УДК 004.852:621.391

Д. Д. Чернышов

МИРЭА — Российский технологический университет, Москва, Российская Федерация

А. А. Сухов

МИРЭА — Российский технологический университет, Москва, Российская Федерация

Е. Е. Юдина

Московский физико-технический институт (национальный исследовательский университет), Долгопрудный, Российская Федерация

УСТОЙЧИВАЯ К ВРЕМЕННОЙ РАССИНХРОНИЗАЦИИ ПАРАМЕТРИЧЕСКАЯ ИДЕНТИФИКАЦИЯ НЕЛИНЕЙНОГО АУДИОЭФФЕКТА НА ОСНОВЕ СИАМСКОЙ СЕТИ И ДИФФЕРЕНЦИРУЕМОГО НЕЙРОСЕТЕВОГО СУРРОГАТА

Аннотация. Рассматривается задача автоматического восстановления управляющих параметров нелинейного аудиоэффекта по паре исходного и обработанного сигналов. Предложена процедура параметрической идентификации, объединяющая сиамский оценщик с общим кодировщиком, позднее слияние признаков и замороженный дифференцируемый нейросетевой суррогат целевого процессора. Ключевой элемент процедуры — асинхронно-синхронная стратегия обучения: оценщик анализирует расширенные окна со случайным временным смещением, а аудиодоменная функция потерь вычисляется на строго синхронизированных фрагментах. В качестве целевого устройства использована программная модель дисторшна класса Boss DS-1 с параметрами Distortion, Tone и Level. Обучающие пары сформированы из открытых наборов гитарных и бас-гитарных сигналов; предусмотрены компенсация групповой задержки, зеркальная предыстория, амплитудная аугментация и фильтрация пауз. На отложенной выборке предложенный метод снизил медианную многомасштабную спектральную ошибку с 1,31 до 1,26 по

сравнению с моделью раннего слияния. При искусственной задержке до 1000 мс сиамская архитектура сохраняла стабильное качество реконструкции, тогда как базовая CNN демонстрировала резкий рост ошибки уже при 200 мс. Полученные результаты подтверждают эффективность оптимизации параметров непосредственно по акустическому отклику устройства.

Ключевые слова: параметрическая идентификация, аудиоэффекты, дисторшн, сиамская нейронная сеть, дифференцируемая цифровая обработка сигналов, DDSP, цифровой двойник, анализ через синтез.

Abstract. The paper addresses automatic recovery of control parameters of a nonlinear audio effect from a clean/processed signal pair. The proposed procedure combines a shared-weight Siamese estimator, late feature fusion, and a frozen differentiable neural surrogate of the target processor. Its central component is an asynchronous–synchronous training strategy: the estimator receives extended contexts with random temporal offsets, while the audio-domain loss is evaluated on strictly aligned fragments. A DS-1-type distortion plug-in with Distortion, Tone, and Level controls is used as the target processor. Training pairs are generated from open guitar and bass datasets with latency compensation, reflected signal history, amplitude augmentation, and silence filtering. On the held-out set, the method reduces the median multi-resolution STFT loss from 1.31 to 1.26 compared with an early-fusion CNN. The Siamese estimator preserves stable reconstruction quality for artificial delays up to 1000 ms, whereas the baseline error rises sharply at 200 ms. The results demonstrate the effectiveness of optimizing processor parameters through their acoustic response.

Keywords: parametric identification, audio effects, distortion, Siamese network, differentiable signal processing, DDSP, neural proxy, analysis by synthesis.

1. Введение

Параметрически управляемые аудиоэффекты являются основным инструментом формирования тембра в звукозаписи, музыкальном производстве

и виртуальном аналоговом моделировании. При повторном создании эталонного звучания инженер обычно вручную подбирает положения регуляторов, многократно сравнивая результат с референсным сигналом. Для нелинейных процессоров такая настройка особенно сложна: влияние параметров взаимозависимо, а изменение входной амплитуды переводит устройство между квазилинейным режимом, мягкой сатурацией и жестким ограничением.

