УДК: 004.8
OECD: 2.02

Современные методы построения систем искусственного интеллекта для обработки аудиосигналов

Лестенко Н.А.[1*], Вальштейн К.В.[2], Верхова А.А.[3]
[1]Преподаватель, [2]Старший преподаватель, [3]Магистрант,
[1,2,3]Балтийский государственный технический университет «ВОЕНМЕХ»
им. Д. Ф. Устинова, г. Санкт-Петербург, Россия

Аннотация

В данной статье рассматриваются различные подходы для обработки аудиосигнала, в том числе на базе технологий искусственного интеллекта для задач распознавания речи, детектирования объектов, генерации речи и шумоподавления. В статье рассматриваются классические математические методы обработки сигнала, такие как быстрое преобразование Фурье (БПФ), мел-кепстральные коэффициенты (MFCC) и дискретное вейвлет-преобразование (ДВП). Вместе с этим рассматриваются другие подходы к обработке аудиосигнала такие, как системы искусственного интеллекта для выделения признаков сигнала, использующие расширенные причинно-следственные сверточные слои при проектировании архитектуры ИНС, примером которых может быть WaveNet, а также более новую технологию – трансформеры, на которых базируются Whisper и Waveformer. Данная статья акцентирует внимание на переходе от традиционных методов к искусственным нейронным сетям с использованием технологии трансформеров и диффузионных моделей, а также приводит пример использования некоторых из рассматриваемых методов для решения различных задач.

Ключевые слова: искусственная нейронная сеть, быстрое преобразование Фурье, вейвлет-преобразование, мел-кепстральные коэффициенты, DCC, Whisper, обработка аудиосигналов.

Modern methods for building artificial intelligence systems for audio signal processing

Lestenko N.A.[1], Valshtein K.V.[2], Verhova A.A.[3]*
[1]Lecturer, [2]Senior Lecturer, [3]Master student,
[1,2,3]Baltic State Technical University 'VOENMEH', St. Petersburg, Russia

Abstract

This article explores various approaches to audio signal processing, including AI-based technologies for tasks such as speech recognition, object detection, speech generation, and noise suppression. The article covers classical mathematical signal processing methods such as the Fast Fourier Transform (FFT), Mel-frequency cepstral coefficients (MFCC), and discrete wavelet transform (DWT). Additionally, other approaches to audio signal processing are considered, such as artificial intelligence systems for feature extraction using dilated causal convolutional layers in neural network architecture design, exemplified by WaveNet, as well as more recent transformer-based technologies like Whisper and Waveformer. The article emphasizes the transition from traditional methods to artificial neural networks using transformer and diffusion model technologies, as well as demonstrating examples of successful using some of methods for different audio processing tasks.

**E-mail: lazarev30_12@mail.ru (Лестенко Н.А.)*

Keywords: *Artificial Neural Network, Fast Fourier Transform (FFT), Wavelet Transform, Mel-frequency Cepstral Coefficients (MFCC), Dilated Causal Convolution (DCC), Whisper, Audio Signal Processing.*

Введение

Технологии искусственного интеллекта (ИИ) активно внедряются в различные процессы уже на протяжении нескольких десятилетий. Всё это время они развивались, переходя от абстрактных теорий принятия решения и обработки данных к созданию сложных мультимодальных интеллектуальных систем. Широкую популярность обрели большие языковые модели (LLM) и диффузионные модели для создания и обработки изображений. Также существует множество моделей ИИ, использующихся для задач, связанных с обработкой аудио информации. Обзор современного состояния подобных моделей и будет представлен в данной статье.

Прежде чем перейти к обзору существующих моделей, следует кратко затронуть виды задач, связанные с обработкой аудио. К таким задачам можно отнести следующие:

– сигнала (преобразование аудиоинформации в иной формат, например – распознавание речи (STT, speech-to-text), либо построение карты помещения [1]);

– детектирование объектов (определение параметров и типа источника аудиосигнала, например – диаризация диктора [2]);

– генерация сигнала (преобразование текста в речь (TTS, text-to-speech), либо создание сигнала на основе полученной от пользователя информации, например – генерация музыки [3]);

– изменение параметров сигнала (обработка и нелинейная фильтрация существующего сигнала, например – выделение и удаление посторонних шумов [4]).

Конкретная задача диктует выбор метода предварительной обработки входного сигнала и ключевые моменты модели для работы с ним. Эти моменты и будут разобраны в данной статье.

1. Методы предобработки входного сигнала

Входной аудиосигнал как правило представляет собой дискретизированную последовательность амплитуд в каждый момент времени. Эти значения абстрактны и сложны для дальнейшей обработки, поэтому для многих задач выполняется дополнительный этап подготовки, заключающийся в выделении основных характеристик обрабатываемого сигнала. Обычно сигнал разбивается на пересекающиеся кадры фиксированной продолжительности, каждый из которых обрабатывается отдельно, при этом степень их пересечения выбирается таким образом, чтобы не допустить образования неполных кадров.

Конкретные методы получения дополнительной информации различаются в зависимости от задачи обработки сигнала. Классическим является алгоритм быстрого преобразования Фурье (БПФ), однако он часто дополняется или заменяется иными инструментами. Так в статье [5] для задач, связанных с обработкой речи, показана эффективность мел-кепстральных коэффициентов – их использование позволило обработать и распознать речь в реальном времени. В статье [6] для обработки речи используется метод Residual Vector Quantization, который получает характеристики аудиосигнала в виде отдельных векторов, использование данного метода позволило достичь естественного звучания генерируемого речевого сигнала. В то же время в статье [7] показана эффективность вейвлет преобразований для решения различных

классов задач – в ней показано что вейвлет преобразования при меньшей вычислительной сложности оказываются столь же эффективны, как и БФТ или мел-кепстральные коэффициенты.

Некоторые из перечисленных методов следует рассмотреть подробнее для большего понимания структуры входного сигнала. Стоит начать с классического алгоритма быстрого преобразования Фурье.

В качестве примера будет использован специально записанный аудиофайл с речью и уличным шумом. Это позволит продемонстрировать эффекты от применения различных методов предобработки сигнала. Общий вид исследуемого сигнала показан на рисунке 1. Следует заметить, что была проведена нормализация амплитуды сигнала с целью удобства обработки. Для построения изображений сигнала с использованием классических методов предобработки была написана отдельная программа на языке Python.

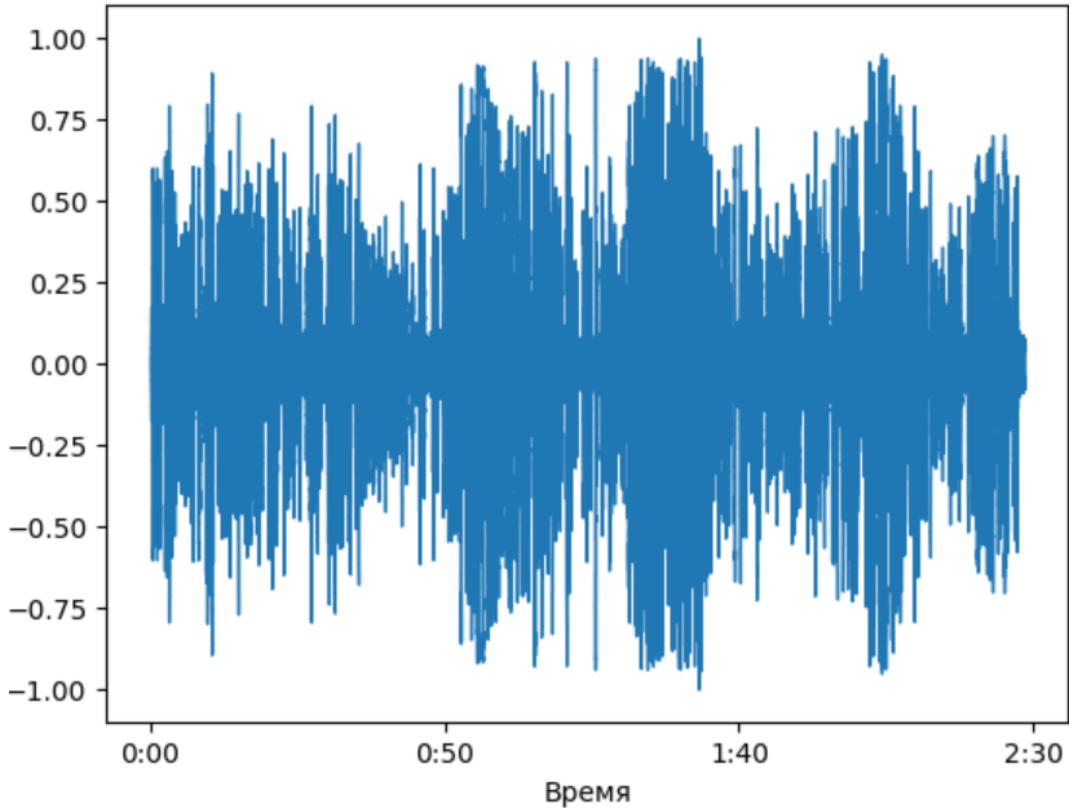


Рис. 1. Исходный сигнал

Классическое преобразование Фурье позволяет разложить сигнал любой формы на составляющие амплитуды и фазового сдвига сигнала, что позволяет определить, какие гармонические колебания и в какой частоте присутствуют в исходном сигнале.

