрик AD и COMP предварительно инвертировались). Лучшие показатели для метрик FF, AI, AG, FidIn, AD выделены зеленым, худшие – оранжевым. Для метрик SPS и COMP лучшие показатели выделены голубым, худшие – желтым

Метод LIME лучше всего сочетается с бинарной маской по полной карте атрибуций. LIME+bin продемонстрировал лучшие результаты по метрикам AI и AG для обеих моделей. Для модели STFT-dB данное сочетание имеет второй лучший показатель по метрике AD, а для модели Mel-dB – самый лучший, причем с большим отличием от результатов других сочетаний. Стабильно высокие результаты достигаются и по метрике FF. Значительно отличаются показатели FidIn для обеих моделей: для STFT-dB точность модели на маскированных спектрограммах снижается почти на 10%, а для Mel-dB точность модели стала больше. Такое различие можно объяснить тем, что первая модель изначально имела более низкую точность на тестовой выборке и, вероятно, хуже запомнила признаки обучающей выборки. Увеличение качества второй модели на маскированном входе является крайне положительным результатом и говорит о высокой степени согласованности сочетания LIME+bin с изучаемой моделью. Однако, судя по метрикам COMP и SPS, интерпретации получаются сложными и недостаточно сфокусированными. Это значит, что данное сочетание не справляется с компромиссом между согласованностью и лаконичностью.

Худшие результаты среди наиболее удачных сочетаний показывает метод SHAP с маской topK_50_pos для модели STFT-dB и маской sigmoid для модели Mel-dB. Судя по метрике AD, при маскировании спектрограммы уверенность снижается. Для модели STFT-dB данное сочетание показало высокие результаты по метрикам FF и SPS, но судя по показателю FidIn точность модели на интерпретациях снизилась до 11.75%. Судя по метрикам SPS и COMP, интерпретации получились одновременно сложными и более разреженными, чем для других сочетаний. При этом нельзя сказать, что показатель SPS оказался достаточно высоким. Для модели Mel-dB результат по метрике FF крайне низкий, точность модели на интерпретациях снизилась до 63.50%, а судя по COMP и SPS интерпретации получились сложными и совершенно не лаконичными. Можно сделать вывод, что интерпретации содержат много компонентов с малоразличимыми слабыми вкладами и относительно небольшое число “пиков”. При этом, несмотря на большое количество сохраненных признаков, этого недостаточно для того, чтобы положительно повлиять на уверенность модели.

Для обеих моделей сочетание Saliency+ topK_50_pos достигает высоких результатов по метрикам FF и AI и относительно низких по метрике AD. Результаты по метрике AG худшие среди всех приведенных сочетаний, а при подаче на вход модели маскированных спектрограмм ее точность снижается на 26,25-26,5%. В сочетании с низкой разреженностью (SPS) и невысокой сложностью (COMP) это говорит о том, что интерпретации Saliency+topK_50_pos сохраняют большое количество признаков, причем лишь относительно малое их число вносит значимый по сравнению с остальными вклад в предсказание. Grad-CAM+topK_50_pos на STFT-dB и Grad-CAM+bin на Mel-dB в целом повторяют эти тенденции, но маскирование спектрограмм не снижает точность модели больше чем на 15%. По метрике AD достигнуто высокое качество, но показатели FF относительно низкие. Таким образом, интерпретации размытые, с небольшим числом сильных компонент, однако маскированная часть содержит больше, чем в предыдущем случае, действительно важных участков спектрограмм, но при этом ощутимое число значимых компонентов остается вне маскированной области.

B. Эксперимент 2

Основные результаты эксперимента 2 приведены в сводных таблицах IV-VI. Выбор лучших сочетаний “метод+маска” проводился по той же процедуре, что и для эксперимента 1. Аналогичным образом выделены цветом лучшие и худшие показатели по метрикам.

Следует отметить, что сочетания, продемонстрировавшие лучшие усредненные результаты для чистой выборки, оказались лучшими и для выборок со всеми типами шумов.