Автоматическая параметрическая идентификация переводит данный процесс в задачу обратного моделирования. По исходному сигналу и результату его обработки требуется восстановить интерпретируемый вектор настроек устройства. В отличие от непараметрического профилирования, результатом становится не только имитация тембра, но и набор управляющих значений, который можно перенести в программный или аппаратный процессор.

Современные нейросетевые методы успешно аппроксимируют нелинейные аудиосистемы [2, 3], а свёрточные модели по спектрограммам решают задачи распознавания эффектов и оценки их параметров [4]. Развитие дифференцируемой цифровой обработки сигналов сделало возможным обучение оценщика по ошибке в аудиодомене [5–10]. Для практической студийной работы определяющей остаётся устойчивость к временному рассогласованию сигналов, возникающему из-за задержки плагина, передискретизации, аппаратной маршрутизации и неточного совмещения записей.

Цель исследования — повысить устойчивость и акустическую точность параметрической идентификации нелинейного аудиоэффекта при наличии временной рассинхронизации исходного и целевого сигналов.

Научная новизна работы определяется сочетанием четырёх взаимосвязанных решений: сиамского оценщика с общими весами и поздним слиянием признаков; явного вектора разности латентных представлений; асинхронно-синхронной схемы формирования обучающих окон; динамически

взвешиваемой функции потерь, вычисляемой через дифференцируемый нейросетевой суррогат процессора. Экспериментально исследована робастность процедуры к задержкам в диапазоне 0–1000 мс.

Практическая значимость состоит в возможности автоматически переносить настройки Distortion, Tone и Level по паре аудиозаписей, сохраняя спектральную структуру, микродинамику и общий уровень выходного сигнала.

2. Связанные работы

Методы моделирования аудиосистем традиционно подразделяются на физические модели «белого ящика», аппроксиматоры «чёрного ящика» и гибридные модели. Физический подход воспроизводит электрическую схему процессора и обеспечивает прямую интерпретируемость внутренних параметров, однако требует подробной схемотехнической информации и значительных вычислительных затрат [1]. Нейросетевые аппроксиматоры обучаются на парах «вход–выход» и эффективно моделируют ламповые усилители, дисторшны и динамические эффекты [2, 3].

Comunità, Stowell и Reiss показали результативность двумерных свёрточных сетей при распознавании гитарных эффектов и оценке непрерывных управляющих параметров [4]. Такой подход использует спектрограмму как двумерную карту характерных гармонических и динамических признаков. При раннем объединении исходного и обработанного сигналов сеть получает сильный локальный признак их взаимного соответствия, но одновременно становится чувствительной к временной позиции спектральных событий.

DDSP интегрирует синтез или обработку сигнала в граф автоматического дифференцирования [5]. Если аналитическая реализация эффекта неизвестна, её заменяют нейронным прокси, обученным имитировать зависимость выхода от входа и управляющих параметров [6–8]. Steinmetz, Bryan и Reiss применили общий спектральный кодировщик и аудиодоменную оптимизацию для переноса стиля обработки [8]. Grant исследовал сиамскую архитектуру для

управления недифференцируемыми аудиоэффектами [9], а Peladeau и Peeters показали преимущество автоэнкодерной постановки с оптимизацией качества восстановленного аудио [10].

Предлагаемая процедура развивает направление «анализ через синтез» применительно к нелинейному дисторшну и фокусируется на инвариантности к задержке. Оценщик получает независимые контекстные представления сигналов, а корректность параметров проверяется посредством синтеза на строго выровненном коротком окне.

3. Постановка задачи

Пусть $x \in \mathbb{R}^N$ — фрагмент исходного аудиосигнала, F — целевой процессор, а $p^* = [pD, pT, pL]^T \in [0,1]^3$ — нормализованные положения регуляторов Distortion, Tone и Level. Обработанный сигнал задаётся выражением

$$y = F(x, p^*). \quad (1)$$

В реальном тракте наблюдаемый целевой сигнал может быть смещён во времени:

$$y_\tau[n] = y[n - \tau], \quad (2)$$

где τ — неизвестная задержка. Требуется построить оценщик E_θ, возвращающий параметры $\hat{p} = E_\theta(x, y_\tau)$, для которых результат повторного синтеза $F(x, \hat{p})$ акустически близок к y. Поскольку целевой VST-процессор не входит в граф дифференцирования, используется нейросетевой суррогат S_φ:

$$\hat{y} = S_\varphi(x, \hat{p}) \approx F(x, \hat{p}). \quad (3)$$

Оптимизация осуществляется не только по расстоянию между $\hat{p}$ и p^*, но и по различию между y и $\hat{y}$. Такая постановка учитывает неравномерную чувствительность тембра к отдельным регуляторам и допускает акустически эквивалентные комбинации настроек.