Но ввиду дискретности сигналов реального мира, где сигнал представляется конечной последовательностью чисел во временной и в частотной области, для разложения сигнала на составляющие применяют Дискретное преобразование Фурье (ДПФ).

$$\sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} = \sum_{n=0}^{N-1} x_n \left(\cos\left(\frac{2\pi kn}{N}\right) - i * \sin\left(\frac{2\pi kn}{N}\right)\right), \; k = 0, \ldots, N-1, \qquad (1)$$

где N – количество значений сигнала, измеренных за период, а также количество компонент разложения;

x_n – измеренные значения сигнала;

k – индекс частоты. Частота k-го сигнала равна K/T,

i – мнимая единица,

Так как на выходе преобразования получаются конечные суммы, что подходит для использования в цифровой форме, в частности, для алгоритмов цифровой обработки сигналов. Но данный подход оказывается затратным по вычислительным мощностям, так как его сложность оценивается как $O(N^2)$. Поэтому на практике применяется оптимизированный алгоритм ДПФ, известный как быстрое преобразование Фурье. Вычислительная сложность БПФ оценивается как $O(Nlog(N))$, что делает БПФ особенно полезным для обработки сигналов в реальном времени и работы с большими данными. На выходе после БПФ получается массив комплексных чисел, из данного массива можно извлечь амплитудную и частотную информацию о входном сигнале, а также фазовый спектр сигнала. Результат может быть представлен графически, что отображено на рисунке 2.

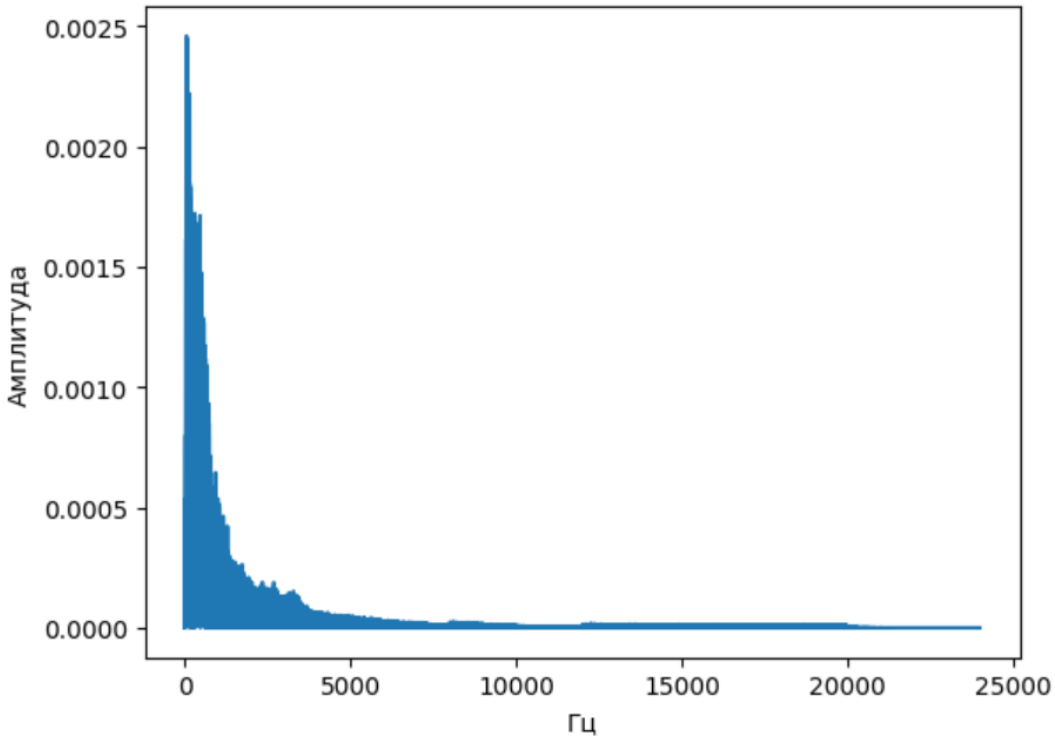


Рис. 2. Преобразование Фурье

Полученные данные служат универсальным представлением исходного сигнала в сжатом виде без потери информации, однако для многих задач предпочтительнее использовать более специализированные механизмы. Так как спектр, полученный с помощью БПФ не содержит временных характеристик сигнала, то для его последовательной обработки часто используют оконное преобразование Фурье (ОПФ, Short Time Fourier Transform – STFT), формирующее спектрограмму сигнала.

По сути, алгоритм ОПФ прост – специальная оконная функция (вид которой зависит от желаемых характеристик) перемножается на фрагмент исходного сигнала и над результатом выполняется БПФ. После этого окно «сдвигается» по входному сигналу и процесс повторяется. Результирующие значения частоты и амплитуды оказываются связаны с конкретной отсечкой по времени и могут полноценно описывать динамику изменения выходного сигнала со временем, при этом используя для хранения информации

меньше данных, нежели исходный аудиосигнал. На основе полученных данных можно построить спектрограмму сигнала, подобную показанной на рисунке 3.

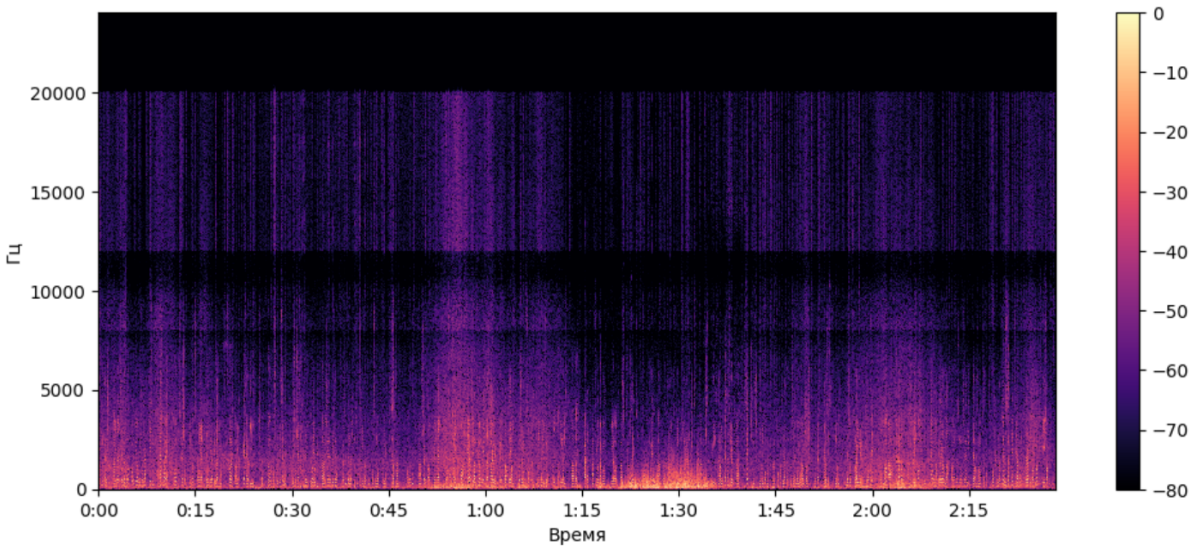


Рис. 3. Спектрограмма аудиосигнала, полученная с помощью ОПФ

Недостатков у STFT несколько: во-первых, вид оконной функции и размер окна не могут изменяться с течением времени, что не позволяет динамически менять разрешающую способность полученной спектрограммы, а во-вторых, невозможно полноценное восстановление сигнала из спектрограммы, что ограничивает использование подобных методов предобработки сигнала. Также для ряда задач использование классических оконных функций оказывается недостаточно, требуется иной алгоритм.

