УДК 621.391 **DOI** 10.15622/ia.24.2.3

И.И. Чижов

АУДИОКОДЕК НА ОСНОВЕ ПЕРЦЕПТУАЛЬНОГО РАВЕНСТВА ИСХОДНОГО И ВОССТАНОВЛЕННОГО ЗВУКОВОГО СИГНАЛА

Чижов И.И. **Аудиокодек на основе перцептуального равенства исходного и восстановленного звукового сигнала.**

Аннотация. Представлен метод сжатия аудиоданных с потерями (аудиокодек), позволяющий улучшить объективное качества восстановленного аудиосигнала на 25% для битрейта 390 кбит/с и 55% для битрейта 64кбит/с по сравнению с форматом AAC MPEG-4. Предлагаемый метод сжатия аудиоданных, базируется на развитии положений теории сжатия аудиоданных с потерями (ТСАП). Повышение объективного качества восстановленного звукового сигнала (по стандартизованной мере PEAQ) достигается за счет того, что ТСАП устраняет несовершенства современных методов сжатия аудиоданных с потерями в части использования психоакустических принципов восприятия звука человеком, в том числе после преодоления «психоакустического предела сжатия» аудиосигнала (т.е. момента в перцептуальном кодировании, когда имеющегося бюджета бит недостаточно для кодирования всех спектральных компонент с необходимой с точки зрения психоакустики точностью), и позволяет достичь перцептуального равенства восприятия исходного и восстановленного аудиосигналов. В качестве анализа состояния вопроса рассмотрены решения по сжатию аудиоданных без потерь и с потерями, а также с использованием искусственного интеллекта. Во всех современных методах сжатия аудиоданных с потерями процедура выбора спектральных компонент, которые необходимо сохранить, а также допустимой погрешности квантования их по уровню выполняется путем ряда достаточно сложных процедур, носящих общее название «психоакустическая модель метода сжатия аудиоданных с потерями». В строгом смысле, перцептуальное равенство спектров исходного и восстановленного сигналов ни одна из групп исследователей не доказала и как следствие – не может его гарантировать. Независимые эксперты регулярно публикуют тесты, показывающие, что современные аудиокодеки имеют проблемы на ряде сигналов. В статье предложен аудиокодек на основе перцептуального равенства исходного и восстановленного звукового сигнала, который базируется на новых идеях теории сжатия аудиоданных с потерями (ТСАП). Эти идеи гарантируют достижение перцептуального равенства восприятия исходного и восстановленного аудиосигналов на различных битовых скоростях, поэтому построенный на ее основе аудиокодек свободен от указанных выше недостатков и, как следствие, существенно превосходит современные кодеки в смысле объективного качества восстановленного аудиосигнала по мере PEAQ.

Ключевые слова: сжатие аудиоданных, психоакустическая модель, спектр, субполосное разделение, перцептуальное равенство сигналов.

1. Введение. Обычно, сжатие аудиосигнала с потерями, т.е. процедура выбора спектральных компонент, которые необходимо сохранить, а также допустимой погрешности квантования их по уровню, выполняется путем ряда последовательных, достаточно сложных процедур, носящих общее название «психоакустическая модель метода сжатия аудиоданных с потерями» [1 – 5].

В строгом смысле, перцептуальное равенство спектров исходного и восстановленного сигналов ни одна из групп исследователей не доказала, поэтому ни один из современных коммерческих кодеков и не может его гарантировать. Другими словами, современные методы сжатия аудиоданных с потерями не могут гарантировать, что все возможные аудиосигналы будут сжаты корректно с точки зрения соблюдения принципов психоакустики. В качестве доказательства этого утверждения, стоит напомнить, что независимые эксперты регулярно проводят тесты на сложных для аудиокодеков файлах (“codec killer test”), на которых у всех методов сжатия аудиоданных, возникают проблемы на ряде сигналов.

Из всего вышеизложенного следует, что построение новой Теории сжатия аудиоданных с потерями (ТСАП), обеспечивающей достижение перцептуального равенства восприятия исходного и восстановленного аудиосигналов на любых битовых скоростях, представляется важной научной проблемой, существенной для такой области знания как кодирование звуковых сигналов.