4. Предлагаемая процедура идентификации

4.1. Формирование обучающих данных

Исходным материалом служили открытые наборы GuitarSet, IDMT-SMT-Guitar, IDMT-SMT-Bass и IDMT-SMT-Audio-Effects [12–15]. Они охватывают одиночные ноты, аккорды, медиаторную и пальцевую технику, слайды, бенды, вибрато, приглушённые звуки, электрогитарные и бас-гитарные тембры. Такое сочетание обеспечивает разнообразие спектрального состава и переходных процессов.

Аудиофайлы приводились к монофоническому представлению с частотой дискретизации 44,1 кГц и сохранялись в формате 32-bit Float WAV. Для автоматической генерации целевых данных применялась программно управляемая VST-модель bx_distorange, воспроизводящая структуру дисторшна класса Boss DS-1. Значения трёх регуляторов стохастически выбирались в полном рабочем диапазоне; сформированный корпус содержал десятки тысяч парных аудиофрагментов.

Каждый двухсекундный фрагмент дополнялся зеркально отражённой предысторией. Она создаёт C^0-непрерывную границу и переводит внутренние состояния рекурсивных фильтров в установившийся режим до начала полезного окна. Собственная групповая задержка VST определялась по импульсному отклику и компенсировалась поотсчётным сдвигом целевого сигнала.

Для покрытия нелинейной амплитудной характеристики использовалось случайное масштабирование входа в диапазоне от −12 до +12 дБ. Фрагменты с пиковой амплитудой менее 0,005 исключались детектором активности. Это концентрирует обучение на информативных атаках и затуханиях и предотвращает аппроксимацию случайного фонового шума.

Таблица 1 — Параметры формирования данных и спектрального представления

Параметр	Значение
Частота дискретизации	44 100 Гц
Формат хранения	WAV, 32-bit Float, mono
Длина базового фрагмента	2 с
Контекст сиамского оценщика	131 072 отсчёта (≈ 2,97 с)
Амплитудная аугментация	от −12 до +12 дБ
Порог активности	0,005 по пиковой амплитуде
STFT	n_fft = 1024, hop = 512
Мел-представление	256 полос, диапазон −80…+20 дБ
Управляющие параметры	Distortion, Tone, Level

4.2. Базовая модель раннего слияния

В качестве базовой архитектуры использована Distortion2DNet. Мел-спектрограммы исходного и целевого сигналов объединяются вдоль оси каналов в тензор [B, 2, F, T]. Экстрактор содержит четыре свёрточных блока с 32, 64, 128 и 256 каналами. В каждом блоке применяются свёртка 3×3, Instance Normalization, ReLU и пространственная субдискретизация. После глобального адаптивного усреднения 256-мерный вектор поступает в регрессионную голову

из полносвязного слоя на 128 нейронов, Dropout 0,3 и выходного сигмоидального слоя.

Модель обучается прямой регрессией по L1-ошибке нормализованных параметров. Раннее слияние обеспечивает точное сопоставление синхронных спектральных паттернов и формирует содержательную базовую линию для оценки устойчивости к временным сдвигам.

4.3. Дифференцируемый нейросетевой суррогат

Суррогат DistortionModel работает непосредственно с временной формой сигнала и аппроксимирует зависимость F(x,p). Основу сети образуют 27 остаточных блоков: девять коэффициентов расширения от 1 до 256 повторяются тремя циклами. Расширенные одномерные свёртки формируют широкое рецептивное поле, необходимое для моделирования нелинейной фильтрации и переходных процессов.