Сочетание Saliency+topK_50_pos вновь дает достаточно простые, но недостаточно разреженные интерпретации. При отбрасывании важных компонентов уверенность модели падает, о чем говорит высокий показатель FF. Тенденция увеличения уверенности вновь положительная, но фактический прирост крайне мал. Интересными представляются результаты по метрике

[1] https://freesound.org/people/theundecided/sounds/165058/
[2] https://freesound.org/people/mzui/sounds/203297/
[3] https://freesound.org/people/foxen10/sounds/149024/

FidIn. На примерах с синтетическим белым шумом и фоновым шумом индустриального помещения результаты по данному показателю оказались значительно лучше, чем для аналогичных примеров без

Таблица II
Результаты лучших сочетаний “метод + маска” для STFT-dB

Метод + маска	FF ↑	AI ↑	AG ↑	FidIn ↑	SPS ↑	AD ↓	COMP ↓
Saliency + topK_50_pos	0.8042	21.2500	0.0054	0.5150	0.2462	48.3761	5.2216
Grad-CAM + topK_50_pos	0.7246	24.7500	0.2179	0.7025	0.2250	31.0750	4.7710
LIME + bin	0.7903	26.0000	14.2055	0.6850	0.2727	32.0287	10.9316
SHAP + topK_50_pos	0.8021	3.7500	0.3779	0.1175	0.4487	88.0014	9.5154

Таблица III
Результаты лучших сочетаний “метод + маска” для Mel-dB

Метод + маска	FF ↑	AI ↑	AG ↑	FidIn ↑	SPS ↑	AD ↓	COMP ↓
Saliency + topK_50_pos	0.9235	17.0000	0.0108	0.6600	0.1738	34.2287	3.3646
Grad-CAM + bin	0.7189	19.5000	2.3121	0.7825	0.1955	24.3636	4.9185
LIME + bin	0.8992	28.7500	12.3594	0.9325	0.2004	8.0670	9.5342
SHAP + sigmoid	0.4115	1.7500	0.7060	0.6350	0.0016	45.8593	10.3755

Таблица IV
Результаты лучших сочетаний “метод + маска” для Mel-dB(white noise), точность 64.25%

Метод + маска	FF ↑	AI ↑	AG ↑	FidIn ↑	SPS ↑	AD ↓	COMP ↓
Saliency + topK_50_pos	0.7112	32.5000	0.0097	0.7450	0.1275	25.3430	2.4690
Grad-CAM + bin	0.5262	15.7500	2.5479	0.5100	0.2917	49.5738	5.3962
LIME + bin	0.6988	34.2500	9.1746	0.8025	0.1878	19.6633	8.0352
SHAP + sigmoid	0.4797	5.0000	1.9955	0.3800	0.0011	69.0501	10.3755

Таблица V
Результаты лучших сочетаний “метод + маска” для Mel-dB(room), точность 84.50%

Метод + маска	FF ↑	AI ↑	AG ↑	FidIn ↑	SPS ↑	AD ↓	COMP ↓
Saliency + topK_50_pos	0.8507	27.2500	0.0107	0.6850	0.1575	31.0446	3.0499
Grad-CAM + bin	0.6246	16.7500	2.4583	0.6450	0.2628	37.1844	5.5138
LIME + bin	0.8346	22.7500	10.0662	0.8225	0.2097	21.2241	9.7336
SHAP + sigmoid	0.5062	3.0000	1.7427	0.4875	0.0010	61.3289	10.3755

Таблица VI
Результаты лучших сочетаний “метод + маска” для Mel-dB(horse), точность 73.00%

Метод + маска	FF ↑	AI ↑	AG ↑	FidIn ↑	SPS ↑	AD ↓	COMP ↓
Saliency + minmax	0.7745	10.7500	7.4014	0.2775	0.0405	72.5531	10.3718
Grad-CAM + bin	0.6468	13.7500	2.4820	0.5950	0.3117	42.4258	6.8341
LIME + bin	0.7802	40.0000	20.3605	0.8275	0.2912	16.9393	9.8295
SHAP + sigmoid	0.5248	2.0000	0.6522	0.4150	0.0011	68.3246	10.3755

примесей. В то же время, на интерпретациях примеров, к которым был добавлен звук копыт и ржания лошади, метод Saliency лучше сработал с маской minmax. Точность модели значительно стала ниже, а сами интерпретации сложные и недостаточно разреженные. При этом, несмотря на в целом низкую степень положительного влияния маскирования (AI и AD), фактический прирост уверенности (AG) выше, чем для маски topK_50_pos в двух других сценариях.