Отметим, что метод сжатия аудиоданных, построенный на основе ТСАП, будет свободен от указанных выше недостатков и поэтому должен существенно превосходить современные кодеки в смысле объективного качества восстановленного аудиосигнала по мере PEAQ (“Objective Measurements of Perceived Audio Quality” – стандартизована Международным союзом электросвязи) [6].

2. Аналитический обзор аудиокодеков. В настоящее время существует два различных по сути подхода к сжатию аудиоинформации: сжатие аудиосигнала без потерь и сжатие с потерями.

2.1. Методы сжатия аудиоданных без потерь. Данные методы основываются на удалении статистической избыточности исходного сигнала и не вносят изменений в восстановленный сигнал по отношению к исходному. Они представляют слабый интерес с точки зрения значительного уменьшения битового представления исходного аудиосигнала. К основным алгоритмам сжатия аудиоданных (и вообще любых данных) без потерь относятся кодирование по методу Хаффмана [7, 8] и арифметическое кодирование [9].

На современном рынке представлен ряд подобных кодеков:

- FLAC [10] (Free Lossless Audio Codec) позволяет осуществить сжатие аудиоданных до 50%.
- Кодек WavPack [11], разработанный Дэвидом Бриантом, позволяет осуществлять компрессию на 30-70%.

– Кодек Monkey’s Audio [12] в настоящее время активно дорабатывается. Последнее обновление было опубликовано на официальном сайте 29 сентября 2023 года. Сжимает исходный файл на 40-60%.

– Кодек Apple Lossless Audio Codec (ALAC) [13], разработанный компанией Apple, впервые был представлен в 2004 году и позволяет осуществить сжатие аудиоданных на 40-60%.

Основными методами уменьшения битового представления аудиоданных в настоящее время являются методы и алгоритмы сжатия с потерями. В основе данных методов сжатия аудиоданных лежит учет особенностей слуха человека, в частности неспособность человека различать звуки при определенных условиях. Например, невозможность различать тихие аудиосигналы после громких, неспособность слышать очень высокие и очень низкие звуки и т.д. Таким образом, психоакустическая модель, учитывающая особенности слуха человека, положена в основу методов и алгоритмов сжатия аудиоданных с потерями, которые реализованы в виде различных кодеков.

Большинство кодеков, использующихся для сжатия аудиоданных, в своей основе ориентированы на следующий алгоритм:

1. Разбиение исходного аудиосигнала на отдельные фреймы (кадры). Длина фрейма может быть различной, но, как правило, исчисляется в миллисекундах. При дальнейшем анализе входящей аудиоинформации зачастую используется не только текущий фрейм, но и предыдущий, либо фреймы берутся с перекрытием, чтобы обеспечить более полный и адекватный анализ сигнала.

2. Спектральный анализ аудиосигнала по каждому из фреймов. Преимущественно для получения информации о спектральных составляющих фрейма аудиосигнала используется модифицированное дискретное косинус-преобразование (МДКП). Также перед реализацией МДКП в некоторых алгоритмах осуществляется разделение спектра исходного сигнала на спектральные полосы, а вычисление непосредственно МДКП осуществляется по каждой отдельной полосе (как например в MPEG-1 Layer III [1]).

3. Использование психоакустической модели для принятия решения о том, какие из составляющих необходимо оставить. Использование психоакустической модели позволяет обеспечить приемлемое качество полученного аудиосигнала, основываясь на несовершенстве слуха человека, эффектах временной и частотной маскировки одних звуков другими. Таким образом, данные действия

позволяют в значительной мере уменьшить объём исходной аудиоинформации.

4. Квантование по уровню спектральных коэффициентов. Методы, используемые для квантования, а также количество уровней квантования влияют на объем и качество аудиосигнала.

5. В заключение для получения еще меньшего представления аудиосигнала используется какой-либо метод сжатия данных без потерь (метод Хаффмана или арифметическое кодирование).

2.2. Методы сжатия аудиоданных с потерями. Сжатие с потерями является на сегодняшний день «де-факто» стандартом в отрасли и большинство стандартизованных методов сжатия работают именно в этой парадигме.