Например, в задачах анализа речи часто используются мел-кепстральные коэффициенты (MFCC), позволяющие построить мел-спектрограмму. В стандартной спектрограмме частотная ось линейна и измеряется в герцах (Гц). Однако слуховая система человека более чувствительна к изменениям на низких частотах, чем на высоких, и эта чувствительность уменьшается логарифмически с увеличением частоты. Шкала Мел - это перцептивная шкала, которая аппроксимирует нелинейную частотную характеристику человеческого уха. Для создания мел-спектрограммы используется ОПФ, при этом аудиосигнал разбивается на короткие сегменты для получения последовательности частотных спектров. Кроме того, каждый спектр пропускается через набор фильтров, так называемый банк фильтров мела, для преобразования частот в Мел шкалу. По сути, главное отличие MFCC от обычного ОПФ – использование логарифмической шкалы, что позволяет решить один из недостатков классического ОПФ – однородность получаемой информации. Пример мел-спектрограммы показан на рисунке 4.

Как видно из рисунка, использование MFCC позволило получить более подробные сведения о частотном диапазоне, соответствующем звучащей речи в обрабатываемом сигнале. Данный метод предобработки показывает себя крайне эффективно в случае, если поставленная задача связана исключительно с обработкой речи, однако в случае универсального решения их диапазона не хватает для единообразного представления всех частот сигнала. В этом случае помимо уже упоминавшегося БПФ используют иные техники, например, вейвлет преобразование.

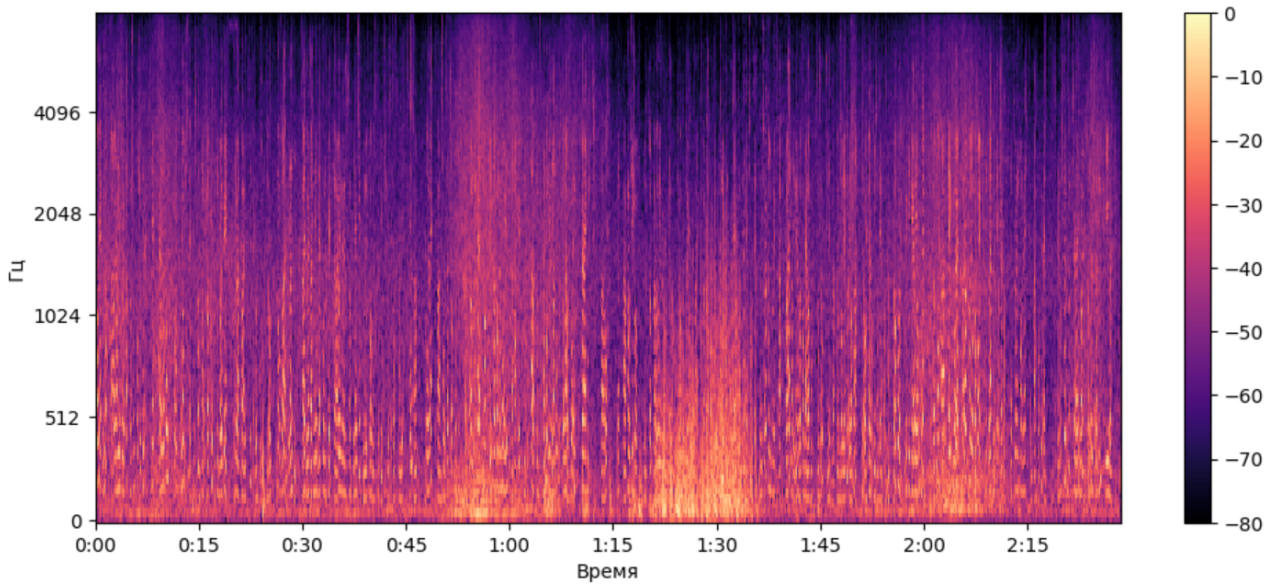


Рис. 4. Мел-спектрограмма

Вейвлет преобразования представляют собой семейства вейвлет функций, каждая из которых позволяет детектировать требуемые для конкретной задачи свойства исследуемого сигнала и их изменения со временем. Общая формула таких преобразований представлена ниже:

$$\psi_{ab}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right), \tag{2}$$

где a – положительное число, описывающее временной масштаб,

b – натуральное число, описывающее временной сдвиг,

t – текущее время,

$\psi(t)$ – материнская вейвлет функция, задающая исследуемые характеристики сигнала.

Конкретная функция ψ зависит от исследуемой характеристики сигнала. Математический аппарат для вычисления подобных функций достаточно сложен и требователен к ресурсам, поэтому на практике применяется так называемое дискретное вейвлет преобразование (ДВП), которое является альтернативой БПФ. Данное семейство преобразований применяется при известном типе сигнала и позволяет получить из него информацию о присутствии определённого спектра частот в различные моменты времени для дальнейшего анализа. Результатом ДВП является набор так называемых аппроксимирующих (cA) и детализирующих (cD) коэффициентов. При этом часто ДВП применяется повторно, что позволяет получить единый набор cA и несколько наборов cD. При этом каждый следующий набор детализирующих коэффицентов меньше предыдущего в два раза, но при этом позволяет лучше отобразить общую картину сигнала. Графически результат ДВП можно представить в виде тепловой картины, содержащей полученные наборы коэффициентов. Пример отображён на рисунке 5.

Рис. 5. Результат применения пяти уровней ДВП к сигналу

Ряд методов предобработки сигнала сводятся не к математическим преобразованиям, а к использованию системы ИИ для выделения признаков сигнала. Развитие нейронных сетей привело к появлению так называемых расширенных причинно-следственных свёрточных слоёв (dilated causal convolution), которые активно используются для преобразования исходного сигнала в скрытое пространство, хранящее обработанные признаки входного сигнала и позволяющее анализировать сигнал в реальном времени. Обычные свёрточные сети хорошо справляются с обработкой аудиосигнала [8-10] – для этого обычно используется так называемая одномерная свёрточная сеть (Convolution1D), которая проходит одномерным ядром свёртки по исходному сигналу и генерирует карту признаков как показано на рисунке 6.

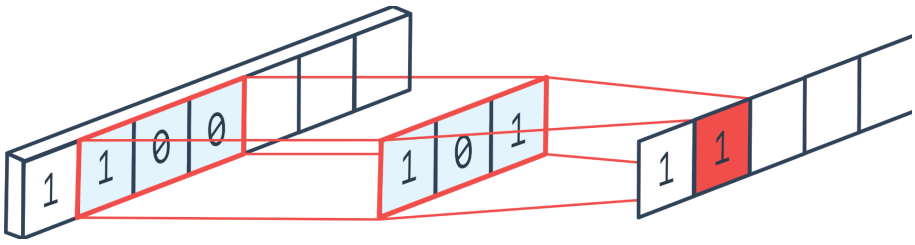


Рис. 6. Пример работы одномерного свёрточного слоя [11]

Однако для аудиосигнала важно, чтобы свёрточная сеть не могла «заглянуть в будущее» и учитывала только предыдущие параметры исследуемого сигнала при генерации карты признаков. Это привело к появлению причинно-следственных слоёв свёртки (causal convolution), описанных в [12]. Изображение из данного источника приведено на рисунке 7.