Поведение Grad-CAM+bin на примерах с шумом проявляется иначе, чем на чистой выборке. Показатели FF, AI, AD и FidIn ухудшились, что в целом говорит о сниженной тенденции приближения целевого классификатора. Тем не менее, показатель AG оказался немного выше, выросли одновременно сложность и разреженность интерпретаций. Можно заключить, что интерпретации, несмотря на общую размытость, по-прежнему содержат “пики”, которые в сценариях с шумом имеют больший вес, чем в сценарии без шума.

Сочетание LIME с бинарной маской вновь оказывается наилучшим относительно метрик согласованности с моделью, но интерпретации слишком сложные и недостаточно прицельные. Маскирование приводит к высокому приросту уверенности и низкому по сравнению с другими методами ее снижению. Судя по показателям AI и AG, метод лучше всего справился с

отделением важных признаков при загрязнении прерывистыми звуками лошади.

Метод SHAP в сочетании с sigmoid-маской вновь показывает низкие результаты относительно всех метрик. Интерпретации не отражают важных для модели признаков, получаются сложными и нелаконичными. В сценариях с шумом улучшились показатели FF и AI, а для непрерывных шумов – AG, однако в контексте общих негативных тенденций такие наблюдения вряд ли указывают на значимые и неслучайные улучшения.

Можно сказать, что общие тенденции могут отличаться в зависимости от того, был ли загрязняющий шум непрерывным или прерывистым.

На Рис. 2 представлены примеры интерпретаций для спектрограммы класса cat. Первый ряд иллюстрирует интерпретации для модели STFT-dB на примере без примесей. Второй ряд содержит интерпретации для модели Mel-dB на примере без примесей. Последующие ряды содержат интерпретации для модели Mel-dB на примере с тремя различными фоновыми примесями. В репозитории [21] доступен полный набор визуальных и прослушиваемых интерпретаций для данного примера.

I. Заключение

В работе рассмотрен способ построения апостериорных интерпретаций моделей, классифицирующих аудиоданные, в визуальной и прослушиваемой форме, посредством выделения наиболее важных для предсказания участков спектрограмм, маскирования и восстановления сигнала. Эффективность подхода проверена на тестовой выборке без примесей, а также на данных, к которым были добавлены фоновые шумы. Результаты показали, что на текущем этапе рассматриваемый подход ограничен с точки зрения сохранения разумного компромисса между высокой степенью согласованности с моделью и лаконичностью генерируемых интерпретаций.

II. Обсуждение

A. *Основные выводы*

Метод SHAP не подходит для решения задачи и требует модификаций для получения более качественных результатов. Метод Saliency в рамках рассмотренного подхода дает простые, но недостаточно сфокусированные интерпретации. Метод LIME оказался лучшим с точки зрения согласованности с моделью, но интерпретации получаются сложными и избыточными. Метод Grad-CAM продемонстрировал промежуточные результаты: он выделяет меньше посторонних компонентов, чем Saliency, но уступает LIME по степени приближения модели. При добавлении примесей к входным данным общие тренды остаются неизменными, но характер загрязняющего шума влияет на поведение модели и, следовательно, на особенности интерпретаций.