Рассмотрим основные стандартизованные методы сжатия аудиоданных с потерями, в основе которых лежит психоакустический подход к сжатию:

Формат MP3 [1, 2, 4]**.** Довольно старым, но тем не менее популярным в настоящее время является формат MP3 (полное название MPEG-1 Audio Layer III, MPEG-2 Audio Layer III и MPEG-2.5 Audio Layer III). Исторически он является лидером по распространенности и имеет хорошую совместимость с различными программами и аудиоустройствами. По техническим параметрам формат MP3 уступает большинству современных форматов.

Формат WMA (Windows Media Audio) [14]**.** Данный формат сжатия аудиоданных был создан компанией Microsoft в качестве кодека, конкурирующего с MP3. Данный формат довольно распространен, но так как он является закрытым, то, соответственно, возникают проблемы с его использованием вне Windows систем.

Формат Vorbis [15]. Данный формат был создан Xiph в качестве свободного формата. Vorbis является похожим на MP3, однако имеются отличия как в психоакустической модели формата, так и в математических методах и их практической реализации в используемой модели.

Аудиокодек Opus [16]. Opus разработан Internet Engineering Task Force (IETF). При создании аудиокодека Opus была использована комбинация технологий, а именно кодека CELT (Constrained Energy Lapped Transform) от организации Xiph и кодека SILK от компании Skype. На сегодняшний день широко внедрен и является лидером среди бесплатных кодеков. Также сравнение результатов работы кодеков Opus и OGG Vorbis приводятся в [17].

MPEG Advanced Audio Coding (AAC). В настоящее время именно данный алгоритм является основным стандартом в отрасли.

Этот кодек базируется на подходах, апробированных в формате MP3, но использует значительно более эффективные технологии сжатия, чем его предшественник.

Спецификации AAC претерпели ряд изменений с момента подачи первого предложения. Для определения системы AAC, был выбран модульный подход, при котором вся система разбивается на ряд автономных модулей или инструментов, где инструмент определяется как модуль кодирования, который можно использовать как отдельный модуль, компонент общей системы.

Некоторые из этих инструментов взяты из более раннего стандарта MPEG Audio, в то время как другие являются новыми.

Если говорить очень обобщенно и оставить в стороне модули, отвечающие за стереокодирование (такие как кодирование суммарного и разностного каналов) и улучшение качества сигнала (такие как уменьшение размера используемого окна анализа в зависимости от наличия в сигнале резких всплесков), то схема работы метода AAC для каждого фрейма (1024 отсчета аудиосигнала) выглядит достаточно типично для современных методов сжатия с потерями:

- Психоакустическая модель определяет пороги маскирования, т.е. то на сколько можно изменить каждое значение спектрального компонента, чтобы эти искажения были не заметны на слух.
- Параллельно фрейм переводится в спектральную область путем применения Модифицированного дискретного косинусного преобразования (МДКП) и далее разбивается на субполосы в соответствии с Барк-шкалой.
- В каждой субполосе производится нормировка и квантование по уровню каждого спектрального компонента, таким образом, чтобы не нарушить полученные от психоакустической модели пороги маскировки (маскирования), применяя кантователи и масштабные коэффициенты (для нормировки) в соответствии со стандартом – данные значения берутся из соответствующих таблиц стандарта. Здесь стоит выделить применение нормировки, т.к. другие методы сжатия с потерями ее не используют.
- Далее применяется энтропийное кодирование (по методу Хаффмана или арифметическое).

Описанную выше часть метода сжатия с потерями принято называть перцептуальным сжатием аудиосигнала, в то время как остальные модули являются вспомогательными, однако именно они занимают значительную часть стандарта.

Более подробно с описанными процедурами можно ознакомиться, например, в рекомендациях ITU-R BS.1196-8 [18].

Здесь стоит отметить, что в указанной схеме есть несколько явных нарушений ТСАП (подробное описание будет дано в разделе 4.2).

В настоящее время используется современная версия стандарта AAC MPEG-4 [18]. Важно отметить, что существуют более современные версии кодека AAC – это AAC High Efficiency v1/2 (AAC HE v1/2) [18].