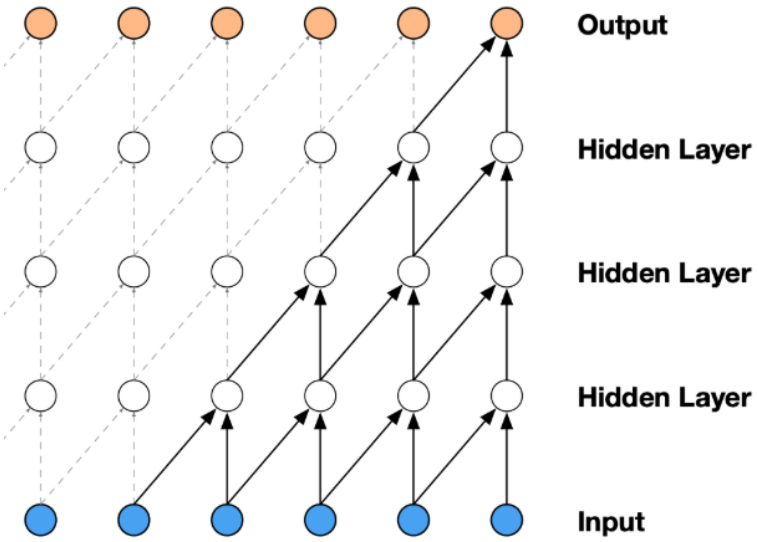


Рис. 7. Пример работы причинно-следственного свёрточного слоя

Фильтр подобных слоёв обрабатывает информацию только об уже прошедших во времени сигналах, не заглядывая в «будущее». Для большего охвата входного сигнала, смещение обрабатываемых значений изменяется в зависимости от слоя и числа сигналов, образуя так называемые расширенные причинно-следственные свёрточные слои, изображение которых заимствовано из [12], показано на рисунке 8.

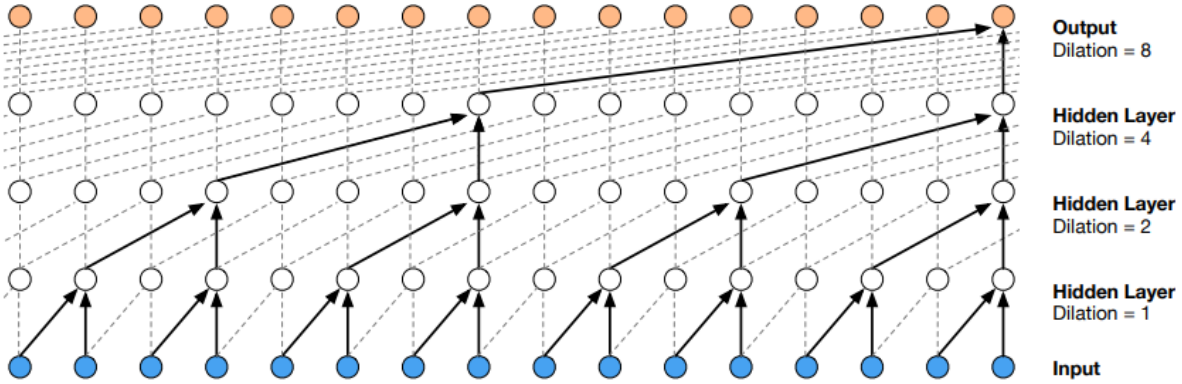


Рис. 8. Пример работы DCC

Подобные слои позволяют сократить вычисления, при этом сохранив достаточно информации для принятия сетью корректного решения. Аудиосигнал, обработанный с помощью набора DCC переводится в так называемое скрытое пространство (latent space), которое содержит информацию о выявленных в процессе свёртки признаках и пригодно для дальнейшей обработки.

После любого из методов предобработки сигнал сохраняет основную информацию и свойства, главные из которых – непрерывность во времени и наличие зависимостей между элементами последовательности. Каждый кадр исходного сигнала предобрабатывается согласно выбранному методу и передаётся в модель ИИ, которая может использовать как классические методы машинного обучения (например, скрытые марковские модели или байесовский классификатор), так и методы глубокого обучения, основанные на построении многослойных нейронных сетей для обработки сигналов. Современные архитектуры глубоких нейронных сетей для обработки аудиосигнала и будут рассмотрены далее.

2. Архитектуры современных систем искусственного интеллекта для обработки аудиосигнала на базе глубоких нейронных сетей

В последние десять лет область ИИ, связанная с глубоким обучением, бурно развивается. В рамках одной статьи практически невозможно разобрать всё множество появившихся в последнее время архитектур и моделей, но можно выделить наиболее характерные для решаемых задач и подходов к организации обработки входных данных. В статье [13] приводится экспериментальное сравнение нескольких архитектур искусственных нейронных сетей (ИНС) для задачи программной эмуляции аналогового усилителя сигнала электрогитары. Сравнение показало, что современные архитектуры свёрточных и рекуррентных сетей способны полноценно заменить аналоговые усилители для обработки предварительно записанного сигнала, однако работа в реальном времени требует значительных ресурсов. Но уже относительно простые рекуррентные модели, построенные на основе LSTM слоёв способны достаточно эффективно эмулировать гитарные педали в реальном времени, будучи использованы на одноплатном компьютере Raspberry Pi 4 [14].

Рассмотренная в статье [14] генеративная ИНС WaveNet [12] от компании Google DeepMind, заслуживает отдельного обзора, как характерный пример многозадачной ИНС, успешно применяющейся как для обработки сигнала, так и для задач TTS и STT. Основой модели являются упомянутые ранее слои DCC, а полная архитектура указана на рисунке 9.

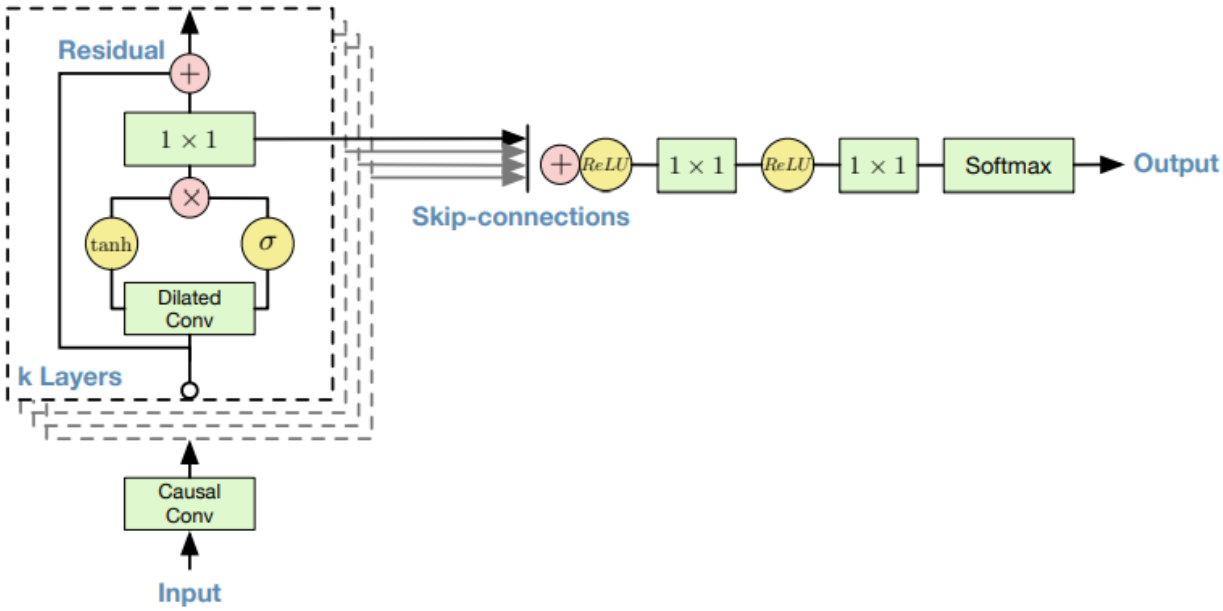


Рис. 9. Архитектура сети WaveNet [12]

Для задачи TTS на вход сети подаются предсказанные на основе текста мел-кепстральные коэффициенты, описывающие произносимую фразу, вместе с полученными из эталона характеристиками диктора. На выходе сеть генерирует аудиосигнал, соответствующий желаемым характеристикам. Для задач обработки сигнала во время обучения, построенная по данной архитектуре модель учится характерным образом фильтровать входной сигнал. Получив характеристики входного сигнала, она обучается изменять их, что приводит к добавлению желаемого звукового эффекта. Данная архитектура до сих пор популярна, однако сложна в обучении и требовательна к ресурсам, что ограничивает её применимость.