B. *Перспективы дальнейших исследований*

На основе результатов работы и существующих ограничений можно выделить следующие направления будущих исследований:

- Модификация методов интерпретаций с учетом особенностей частотно-временных представлений

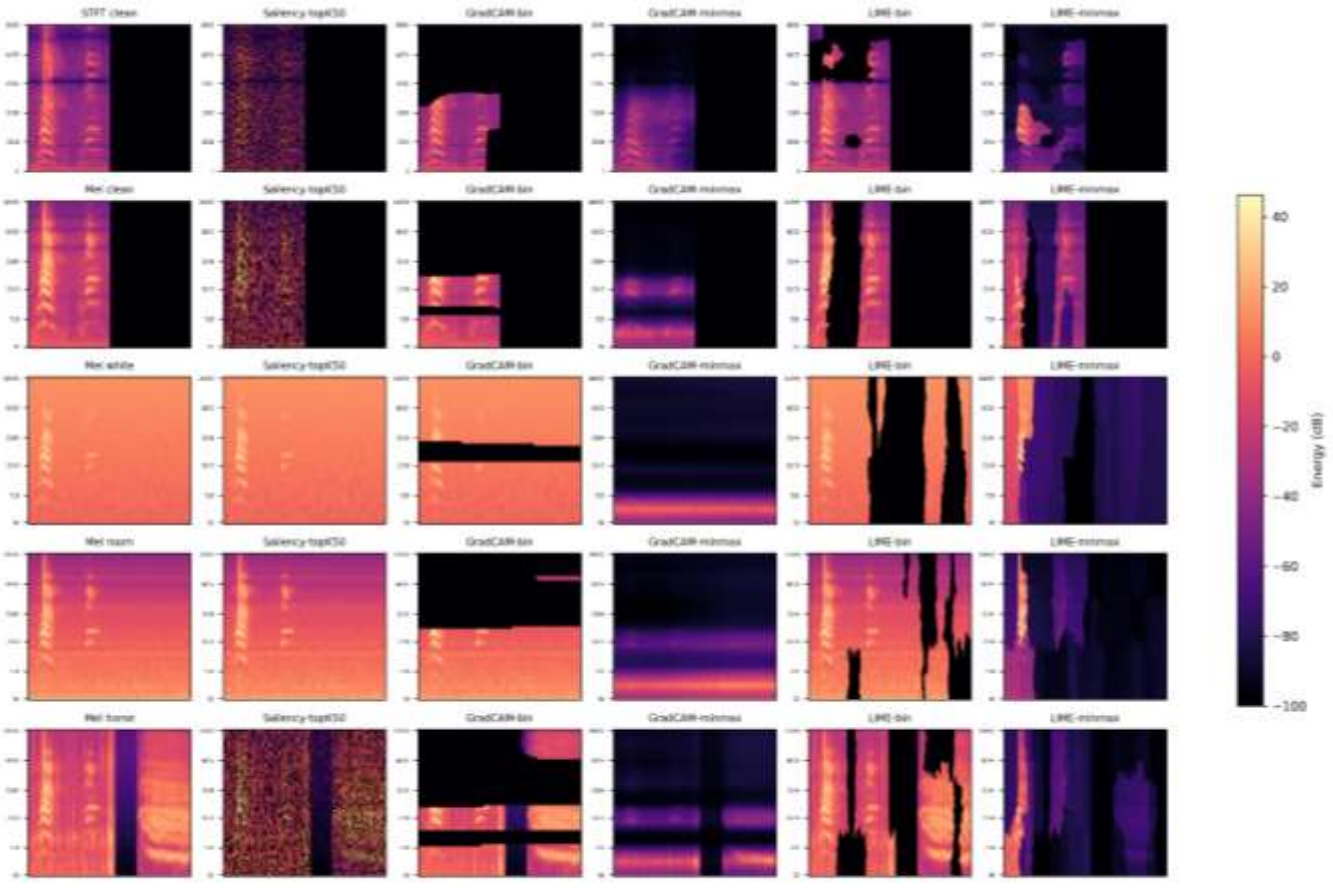

Рис. 2: Примеры визуальных интерпретаций для спектрограмм класса cat.

аудиосигналов (оси частот и временных шагов отражают физические свойства сигнала, а не пространственные координаты);

- Разработка универсальных методов построения прослушиваемых интерпретаций, не зависящих от типа представления входных данных;
- Проведение анализа качества восстановленного аудио с помощью набора объективных метрик или субъективных оценок слушателей;
- Распространение подхода на другие типы данных (музыка, речь, акустические сцены) и другие типы задач (многоклассовая классификация с пересекающимися классами, сегментация аудио).