Однако, с точки зрения перцептуального сжатия аудиосигнала эти форматы не имеют отличий от AAC, а повышение качества достигается в случае AAC HE v1 за счет технологии восстановления высоких частот на основе имеющихся нижних и средних, именуемой SBR (Spectral Band Replication), в AAC HE v2 еще дополнительно появляется технология Параметрического кодирования стереопанорамы – PS (Pyranometric Stereo).

Достаточно очевидно, что эти технологии имеют крайне отдаленное отношения к перцептуальному сжатию аудио с потерями и поэтому не являются частью настоящего исследования.

2.3. Методы сжатия аудиоданных с использованием нейронных сетей. В настоящее время многие группы разработчиков, в первую очередь за рубежом, пытаются применить технологии нейросетей для компрессии звука и речи [19].

Так же вызывает большой интерес прогресс в смежных областях, таких как генерация речи из его мел-представления (что может быть использовано в качестве Декодера в системе сжатия речи) [20, 21], а так же попытка применения генеративных моделей на основе диффузии (“Diffusion”) [22] для улучшения качества речи, что может быть применено как пост-обработка в системах сжатия речи (и потенциально аудио). Надо отметить, что в сжатии речи ими были достигнуты определенные успехи, однако в сжатии аудиосигналов пока применение указанных технологий не дало существенного результата.

Основными кодеками, в основе которых лежит применение искусственного интеллекта, являются SoundStream и LPCNet.

LPCNet. LPCNet является гибридным кодеком [23, 24], использующим в своем составе как традиционные подходы к сжатию (в данном случае это сжатие на основе линейного предсказания значений речевого сигнала – LPC (linear predictive coding), широко применяемое в традиционных речевых кодеках), так и блок восстановления спектра речевого сигнала на основе нейросетевого

подхода (в случае LPCNet применена многослойная рекуррентная архитектура нейросети) [25, 26].

SoundStream [27, 28]. Основным компонентом аудиокодека SoundStream является нейронная сеть, состоящая из энкодера, декодера и квантователя, каждый из которых обучен сквозным способом (end-to-end). Кодер преобразует входной аудиопоток в закодированный сигнал, который сжимается с помощью квантователя, а затем преобразуется обратно в аудио с помощью декодера. Для повышения субъективного качества восстановленного сигнала SoundStream использует генеративно-состязательное обучение.

Дискриминатор вычисляет комбинацию функций состязательных и генеративных искажений, что повышает субъективное качество сигнала. После обучения энкодер и декодер можно запускать на отдельных машинах для эффективной передачи высококачественного звука по сети.

Кодер SoundStream создает латентные векторы, которые могут принимать нефиксированное количество значений. Чтобы передать их приемнику, используя ограниченное количество бит, необходимо заменить их близкими векторами из конечного набора (называемого кодовой книгой). Данный процесс известен как векторное квантование. Для повышения эффективности векторного квантования в SoundStream используется векторное квантование остаточных векторов – RVQ (Residual vector quantizer). Данный подход сейчас применяется большинством методов сжатия речи на основе нейронных сетей.

Стоит отметить, что существует более эффективный, чем SoundStream, метод сжатия, называющийся **EnCodec** (компания Meta) [29], но так как оба метода достаточно конструктивно схожи, приводить его подробно не имеет смысла. Отметим только, что в настоящее время именно EnCodec демонстрирует наиболее высокое качество восстановленного аудиосигнала из всех методов подобного класса. Достигается это за счет применения новых функций потерь и автоматического «балансера» весовой функции в суммарной функции потерь.

Однако, все указанные методы сжатия на основе искусственного интеллекта в настоящее время хотя и дают хорошие результаты для сжатия речи, но при сжатии аудиофайлов (например, музыки) не дают приемлемых результатов. Уместно предположить, что указанные различия в результатах обусловлены тем, что в речи есть явно выраженные паттерны (гармоники, обертона и т.д.), а в музыке либо

их вообще нет, либо они слишком разнообразны, что не позволяет нейронной сети эффективно обучиться для их генерации.