Управляющий вектор внедряется в каждый остаточный блок посредством FiLM модуляции [18]. Для карт признаков h слой вычисляет коэффициенты γ(p) и β(p):

$$FiLM(h \mid p) = \gamma(p) \odot h + \beta(p). \quad (4)$$

Суррогат предварительно обучается по сумме многомасштабной спектральной ошибки и L1-ошибки во временной области. После сходимости параметры φ фиксируются. Замороженная модель сохраняет дифференцируемость по управляющему вектору и передаёт градиент акустической ошибки к оценщику.

4.4. Сиамский оценщик с поздним слиянием

SiameseParamEstimator независимо кодирует исходный и целевой сигналы одним свёрточным экстрактором с общими весами. Общий кодировщик проецирует оба сигнала в единое латентное пространство, поэтому одинаковые спектральные структуры получают сопоставимые представления независимо от их абсолютной временной позиции.

После свёрточных слоёв используется пространственное внимание на основе свёртки 1×1 и сигмоидальной активации. Карта внимания подавляет участки тишины и усиливает информативные транзиенты. Вместо глобального сведения временной оси к одной точке применяется адаптивная темпоральная подвыборка с двумя выходными шагами, сохраняющая различие между атакой и установившейся частью звука.

Латентные векторы Vx и Vy, а также их направленная разность ΔV = Vy − Vx конкатенируются в 768-мерное представление. Разность интерпретируется как компактное описание преобразования, внесённого процессором. Регрессионная голова с LayerNorm и LeakyReLU отображает объединённый вектор в диапазон $[0,1]^3$.

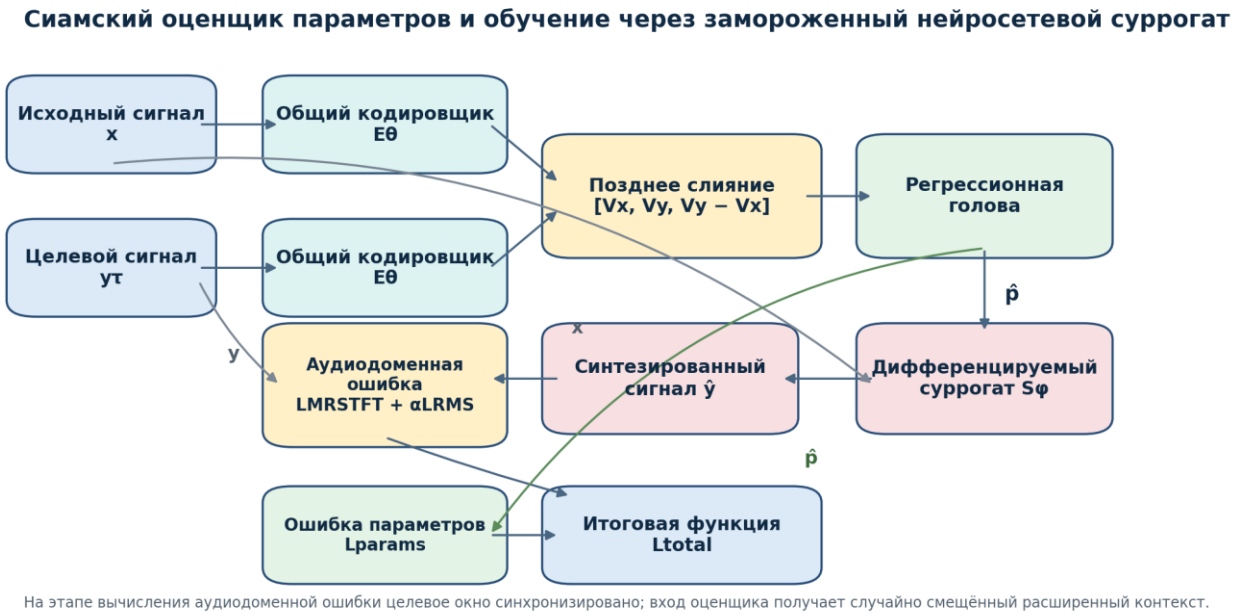

Рисунок 1 — Архитектура сиамского оценщика и обучение через дифференцируемый суррогат

4.5. Асинхронно-синхронная стратегия обучения

Для оценщика и функции потерь используются разные временные представления одной обучающей пары. Входные ветви оценщика получают расширенные контекстные окна; целевое окно случайно смещается относительно исходного. Сеть вынуждена извлекать глобальные признаки тембра и динамики, не используя попиксельное совпадение спектрограмм.