С развитием архитектуры трансформеров, появились исследования о применимости ИНС, использующих механизм внимания [15] для задач обработки аудиосигнала. Характерными примерами являются популярные ИНС Whisper, Wav2Vec, xtts_v2 и многие другие.

Whisper [16] – семейство популярных моделей для распознавание речи от компании OpenAI. Важной особенностью является наличие версий модели, оптимизированных для работы в реальном времени на низкопроизводительных устройствах, таких как одноплатные компьютеры (например, Raspberry Pi) или мобильные телефоны, что значительно расширяет возможности применения данных моделей. В основе моделей лежит технология построения сетей – трансформеров GPT, обрабатывающая мел-кепстральные коэффициенты входного сигнала. Архитектура сети представлена на рисунке 10.

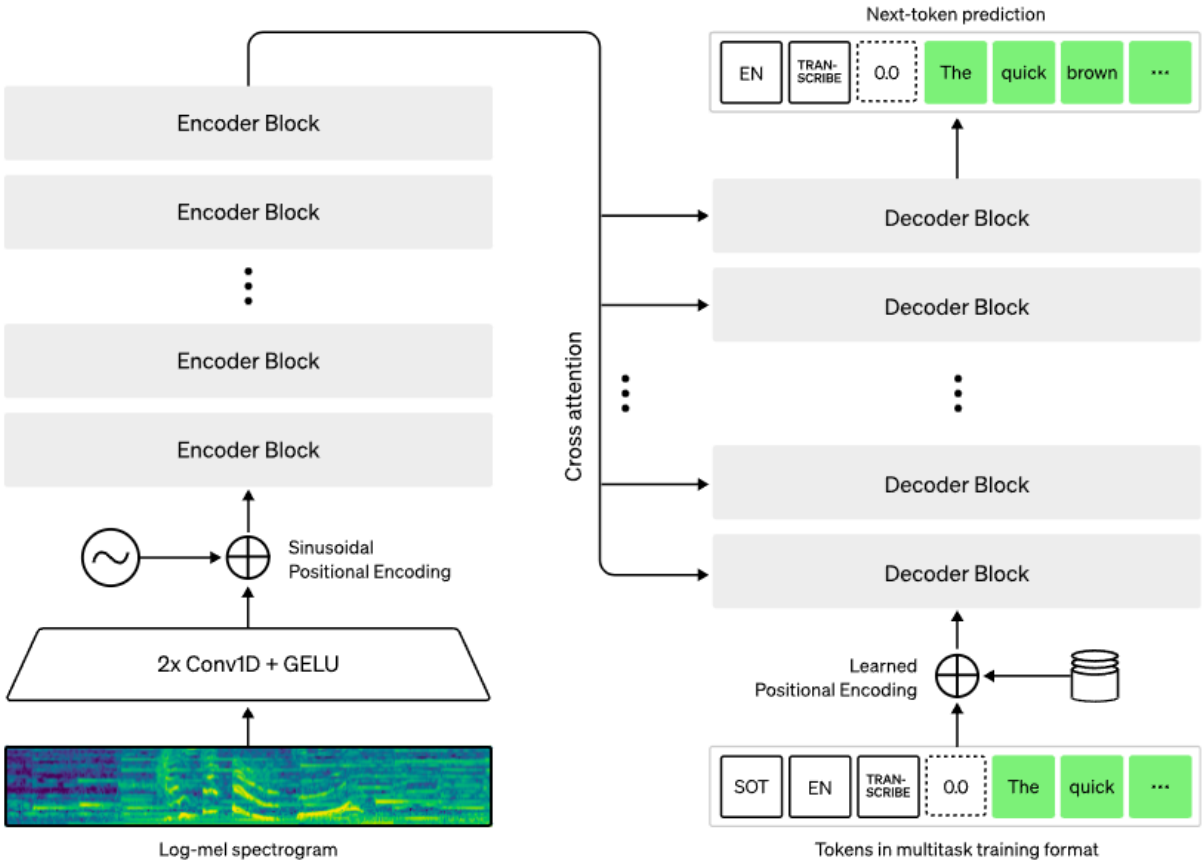


Рис. 10. Архитектура сети Whisper [16]

Данная архитектура является одной из ведущих при проектировании систем распознавания речи. Важным достоинством является соблюдение правил пунктуации при генерации результирующего текста и поддержка нескольких языков, включая русский. При этом модель small всего с 74 миллионами параметров обеспечивает приемлемое качество и возможность работы в реальном времени на одноплатном компьютере Raspberry Pi 5.

xtts_v2 [17] от компании Coqui – популярная генеративная модель TTS, позволяющая копировать эталонный голос на основе шестисекундного отрывка. Однако, хотя веса модели распространяются свободно, её архитектура закрыта и узнать можно только некоторые её особенности. В её основе также лежит

GPT модель для обработки входного текста и преобразования его в скрытое пространство, восстанавливаемое при помощи VQ-VAE в целевой аудиосигнал, изменённый с учётом токенов, полученных при помощи того же VQ-VAE из эталонного аудиофайла. Эксперимент по использованию данной модели, проведённый при написании данной статьи, показал хорошее по мнению опрошенных слушателей качество результирующего сигнала при генерации текста на русском языке и небольшое время работы при использовании GPU Nvidia 4080 RTX. На рисунке 11 показана мел-спектрограмма эталонного сигнала, а на рисунке 12 – сгенерированного моделью. Стоит заметить, что сгенерированный сигнал лишён фоновых шумов и, ввиду скопированной манеры разговора диктор, содержит набор пауз.