Библиография

[1] Z. Bai, X.-L. Zhang, *Speaker recognition based on deep learning: An overview*. Neural Networks, 2021, vol. 140, pp. 65–99. DOI: 10.1016/j.neunet.2021.03.004.

[2] A. Srivastava, S. Jain, R. Miranda, S. Patil, S. Pandya, K. Kotecha, *Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease*. PeerJ Computer Science, 2021, vol. 7, p. e369. DOI: 10.7717/peerj-cs.369.

[3] E. Dufourq, C. Batist, R. Foquet, I. Durbach, *Passive acoustic monitoring of animal populations with transfer learning*. Ecological Informatics, 2022, vol. 70, p. 101688. DOI: 10.1016/j.ecoinf.2022.101688.

[4] L. K. D. Katsis, A. P. Hill, E. Piña-Covarrubias, P. Prince, A. Rogers, C. P. Doncaster, J. L. Snaddon, *Automated detection of gunshots in tropical forests using convolutional neural networks*. Ecological Indicators, 2022, vol. 141, p. 109128. DOI: 10.1016/j.ecolind.2022.109128.

[5] F. Doshi-Velez, B. Kim, *Towards a rigorous science of interpretable machine learning*, 2017, url: https://arxiv.org/abs/1702.08608. DOI: 10.48550/arXiv.1702.08608.

[6] C. Molnar, *Interpretable machine learning: A guide for making black-box models explainable*. Leanpub, 2020. ISBN: 978-0244768522.

[7] A. V. Oppenheim, R. W. Schafer, *Discrete-time signal processing*. Prentice Hall, 1989. ISBN: 978-0132162920.

[8] A. Akman, B. W. Schuller, *Audio explainable artificial intelligence: A review*. Intelligent Computing, 2024, vol. 3, p. 0074. DOI: 10.34133/icomputing.0074.

[9] A. N. Zereen, A. Das, J. Uddin, *Machine fault diagnosis using audio sensors data and explainable AI techniques-LIME and SHAP*. Computers, Materials and Continua, 2024, vol. 80, no. 3, pp. 3463-3484. DOI: 10.32604/cmc.2024.054886.

[10] M. T. Ribeiro, S. Singh, C. Guestrin, *"Why should I trust you?": Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[11] S. M. Lundberg, S. I. Lee, *A unified approach to interpreting model predictions*. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4768-4777.

[12] J. Kim, J. Oh, T. Y. Heo, *Acoustic scene classification and visualization of beehive sounds using machine learning algorithms and Grad-CAM*. Mathematical Problems in Engineering, 2021, vol. 2021, no. 1, p. 5594498. DOI: 10.1155/2021/5594498.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, *Grad-CAM: Visual explanations from deep networks via gradient-based localization*. Proceedings of the IEEE international conference on computer vision, 2017, pp. 618-626. DOI: 10.1109/ICCV.2017.74.

[14] J. Vielhaben, S. Lapuschkin, G. Montavon, W. Samek, *Explainable AI for time series via virtual inspection layers*. Pattern Recognition, vol. 150, p. 110309. DOI: 10.1016/j.patcog.2024.110309.

[15] A. Wullenweber, A. Akman, B. W. Schuller, *CoughLIME: Sonified explanations for the predictions of COVID-19 cough classifiers*. 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022, pp. 1342-1345. DOI: 10.1109/EMBC48229.2022.9871291.

[16] V. Haunschmid, E. Manilow, G. Widmer, *audioLIME: Listenable explanations using source separation*, 2020. url: https://arxiv.org/abs/2008.00582. DOI: 10.48550/arXiv.2008.00582.

[17] S. Mishra, B. L. Sturm, S. Dixon, *Local interpretable model-agnostic explanations for music content analysis*. ISMIR, 2017, vol. 53, pp. 537-543. DOI: 10.5281/zenodo.1417387.