Одновременно из той же пары выделяется короткое строго синхронизированное окно. Предсказанные параметры подаются вместе с исходным синхронным фрагментом в суррогат, а его выход сравнивается с выровненным целевым сигналом. Таким образом, инвариантность к задержке формируется на входе оценщика, а акустическая точность контролируется по физически корректной паре.

Многомасштабная спектральная функция потерь вычисляется для наборов окон 2048, 1024, 512 и 256 отсчётов и объединяет спектральную сходимость и расстояние логарифмических магнитуд [19]:

$$LMRSTFT = (1/M)\ \Sigma m\ [SCm(y, \hat{y}) + \|log|STFTm(y)| - log|STFTm(\hat{y})|\|_1]. \quad (5)$$

Для точного восстановления параметра Level добавляется ошибка среднеквадратичной огибающей LRMS. Параметрическое расстояние используется как направляющий регуляризатор:

$$Ltotal = LMRSTFT + \alpha LRMS + \lambda e\|p^* - \hat{p}\|^2_2, \quad \lambda e = \lambda_0 exp(-ke). \quad (6)$$

В начале обучения параметрический компонент формирует устойчивое приближение, после чего экспоненциальное уменьшение λe переводит оптимизацию на акустический критерий. Масштабирующий коэффициент RMS-компонента принят равным 1000, что уравновешивает вклад спектральной формы и абсолютного уровня.

5. Экспериментальная установка

Программная реализация выполнена на Python и PyTorch. Обучение проводилось на компьютере с процессором AMD Ryzen 7 6800H, 16 ГБ оперативной памяти и графическим ускорителем NVIDIA RTX 3070 Ti. Сгенерированные данные разделялись на обучающую часть и отложенную тестовую выборку в пропорции 90/10; контрольные точки выбирались по минимуму функции потерь на валидационных данных.

Таблица 2 — Основные гиперпараметры обучения

Модель	Оптимизатор	Скорость обучения	Пакет	Режим
Distortion2DNet	Adam	$5 \cdot 10^{-5}$	128	L1, ReduceLROnPlateau
DistortionModel	AdamW	$3 \cdot 10^{-4}$	32	weight decay 10^{-4}, AMP
SiameseParamEstimator	AdamW	$2 \cdot 10^{-5}$	8	150 эпох, CosineAnnealingLR

Качество оценивалось в двух пространствах. Параметрическая метрика MAE вычислялась после перевода выходов из нормализованного диапазона в шкалу регуляторов 0–10. Акустическая метрика MRSTFT определялась после повторного рендеринга предсказанных настроек через целевой VST-процессор. Дополнительно анализировались АЧХ, форма волны, RMS-уровень и устойчивость при искусственной задержке целевого канала от 0 до 1000 мс.

6. Результаты

6.1. Точность прямой регрессии и нейросетевого суррогата

Distortion2DNet демонстрировала устойчивую сходимость: нормализованная MAE обучения и валидации стабилизировалась около 0,09. На синхронных фрагментах модель точно восстанавливала спектральный контур и динамическую огибающую. Исследование разреженных партий показало, что глобальное усреднение сглаживает короткие атаки и может снижать оценку выходного уровня, что обосновывает использование внимания и двухшаговой темпоральной агрегации.

Обученный DistortionModel воспроизводил форму клиппинга, расположение высших гармоник и фазовую структуру сигнала. Синхронное изменение обучающей и валидационной MRSTFT подтвердило устойчивую

аппроксимацию целевого VST. Полученная точность позволила использовать сеть как замороженный акустический генератор в контуре обучения оценщика.