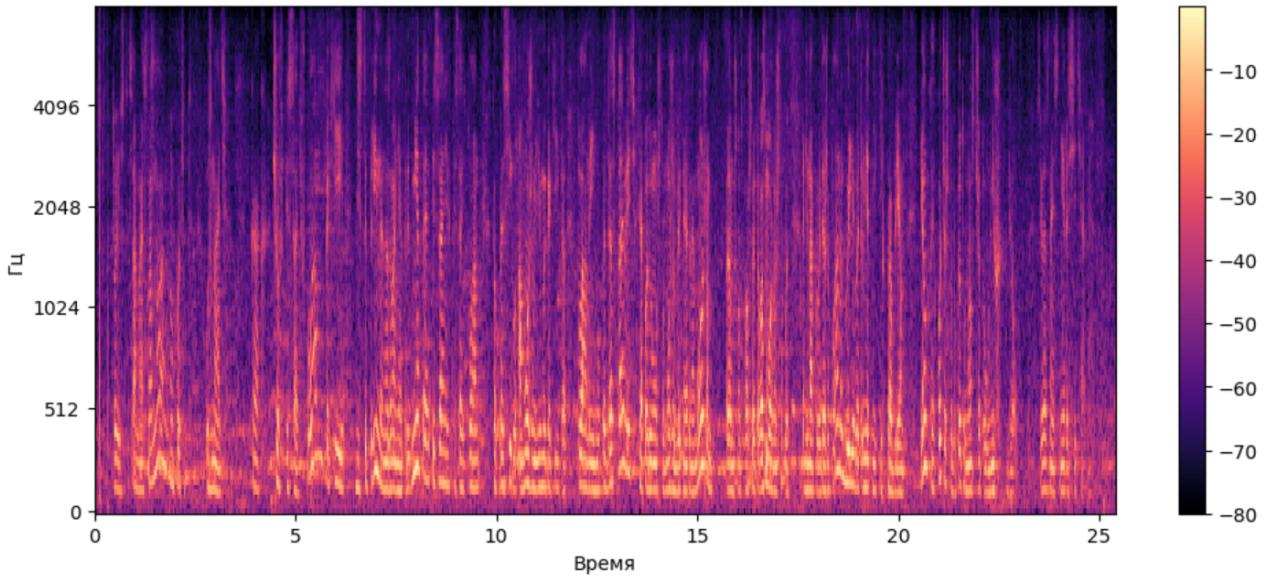


Рис. 11. Мел-спектрограмма эталонного сигнала диктора

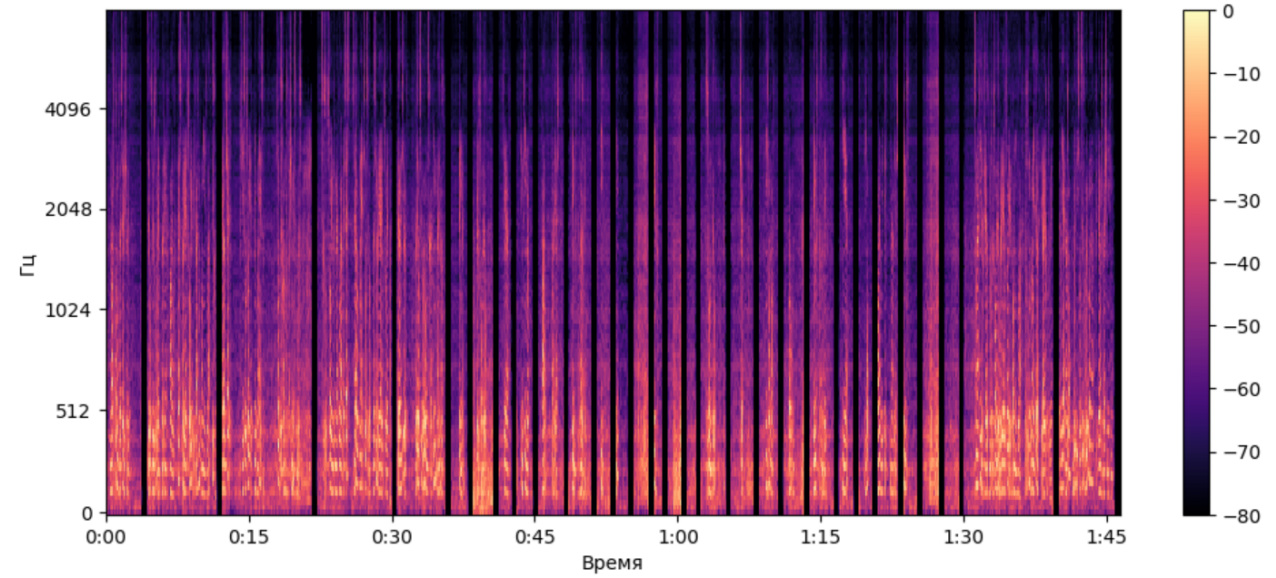
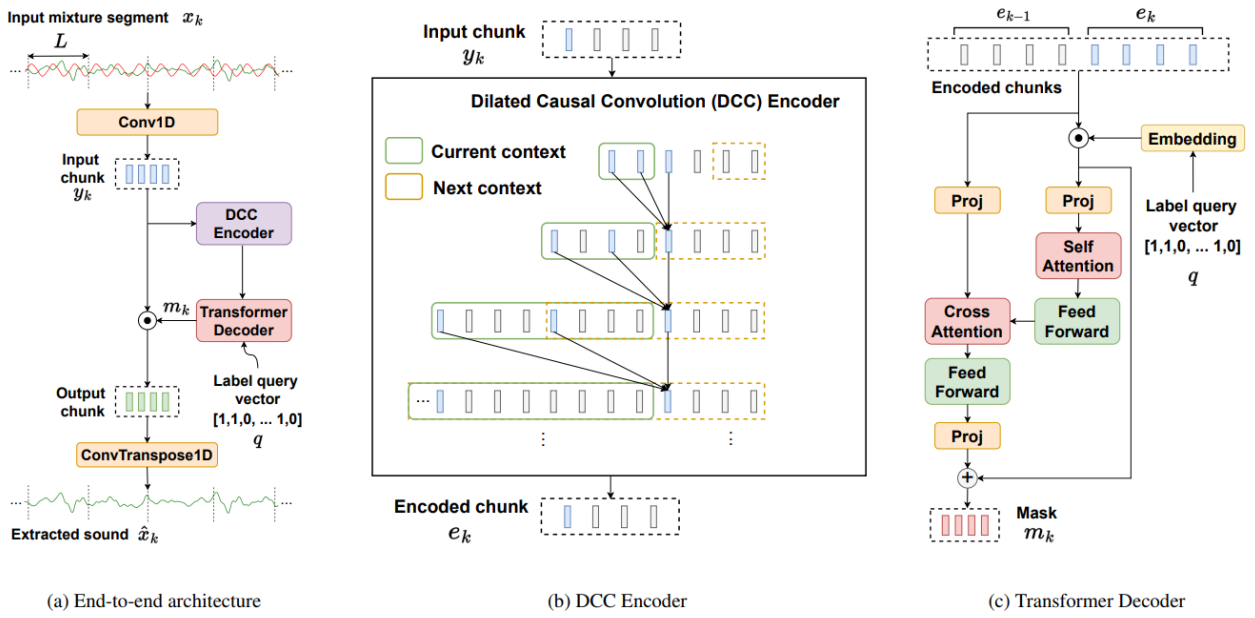


Рис. 12. Мел-спектрограмма сгенерированного сигнала диктора

Для задачи разделения сигнала хорошо себя показывает модель Waveformer [18]. Она позволяет при вычислении на ЦПУ достигать задержки всего в 20 мс. Данная архитектура позволяет обучить модель, подходящую для выделения сигнала от отдельного объекта в исходном аудиосигнале, разделения входного сигнала на все отдельные сигналы, а также задач, связанных с подавлением шумов в обрабатываемом сигнале. Архитектура данной модели показана на рисунке 13.



(a) End-to-end architecture (b) DCC Encoder (c) Transformer Decoder

Рис. 13. Архитектура сети Waveformer [18]

На вход подаётся фрагмент исходного файла в необработанной форме. Полученный сигнал проходит сначала через слой одномерной свёртки, а затем переводится в скрытое пространство посредствам набора DCC слоёв. Скрытое пространство передаётся в декодер трансформера, который с учётом полученного запроса выделяет маску для искомого сигнала. Полученная маска суммируется с входным сигналом и вновь передаётся на слой одномерной свёртки для получения результата.

Для задач шумоподавления существует множество различных подходов, начиная от описанного ранее Waveformer и применения диффузионной модели, популярной для исправления зашумлённости изображений [19]. Так в статье [20] показана эффективность использования архитектуры U-Net для удаления гауссовского шума из аудиосигнала. При написания настоящей статьи был проведён эксперимент с использованием модели aTENNuate [21], которая использует слои, моделирующие пространство состояний (SSM State Space Model), описанные в статье [22]. Данные слои являются модификацией рекуррентных слоёв и показывают себе эффективнее при построении глубоких сетей для обработки изменяющегося во времени сигнала. Модель aTENNuate обучалась по методу автокодировщика и её архитектура представлена на рисунке 14.

В ходе проведённого эксперимента, модель эффективно удалила шумы из ряда аудиосигналов, потратив в среднем 10 секунд на обработку ранее описанного сигнала. Полученный сигнал показан на рисунке 15.

Можно также привести мел-кепстральные коэффициенты полученного сигнала (рисунок 16), которые показывают, что связанные с речью частоты в основном не подверглись изменению, что показывает эффективность данного метода. Визуальные паузы на графике связаны с манерой речи.