Кроме того, с увеличением битрейта качество у Нейросетевых кодеков не растет (а иногда даже немного снижается [29]). Это связано с отсутствием в их составе психоакустической модели, которая определяет, какие спектральные компоненты необходимо передать, а какие будут не слышны человеческим ухом. Таким образом, рост битрейта обеспечивает более точное квантование латентного вектора, что вообще-то не ведет непосредственно к увеличению субъективного качества восстановленного сигнала, так как далее латентный вектор проходит через «глубокий» декодер, который в достаточной мере способен компенсировать погрешность квантования латентного вектора, и точность его восстановления после квантования уже почти не играет никакой роли.

2.4. Постановка проблемы. Сжатие без потерь – интересная технология, но имеет маленький рынок, т.к. степень сжатия не превышает 2-3 раз, что явно недостаточно. И хотя многие стриминговые сервисы добавили их как возможные, но большой популярностью они, по оценкам самих сервисов, в настоящее время не пользуются. Кроме того, в них не используется перцептуальное сжатие, поэтому их рассмотрение выходит за рамки настоящей статьи. Сжатие на основе нейронных сетей – перспективное направление, но для сжатия аудио пока эти технологии применимы слабо (хотя и эффективны при сжатии речи). Различия на слух исходного и восстановленного файлов весьма значительны и нет способа их устранить, т.е. повышение битрейта не ведет к достижению перцептуальной близости восстановленного сигнала к исходному. В этом легко убедиться, прослушав примеры на демо-сайте EnCodec [30]. При этом стоит помнить, что авторы выбрали удачные примеры, но даже в них искажения легко заметны на слух. MP3, Vorbis: в целом эффективные, но устаревшие форматы сжатия аудио с потерями. Исходный и восстановленный файл достаточно легко различить на слух. Opus – перспективный формат, составлявший серьезную конкуренцию AAC, но в настоящее время не развивается и поэтому отстал от AAC. Различить на слух исходный и восстановленный файл возможно (особенно на низких битрейтах). AAC – наиболее совершенный формат сжатия с потерями на сегодняшний день. Исходный и восстановленный сигнал на слух достаточно близки, но объективная оценка (например, по мере PEAQ) сразу выявляет существенные искажения в восстановленном сигнале. Кроме того, в ряде случаев нарушаются психоакустические принципы

восприятия звука человеком с целью снижения необходимого бюджета бит (например, когда удаляются верхние частоты на низких битрейтах). Теоретического доказательства перцептуального равенства исходного и восстановленного сигнала авторы данного метода также не приводят.

Из приведенного анализа следует, что, несмотря на высокую эффективность современных методов сжатия аудиоданных с потерями, ни один из них не гарантирует перцептуального равенства исходного и восстановленного звукового сигнала.

Таким образом, возникает необходимость развития теории сжатия аудиоданных с потерями, позволяющей осуществлять высокоэффективное сжатие аудиоданных с потерями на основе психоакустических принципов восприятия звука человеком и гарантировать перцептуальное равенство исходного и восстановленного аудиосигнала на любых битовых скоростях, поддерживаемых современными стандартизованными аудиокодеками.

3. Уточненная теория сжатия аудиоданных с потерями. Достаточно очевидно, что целью развития теории сжатия аудиоданных с потерями является достижение равенства исходного и восстановленного сигналов. Так как речь идет о сжатии с потерями, то равенство в строгом математическом смысле невозможно и можно говорить лишь о перцептуальном (психоакустическом) равенстве исходного и восстановленного сигналов.

Рассмотрим следующую теорему.

3.1. Теорема о перцептуальном равенстве исходного и восстановленного аудиосигналов.

Теорема: При сжатии аудиоданных с потерями достижение перцептуального равенства исходного и восстановленного аудиосигналов достижимо при разбиении спектра сжимаемого аудиосигнала на субполосы, соответствующие требованиям психоакустики, и выполнении условия неизменности соотношения энергии и фазы исходного и обработанного (сжатого) сигнала в каждой субполосе.

В данном случае фазу необходимо рассматривать как величину, определяющую уровень звукового давления известной звуковой волны в текущий момент времени в измеряемой точке пространства. Другими словами, фаза определяет (совместно с амплитудой) энергию сигнала для каждой частотной группы в данный момент времени, а т.к. длительность фрейма не превышает 20 мс. (что существенно меньше времени реакции слухового аппарата человека на раздражитель), то и среднюю энергию по фрейму допустимо считать мгновенной с точки

зрения перцептуального равенства исходного и восстановленного аудиосигналов.