6.2. Сравнение на отложенной выборке

Таблица 3 — Сравнение точности параметров и качества повторного синтеза

Модель	Медианная MAE параметров, шкала 0–10	Медианная MRSTFT
Distortion2DNet	0,70	1,31
SiameseParamEstimator + surrogate	0,75	1,26

Прямая CNN дала минимальную численную ошибку координат регуляторов. Сиамская модель, оптимизированная через акустический отклик, снизила медианную MRSTFT с 1,31 до 1,26, то есть на 3,8 %. Полученный эффект согласуется с неединственностью обратного отображения: близкие тембры могут формироваться различными комбинациями Distortion, Tone и Level. Аудиодоменная оптимизация выбирает комбинацию, обеспечивающую наилучший повторный синтез, а не только минимальное евклидово расстояние в пространстве регуляторов.

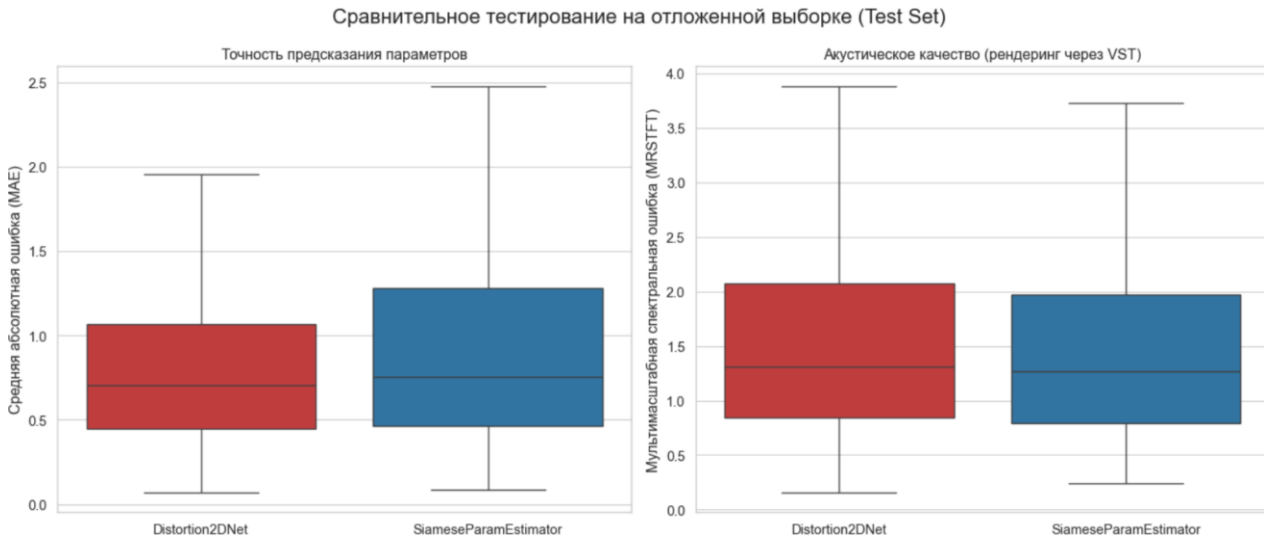

Рисунок 2 — Распределение параметрической и многомасштабной спектральной ошибок на отложенной выборке

6.3. Устойчивость к временной задержке

При рассинхронизации раннее слияние теряет локальное соответствие между каналами. Уже при задержке 50 мс параметрическая ошибка базовой модели возрастала примерно в шесть раз, а при 200 мс MRSTFT достигала около 6,4. Сиамский оценщик сохранял MRSTFT в узком диапазоне и при задержке 1000 мс обеспечивал восстановление спектрального контура и RMS-уровня.