[18] J. Parekh, S. Parekh, P. Mozharovskyi, F. d'Alché-Buc, G. Richard, *Listen to interpret: Post-hoc interpretability for audio networks with nmf*. Advances in Neural Information Processing Systems, 2022, vol. 35, pp. 35270-35283.

[19] F. Paissan, M. Ravanelli, C. Subakan, *Listenable maps for audio classifiers*, International Conference on Machine Learning (ICML), 2024.

[20] K. Simonyan, A. Vedaldi, A. Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps*, 2013. url: https://arxiv.org/abs/1312.6034. DOI: 10.48550/arXiv.1605.01713.

[21] Y. Pak, *Constructing Post-hoc Interpretations for Audio Classification Models*, 2025. Available: https://github.com/exile8/audio-xai.

[22] A. Shrikumar, P. Greenside, A. Kundaje, *Learning important features through propagating activation differences*. Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, vol. 70, pp. 3145-3153.

[23] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, *Not just a black box: Learning important features through propagating activation differences*, 2016. url: https://arxiv.org/abs/1605.01713.

[24] K. J. Piczak, *ESC: Dataset for environmental sound classification*. Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018. DOI: 10.1145/2733373.2806390.

[25] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, *PANNs: Large-scale pretrained audio neural networks for audio pattern recognition*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, vol. 28, pp. 2880-2894. DOI: 10.1109/TASLP.2020.3030497.

[26] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore et al., *Audio Set: An ontology and human-labeled dataset for audio events*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776-780. DOI: 10.1109/ICASSP.2017.7952261.

Constructing post-hoc interpretations for audio classification models

Yu. R. Pak, I. Yu. Teryokhina

***Аннотация*—This paper discusses the task of constructing interpretations for machine learning models that classify audio data. The proposed approach enables the construction of interpretations in both visual and listenable forms by masking spectrograms based on feature attribution maps and subsequently reconstructing the signal. To generate the feature attribution maps, the methods Saliency, Grad-CAM, LIME, and SHAP are employed. These methods are universal and can be applied to various architectures. The effectiveness of the approach is evaluated in terms of the fidelity of interpretations to the model's behavior and their perceptual simplicity. Experiments were conducted with different types of masks as well as with the addition of background noise. The results demonstrate that the main challenge of the proposed approach is achieving a compromise between accurately reflecting the model's behavior and producing simple, interpretable explanations. While the addition of noise does not change global trends, the type of noise affects model behavior and, consequentially, the characteristics of the corresponding interpretations.**

***Ключевые слова*—audio data classification, interpretable machine learning, post-hoc interpretation, spectrogram.**

REFERENCES

[1] Z. Bai, X.-L. Zhang, *Speaker recognition based on deep learning: An overview*. Neural Networks, 2021, vol. 140, pp. 65–99. DOI: 10.1016/j.neunet.2021.03.004.

[2] A. Srivastava, S. Jain, R. Miranda, S. Patil, S. Pandya, K. Kotecha, *Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease*. PeerJ Computer Science, 2021, vol. 7, p. e369. DOI: 10.7717/peerj-cs.369.

[3] E. Dufourq, C. Batist, R. Foquet, I. Durbach, *Passive acoustic monitoring of animal populations with transfer learning*. Ecological Informatics, 2022, vol. 70, p. 101688. DOI: 10.1016/j.ecoinf.2022.101688.

[4] L. K. D. Katsis, A. P. Hill, E. Piña-Covarrubias, P. Prince, A. Rogers, C. P. Doncaster, J. L. Snaddon, *Automated detection of gunshots in tropical forests using convolutional neural networks*. Ecological Indicators, 2022, vol. 141, p. 109128. DOI: 10.1016/j.ecolind.2022.109128.

[5] F. Doshi-Velez, B. Kim, *Towards a rigorous science of interpretable machine learning*, 2017, url: https://arxiv.org/abs/1702.08608. DOI: 10.48550/arXiv.1702.08608.

[6] C. Molnar, *Interpretable machine learning: A guide for making black-box models explainable*. Leanpub, 2020. ISBN: 978-0244768522.