Стоит отметить, что при переходе в спектральную область при помощи ДКП или МДКП в каждый коэффициент такого разложения включена как энергия, так и фаза сигнала. Это является одной из основных причин, почему в сжатии аудиосигналов МДКП является на сегодняшний день наиболее широко применяемым разложением.

Второй причиной является то, что МДКП оперирует в действительной области, а не в комплексной (как, например Дискретное преобразование Фурье (ДПФ)), что существенно упрощает машинные вычисления с применением данного разложения.

Применение МДКП (или любого другого преобразования сохраняющего информацию о фазе сигнала) для перевода аудиосигнала в спектральную область при его сжатии является необходимым условием, т.к. восстановление сигнала без информации о его фазе внесет значительные искажения в восстановленный сигнал и является недопустимым с точки зрения предлагаемой ТСАП.

Именно поэтому методы, использовавшие в качестве разложения ДПФ с последующим вычислением энергетического спектра (и как следствие – потерей фазы), впоследствии были заменены на методы, сохраняющие информацию о фазе, в том числе это касается и методов на основе нейронных сетей. Более того, для восстановления фазы сигнала был разработан специальный метод, именуемый PHASEN [31], и ряд более современных его аналогов, получивший широкое применение в сфере улучшение качества речи с применением нейросетевых подходов.

Таким образом, сохранение фазы сигнала так же является важной научной задачей, обеспечивающей перцептуальное равенство исходного и восстановленного сигналов, а в случае многоканального сигнала – еще и все бинауральные эффекты человеческого слуха, например позиционирование источника сигнала в пространстве.

Доказательство:

Исходя из взаимной однозначности представления сигнала в спектральном и временном доменах

$$X(n)_{исх} = X(n)_{вост}, \quad (1)$$

если

$$F(n)_{исх} = F(n)_{вост}, \quad (2)$$

где X – цифровой сигнал, а F – спектр этого сигнала (здесь и далее большими буквами обозначены вектора вида x(t), т.к. дискретный аудиосигнал – это вектор амплитуды от времени), $n = 1, 2, \ldots, N$, где N – количество временных/спектральных отсчетов.

$$F = \left(f_1, f_2, \ldots, f_N \right), \tag{3}$$

где f – спектральные коэффициенты (в нашем случае – МДКП), N – количество спектральных коэффициентов.

Для достижения (2) воспользуемся тем, что в пределах критической полосы слух интегрирует возбуждение по частоте и не различает тонкой структуры возбуждения. На этом, в частности, строится эффект маскирования в спектральной области. Другими словами, спектр звукового сигнала может рассматриваться как набор сумм значений коэффициентов МДКП (включающих энергию и фазу сигнала) в частотных группах, соответствующих критическим полосам слуха.

Напомним, что при восприятии звука слуховой аппарат человека разделяет его на частотные группы, называемые критическими полосами слуха. В диапазоне от 20 до 22050 Гц число критических полос равно 25. Ширина этих полос меняется от низких к высоким частотам нерегулярным образом.

Таким образом, спектр сигнала можно рассматривать как 25 значений энергии сигнала в соответствующих критических полосах слуха. Для получения указанных значений необходимо применить соответствующее преобразование (например, МДКП) с последующей группировкой частот в соответствии с неравнополосной шкалой БАРК (Bark-scale).

На основании вышеизложенного сформулируем первый Принцип теории сжатия аудиоданных с потерями.

Принцип 1. Необходимо использовать субполосное разделение спектра сжимаемого аудиосигнала в соответствии с критическими полосами слуха (желательно в соответствии с Барк-шкалой), а не на субполосы произвольной ширины.

Стоит отметить, что данный подход применяется во многих современных аудиокодеках, что только подтверждает корректность предлагаемой теории сжатия аудиоданных с потерями.

При переходе к рассмотрению спектра сигнала в субполосном представлении (3) можно записать как

$$F = (Z_1, Z_2, \ldots, Z_R), \quad (4)$$

где $Z_i = (f_1, f_2, \ldots, f_L)$, где L – количество частотных компонент в субполосе, R – количество субполос.