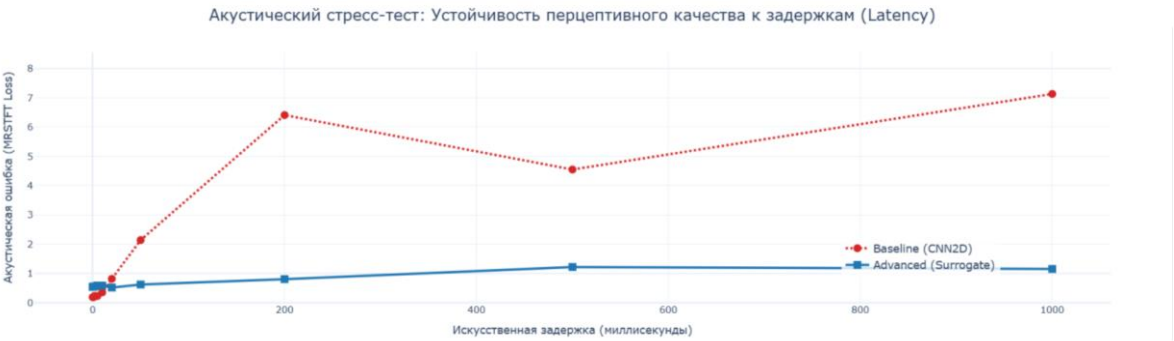

Рисунок 3 — Устойчивость акустического качества к искусственной временной задержке

Характер кривых показывает, что устойчивость обусловлена не компенсацией конкретного фиксированного сдвига, а переходом к глобальным тембральным представлениям. Общий кодировщик и позднее слияние сохраняют смысл разности признаков даже при отсутствии фазового совпадения отдельных атак.

6.4. Вычислительная эффективность

Таблица 4 — Вычислительная сложность нейросетевых компонентов

Компонент	Параметров	Обучаемых параметров	Теоретическая память
Distortion2DNet	422 371	422 371	≈ 23,15

Компонент	Параметров	Обучаемых параметров	Теоретическая память
			МБ
DistortionModel	206 294	206 294	≈ 932,76 МБ
SiameseParamEstimator	558 820	558 820	≈ 5,32 МБ

Каждый модуль содержит менее 600 тыс. параметров. На этапе применения веса суррогата не участвуют в построении графа, поэтому фактическое пиковое потребление видеопамяти не превышало 300 МБ. Время обработки единичного фрагмента на CPU составляло 0,0867 с. Первый запуск на GPU занимал 1,04 с из-за инициализации CUDA, после чего задержка уменьшалась благодаря кэшированию.

7. Обсуждение

Результаты демонстрируют различие между точностью восстановления координат регуляторов и точностью восстановления звучания. Для нелинейного процессора отображение p → y не является равномерным: чувствительность к Tone зависит от уровня клиппинга, а влияние Distortion насыщается в области жёсткого ограничения. Поэтому аудиодоменная функция потерь задаёт более содержательную геометрию обратной задачи.

Асинхронно-синхронная организация данных разделяет две цели обучения. Асинхронный контекст формирует инвариантность оценщика, а синхронное окно обеспечивает корректный градиент через суррогат. В результате модель не пытается сопоставлять координаты отдельных спектральных событий, но сохраняет строгий акустический критерий для предсказанных параметров.

Адаптивная агрегация по двум временным шагам сохраняет различие между атакой и последующим затуханием. Пространственное внимание дополнительно снижает вклад пауз и фонового шума. Эти механизмы особенно важны для дисторшна, поскольку плотность высших гармоник определяется мгновенной амплитудой и формой транзиента.

Процедура допускает масштабирование на другие параметрические процессоры. Для нового эффекта требуется сформировать пары «вход–выход–параметры», обучить условный суррогат и изменить размер выходного слоя оценщика. Архитектура кодировщика, позднее слияние, двухоконная схема и аудиодоменная оптимизация сохраняются без принципиальных изменений.

8. Заключение

Предложена процедура параметрической идентификации нелинейного аудиоэффекта, объединяющая сиамский оценщик, позднее слияние признаков, пространственное внимание, адаптивную временную агрегацию и дифференцируемый нейросетевой суррогат. Разработанная асинхронно-синхронная стратегия позволяет обучать инвариантное представление по рассогласованным контекстам и одновременно вычислять акустически корректную функцию потерь.

На отложенной выборке метод снизил медианную многомасштабную спектральную ошибку повторного синтеза на 3,8 % относительно модели раннего слияния. При задержках до 1000 мс сиамская архитектура сохранила устойчивое качество, тогда как базовая CNN демонстрировала резкий рост ошибки при 200 мс. Модель также обеспечила восстановление спектрального контура, микродинамики и общего RMS-уровня.