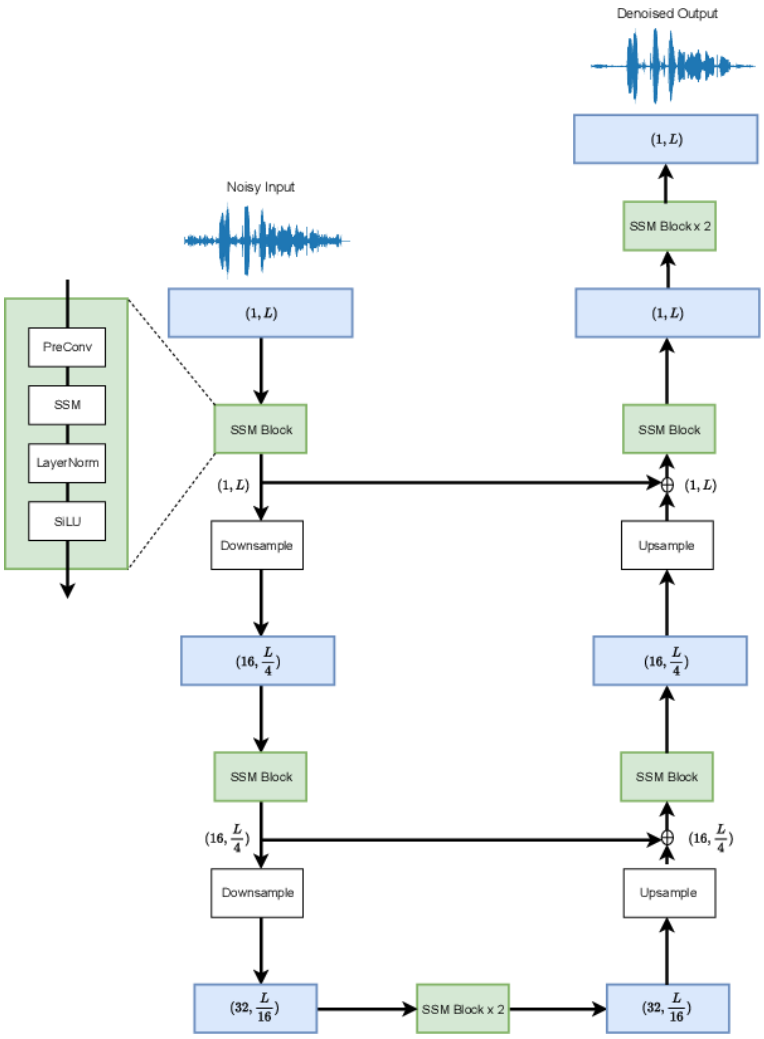


Рис. 14. Архитектура сети aTENNuate [21]

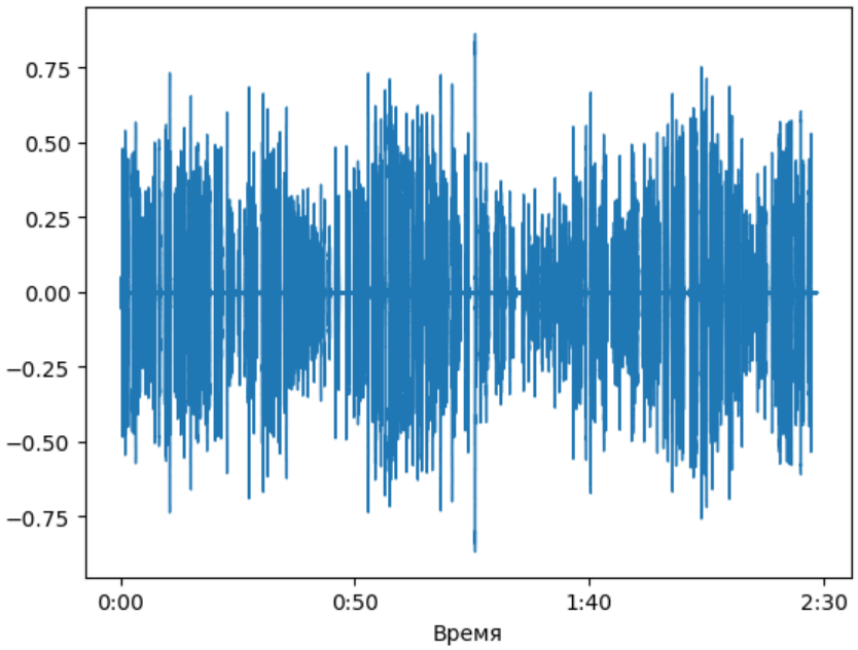


Рис. 15. Сигнал после обработки aTENNuate

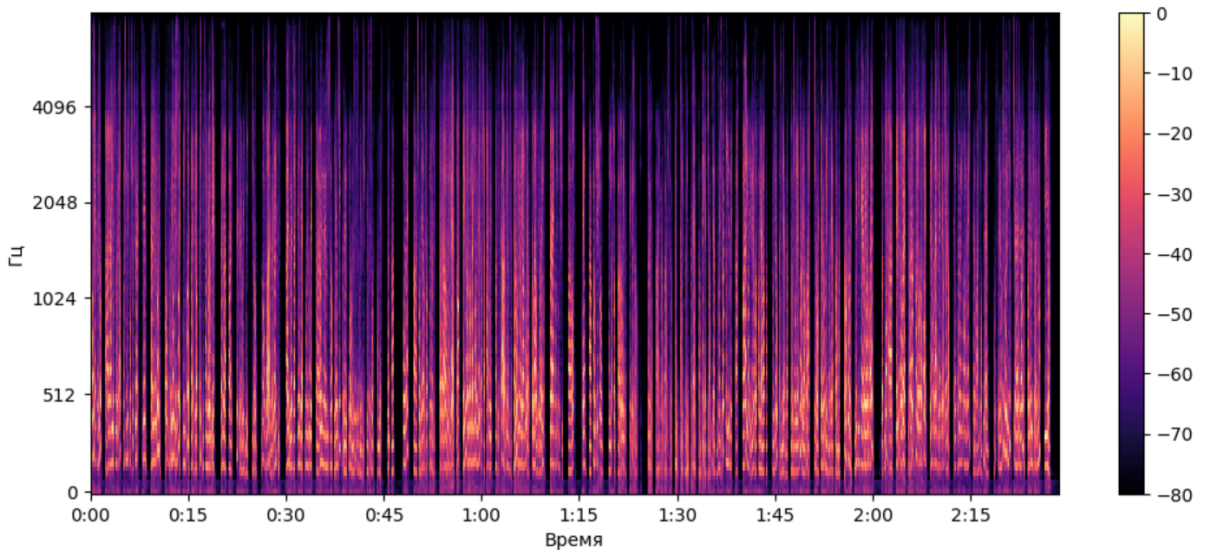


Рис. 16. Мел-спектрограмма сигнала после обработки aTENNuate

Заключение

Применение современных методов построения систем искусственного интеллекта для обработки аудиосигналов позволяют достичь высокой производительности и точности. При этом, хотя обучение глубоких нейронных сетей и требует значительных ресурсных затрат, но современные технологии квантования и оптимизации позволяют использовать предобученные модели ИНС на широком спектре устройств, в том числе с ограниченными вычислительными возможностями.

Методы предобработки сигнала и конкретная архитектура ИНС сильно отличаются в зависимости от выполняемой задачи, что должно быть учтено при проектировании. Современные тенденции развития систем ИИ для обработки аудиосигнала говорят об постепенном уходе от использования БПФ в задачах комплексной обработки сигнала в пользу использования DCC и перевода сигнала в скрытое пространство, позволяющее использовать при обработке сигнала архитектуру ИНС трансформер и диффузионные модели. В то же время задачи обработки речи всё ещё активно подразумевают использование MFCC для построения мел-спектрограммы входного сигнала. Использование DCC показало свою эффективность при удалении шумов из исследуемого аудиосигнала без значительных искажений полезной части сигнала.

Список литературы

1. Yeon I. et al. EchoScan: Scanning Complex Indoor Geometries via Acoustic Echoes //arXiv preprint arXiv:2310.11728. – 2023. DOI: https://doi.org/10.48550/arXiv.2310.11728

2. Sato H. et al. Speakerbeam-ss: Real-time target speaker extraction with lightweight conv-tasnet and state space modeling //arXiv preprint arXiv:2407.01857. – 2024. DOI: https://doi.org/10.48550/arXiv.2407.01857

3. Lam M. W. Y. et al. Efficient neural music generation //Advances in Neural Information Processing Systems. – 2024. – T. 36. DOI: https://doi.org/10.48550/arXiv.2305.15719

4. Pei Y. R., Shrivastava R., Sidharth F. N. U. Real-time Speech Enhancement on

Raw Signals with Deep State-space Modeling //arXiv preprint arXiv:2409.03377. – 2024.DOI: https://doi.org/10.48550/arXiv.2409.03377

5. Zhou, R., Zhao, S., Luo, M. et al. MFCC based real-time speech reproduction and recognition using distributed acoustic sensing technology. Optoelectron. Lett. 20, 222–227 (2024) DOI: https://doi.org/10.1007/s11801-024-3167-5

6. Peng P. et al. Voicecraft: Zero-shot speech editing and text-to-speech in the wild //arXiv preprint arXiv:2403.16973. – 2024. DOI: https://doi.org/10.48550/arXiv.2403.16973

7. Tzanetakis, George & Essl, Georg & Cook, Perry. (2001). Audio Analysis using the Discrete Wavelet Transform. Proceedings of the Conference in Acoustics and Music Theory Applications. 318-323.