[7] A. V. Oppenheim, R. W. Schafer, *Discrete-time signal processing*. Prentice Hall, 1989. ISBN: 978-0132162920.

[8] A. Akman, B. W. Schuller, *Audio explainable artificial intelligence: A review*. Intelligent Computing, 2024, vol. 3, p. 0074. DOI: 10.34133/icomputing.0074.

[9] A. N. Zereen, A. Das, J. Uddin, *Machine fault diagnosis using audio sensors data and explainable AI techniques-LIME and SHAP*. Computers, Materials and Continua, 2024, vol. 80, no. 3, pp. 3463-3484. DOI: 10.32604/cmc.2024.054886.

[10] M. T. Ribeiro, S. Singh, C. Guestrin, *"Why should I trust you?": Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. DOI: 10.1145/2939672.2939778.

[11] S. M. Lundberg, S. I. Lee, *A unified approach to interpreting model predictions*. Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4768-4777.

[12] J. Kim, J. Oh, T. Y. Heo, *Acoustic scene classification and visualization of beehive sounds using machine learning algorithms and Grad-CAM*. Mathematical Problems in Engineering, 2021, vol. 2021, no. 1, p. 5594498. DOI: 10.1155/2021/5594498.

[13] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, *Grad-CAM: Visual explanations from deep networks via gradient-based localization*. Proceedings of the IEEE international conference on computer vision, 2017, pp. 618-626. DOI: 10.1109/ICCV.2017.74.

[14] J. Vielhaben, S. Lapuschkin, G. Montavon, W. Samek, *Explainable AI for time series via virtual inspection layers*. Pattern Recognition, vol. 150, p. 110309. DOI: 10.1016/j.patcog.2024.110309.

[15] A. Wullenweber, A. Akman, B. W. Schuller, *CoughLIME: Sonified explanations for the predictions of COVID-19 cough classifiers*. 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2022, pp. 1342-1345. DOI: 10.1109/EMBC48229.2022.9871291.

[16] V. Haunschmid, E. Manilow, G. Widmer, *audioLIME: Listenable explanations using source separation*, 2020. url: https://arxiv.org/abs/2008.00582. DOI: 10.48550/arXiv.2008.00582.

[17] S. Mishra, B. L. Sturm, S. Dixon, *Local interpretable model-agnostic explanations for music content analysis*. ISMIR, 2017, vol. 53, pp. 537-543. DOI: 10.5281/zenodo.1417387.

[18] J. Parekh, S. Parekh, P. Mozharovskyi, F. d'Alché-Buc, G. Richard, *Listen to interpret: Post-hoc interpretability for audio networks with nmf*. Advances in Neural Information Processing Systems, 2022, vol. 35, pp. 35270-35283.

[19] F. Paissan, M. Ravanelli, C. Subakan, *Listenable maps for audio classifiers*, International Conference on Machine Learning (ICML), 2024.

[20] K. Simonyan, A. Vedaldi, A. Zisserman, *Deep inside convolutional networks: Visualising image classification models and saliency maps*, 2013. url: https://arxiv.org/abs/1312.6034. DOI: 10.48550/arXiv.1605.01713.

[21] Y. Pak, *Constructing Post-hoc Interpretations for Audio Classification Models*, 2025. Available: https://github.com/exile8/audio-xai.

[22] A. Shrikumar, P. Greenside, A. Kundaje, *Learning important features through propagating activation differences*. Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, vol. 70, pp. 3145-3153.

[23] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, *Not just a black box: Learning important features through propagating activation differences*, 2016. url: https://arxiv.org/abs/1605.01713.

[24] K. J. Piczak, *ESC: Dataset for environmental sound classification*. Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015–1018. DOI: 10.1145/2733373.2806390.

[25] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, M. D. Plumbley, *PANNs: Large-scale pretrained audio neural networks for audio pattern recognition*. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, vol. 28, pp. 2880-2894. DOI: 10.1109/TASLP.2020.3030497.

[26] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore et al., *Audio Set: An ontology and human-labeled dataset for audio events*. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 776-780. DOI: 10.1109/ICASSP.2017.7952261.