Компактность нейросетевых компонентов и время применения менее 0,1 с на CPU позволяют использовать процедуру в интерактивных системах звукового дизайна и автоматизированного переноса настроек аудиопроцессоров.

Список литературы

1. Zölzer U. (ed.). DAFX: Digital Audio Effects. 2nd ed. Chichester: Wiley, 2011.

2. Wright A., Damskägg E.-P., Välimäki V. Real-time black-box modelling with recurrent neural networks // Proc. 22nd Int. Conf. on Digital Audio Effects (DAFx-19). Birmingham, 2019.

3. Damskägg E.-P., Juvela L., Thuillier E., Välimäki V. Deep learning for tube amplifier emulation // Proc. IEEE ICASSP. 2019. P. 471–475.

4. Comunità M., Stowell D., Reiss J. D. Guitar Effects Recognition and Parameter Estimation with Convolutional Neural Networks // Journal of the Audio Engineering Society. 2021. Vol. 69, No. 7/8.

5. Engel J., Hantrakul L., Gu C., Roberts A. DDSP: Differentiable Digital Signal Processing // Proc. ICLR. 2020. URL: https://arxiv.org/abs/2001.04643

6. Comunità M., Steinmetz C. J., Reiss J. D. Differentiable Black-Box and Gray-Box Modeling of Nonlinear Audio Effects. 2021. URL: https://arxiv.org/abs/2101.07754

7. Martínez Ramírez M. A., Wang O., Smaragdis P., Bryan N. J. Differentiable Signal Processing with Black-Box Audio Effects // Proc. IEEE ICASSP. 2021. DOI: 10.1109/ICASSP39728.2021.9415103.

8. Steinmetz C. J., Bryan N. J., Reiss J. D. Style Transfer of Audio Effects with Differentiable Signal Processing // Journal of the Audio Engineering Society. 2022. Vol. 70, No. 9. URL: https://arxiv.org/abs/2207.08759

9. Grant K. Style Transfer for Non-Differentiable Audio Effects // Proc. 155th AES Convention. New York, 2023. URL: https://arxiv.org/abs/2309.17125

10. Peladeau C., Peeters G. Blind Estimation of Audio Effects Using an Auto-Encoder Approach and Differentiable Digital Signal Processing // Proc. IEEE ICASSP. 2024. P. 856–860. DOI: 10.1109/ICASSP48485.2024.10448301.

11. Peladeau C., Fourer D., Peeters G. Audio Processor Parameters: Estimating Distributions Instead of Deterministic Values // Proc. 28th Int. Conf. on Digital Audio Effects (DAFx-25). Ancona, 2025.

12. Xi Q., Bittner R. M., Pauwels J., Ye X., Bello J. P. GuitarSet: A Dataset for Guitar Transcription // Proc. 19th ISMIR Conference. Paris, 2018. P. 453–460.

13. Fraunhofer IDMT. IDMT-SMT-Guitar Dataset. URL: https://www.idmt.fraunhofer.de/en/publications/datasets/guitar.html (дата обращения: 25.07.2026).

14. Fraunhofer IDMT. IDMT-SMT-Bass Dataset. URL: https://www.idmt.fraunhofer.de/en/publications/datasets/bass.html (дата обращения: 25.07.2026).

15. Fraunhofer IDMT. IDMT-SMT-Audio-Effects Dataset. URL: https://www.idmt.fraunhofer.de/en/publications/datasets/audio_effects.html (дата обращения: 25.07.2026).

16. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition // Proc. ICLR. 2015. URL: https://arxiv.org/abs/1409.1556

17. Ulyanov D., Vedaldi A., Lempitsky V. Instance Normalization: The Missing Ingredient for Fast Stylization. 2016. URL: https://arxiv.org/abs/1607.08022

18. Perez E., Strub F., de Vries H., Dumoulin V., Courville A. FiLM: Visual Reasoning with a General Conditioning Layer // Proc. AAAI. 2018. P. 3942–3951.

19. Yamamoto R., Song E., Kim J.-M. Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram // Proc. IEEE ICASSP. 2020. P. 6199–6203.

20. Paszke A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems. 2019. Vol. 32.