8. de Benito-Gorron, D., Lozano-Diez, A., Toledano, D.T. et al. Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. J AUDIO SPEECH MUSIC PROC. 2019, 9 (2019). DOI https://doi.org/10.1186/s13636-019-0152-1

9. Попов Владислав Николаевич, Ладыгин Павел Сергеевич, Карев Валентин Витальевич, Борцова Яна Игоревна РАЗРАБОТКА СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ КЛАССИФИКАЦИИ АМПЛИТУДНО-ЧАСТОТНЫХ ХАРАКТЕРИСТИК АУДИОСИГНАЛОВ // Известия АлтГУ. 2022. №1 (123).

10. К.И. Дементьева, А.А. Ракитский Метод для восстановления аудиосигнала с помощью свёрточных нейронных сетей. Вестник НГУ. Серия: Информационные технологии. 2022 Т.20, №3. С. 38–50.

11. Subhash D. Convolution Made Easy [Электронный ресурс] // Medium. – URL: https://darshanasubhash.medium.com/convolution-made-easy-83d90371ed25 (дата обращения: 26.02.2025).

12. Oord A. WaveNet: A Generative Model for Raw Audio //arXiv preprint arXiv:1609.03499. – 2016. DOI: https://doi.org/10.48550/arXiv.1609.03499

13. Wright A. et al. Real-time guitar amplifier emulation with deep learning //Applied Sciences. – 2020. – Т. 10. – №. 3. – С. 766. DOI: https://doi.org/10.3390/app10030766

14. Bloemer K. Neural Networks for Real-Time Audio: Raspberry-Pi Guitar Pedal // Towards Data Science. – 24.05.2021. – URL: https://towardsdatascience.com/neural-networks-for-real-time-audio-raspberry-pi-guitar-pedal-bded4b6b7f31 (дата обращения: 08.02.2025).

15. Vaswani A. Attention is all you need //Advances in Neural Information Processing Systems. – 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762

16. Radford A. et al. Robust speech recognition via large-scale weak supervision //International conference on machine learning. – PMLR, 2023. – С. 28492-28518. DOI: https://doi.org/10.48550/arXiv.2212.04356

17. Coqui. XTTS [Электронный ресурс] // Coqui TTS Documentation. – URL: https://docs.coqui.ai/en/latest/models/xtts.html (дата обращения: 08.02.2025).

18. Veluri B. et al. Real-time target sound extraction //ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2023. – С. 1-5. DOI: https://doi.org/10.48550/arXiv.2211.02250

19. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models //Advances in neural information processing systems. – 2020. – Т. 33. – С. 6840-6851 DOI: https://doi.org/10.48550/arXiv.2006.11239

20. Wang P. et al. Diffusion Gaussian Mixture Audio Denoise //arXiv preprint arXiv:2406.09154. – 2024. DOI: https://doi.org/10.48550/arXiv.2406.09154

21. Pei Y. R., Shrivastava R., Sidharth F. N. U. Real-time Speech Enhancement on Raw Signals with Deep State-space Modeling //arXiv preprint arXiv:2409.03377. – 2024. DOI: https://doi.org/10.48550/arXiv.2409.03377

22. Gu A., Goel K., Ré C. Efficiently modeling long sequences with structured state spaces //arXiv preprint arXiv:2111.00396. – 2021. DOI: https://doi.org/10.48550/arXiv.2111.00396

References

1. Yeon I. et al. EchoScan: Scanning Complex Indoor Geometries via Acoustic Echoes //arXiv preprint arXiv:2310.11728. – 2023. DOI: https://doi.org/10.48550/arXiv.2310.11728

2. Sato H. et al. Speakerbeam-ss: Real-time target speaker extraction with lightweight conv-tasnet and state space modeling //arXiv preprint arXiv:2407.01857. – 2024. DOI: https://doi.org/10.48550/arXiv.2407.01857

3. Lam M. W. Y. et al. Efficient neural music generation //Advances in Neural Information Processing Systems. – 2024. – T. 36. DOI: https://doi.org/10.48550/arXiv.2305.15719

4. Pei Y. R., Shrivastava R., Sidharth F. N. U. Real-time Speech Enhancement on Raw Signals with Deep State-space Modeling //arXiv preprint arXiv:2409.03377. – 2024. DOI: https://doi.org/10.48550/arXiv.2409.03377

5. Zhou, R., Zhao, S., Luo, M. et al. MFCC based real-time speech reproduction and recognition using distributed acoustic sensing technology. Optoelectron. Lett. 20, 222–227 (2024) DOI: https://doi.org/10.1007/s11801-024-3167-5

6. Peng P. et al. Voicecraft: Zero-shot speech editing and text-to-speech in the wild //arXiv preprint arXiv:2403.16973. – 2024. DOI: https://doi.org/10.48550/arXiv.2403.16973

7. Tzanetakis, George & Essl, Georg & Cook, Perry. (2001). Audio Analysis using the Discrete Wavelet Transform. Proceedings of the Conference in Acoustics and Music Theory Applications. 318-323.

8. de Benito-Gorron, D., Lozano-Diez, A., Toledano, D.T. et al. Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset. J AUDIO SPEECH MUSIC PROC. 2019, 9 (2019). DOI: https://doi.org/10.1186/s13636-019-0152-1

9. Popov Vladislav Nikolaevich, Ladygin Pavel Sergeevich, Karev Valentin Vitalievich, Bortsova Yana Igorevna DEVELOPMENT OF CONVOLUTION OF A NEURAL NETWORK FOR CLASSIFICATION OF AMPLITUDE-FREQUENCY CHARACTERISTICS OF AUDIO SIGNALS // News of Altai State University. 2022. No. 1 (123).

10. K.I. Dementieva, A.A. Rakitskiy Method for audio signal restoration using convolutional neural networks. Bulletin of NSU. Series: Information technologies. 2022 T.20, No. 3. pp. 38–50.

11. Subhash D. Convolution Made Easy // Medium. – URL: https://darshanasubhash.medium.com/convolution-made-easy-83d90371ed25

12. Oord A. WaveNet: A Generative Model for Raw Audio //arXiv preprint arXiv:1609.03499. – 2016. DOI https://doi.org/10.48550/arXiv.1609.03499

13. Wright A. et al. Real-time guitar amplifier emulation with deep learning //Applied Sciences. – 2020. – T. 10. – №. 3. – C. 766. DOI: https://doi.org/10.3390/app10030766

14. Bloemer K. Neural Networks for Real-Time Audio: Raspberry-Pi Guitar Pedal // Towards Data Science. – 24.05.2021. – URL: https://towardsdatascience.com/neural-networks-for-real-time-audio-raspberry-pi-guitar-pedal-bded4b6b7f31 (дата обращения: 08.02.2025).

15. Vaswani A. Attention is all you need //Advances in Neural Information Processing Systems. – 2017. DOI: https://doi.org/10.48550/arXiv.1706.03762

16. Radford A. et al. Robust speech recognition via large-scale weak supervision

//International conference on machine learning. – PMLR, 2023. – С. 28492-28518. DOI: https://doi.org/10.48550/arXiv.2212.04356

17. Coqui. XTTS // Coqui TTS Documentation. – URL: https://docs.coqui.ai/en/latest/models/xtts.html

18. Veluri B. et al. Real-time target sound extraction //ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2023. – С. 1-5. DOI: https://doi.org/10.48550/arXiv.2211.02250

19. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models //Advances in neural information processing systems. – 2020. – Т. 33. – С. 6840-6851 DOI: https://doi.org/10.48550/arXiv.2006.11239

20. Wang P. et al. Diffusion Gaussian Mixture Audio Denoise //arXiv preprint arXiv:2406.09154. – 2024. DOI: https://doi.org/10.48550/arXiv.2406.09154

21. Pei Y. R., Shrivastava R., Sidharth F. N. U. Real-time Speech Enhancement on Raw Signals with Deep State-space Modeling //arXiv preprint arXiv:2409.03377. – 2024. DOI: https://doi.org/10.48550/arXiv.2409.03377

22. Gu A., Goel K., Ré C. Efficiently modeling long sequences with structured state spaces //arXiv preprint arXiv:2111.00396. – 2021. DOI: https://doi.org/10.48550/arXiv.2111.00396