На основе Принципа 1 формулу (4) можно записать как

$$F = (z_1, z_2, \ldots, z_R), \quad (5)$$

где $z_i = \sum_{j=1}^{L} f_j^2$, т.е. сумма спектральных компонент в субполосе.

Как уже отмечалось выше, в данном случае речь идет об энергии сигнала в конкретный момент времени, а она определяется амплитудой и фазой сигнала. Таким образом, фаза сигнала всегда присутствует в энергии каждой субполосы в неявном виде, поэтому не требуется ее оценивать дополнительно. Это объясняет тот факт, почему во всех источниках, относящихся к перцептуальному сжатию аудиосигналов (например, [5]) речь всегда идет только об энергии сигнала и не упоминается о его фазе.

Так же отметим, что т.к. фаза не вычисляется в явном виде (а только влияет на энергию), то возведение спектрального компонента в квадрат не может на нее повлиять. Кроме того, МДКП (в отличие от ДПФ) позволяет легко восстановить значение спектрального компонента после его возведения в квадрат – необходимо тривиально взять квадратный корень (т.к. это действительное число) и получить знак из дополнительного потока знаков. Достаточно очевидно, что знаки МДКП коэффициентов необходимо сохранять и включать в сжатый аудиопоток (файл).

В качестве альтернативы МДКП уместно рассмотреть пакетное дискретное вейвлетное преобразование (ПДВП), изучению которого посвящено достаточное количество работ. Указанный подход к переходу в спектральную область позволяет получить частотные группы в соответствии с Барк-шкалой и поэтому в полной мере соответствует Принципу 1.

Перцептуальное сжатие аудиоданных с применением ПДВП так же достаточно изученная область знаний. Значительный вклад в ее развитие внесли д.т.н. Петровский А.А. [17], Фадеев Д.Р., д.т.н. Рогозинский Г.Г. [32], и другие авторы. В упомянутых работах доказано, что ПДВП применимо в качестве базиса для использования в психоакустической модели (ПАМ) и авторы успешно строят ПАМ на его основе.

Вместе с тем, несмотря на полное соответствие Принципу 1 и доказанную возможность строить ПАМ с применением ПДВП, именно вопрос точного сохранения энергии и фазы сигнала при применении ПДВП (например, во время квантования по уровню ПДВП коэффициентов) авторами не рассматривался, и более того в работе [32] идет прямое противопоставление пространства время-частота (используемому, например, в МДКП) пространству время-масштаб при использовании ПДВП.

Данное обстоятельство не позволяет прямо сейчас рекомендовать ПДВП как базис перехода в спектральную область в соответствии с Принципом 1 (о причинах этого будет сказано ниже в формуле (6)), но в случае проведения новых исследований, доказывающих его применимость, планируется вернуться к его рассмотрению.

Если потребовать при осуществлении любых изменений спектра при сжатии аудиосигнала с потерями выполнения условия

$$\frac{z_i}{\tilde{z}_i} = const = a, \tag{6}$$

где $\tilde{z}_i$ – сумма спектральных компонент в субполосе сжатого (искаженного) аудиосигнала, то спектр сжатого аудиосигнала в соответствии с (5) будет равен

$$F_{сж} = \left(\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_R\right), \tag{7}$$

где $\tilde{z}_i = \frac{z_i}{a}$, где a=const (одинаковая для всех субполос), следовательно, спектр восстановленного сигнала будет равен

$$F_{вост} = \left(a\tilde{z}_1, a\tilde{z}_2, \ldots, a\tilde{z}_R\right), \tag{8}$$

т.е.

$$F_{вост} = \left(a\frac{z_1}{a}, a\frac{z_2}{a}, \ldots a\frac{z_R}{a}\right) = \left(z_1, z_2, \ldots, z_R\right) \Rightarrow, \tag{9}$$

$$F_{вост} = F_{исх}. \tag{10}$$

Таким образом, соблюдение требования (6) во время сжатия аудиосигнала с потерями обеспечивает выполнение (2), т.е. исходный сигнал будет равен восстановленному в перцептуальном смысле.

Более того, так как энергия сигнала в субполосе определяется в том числе его фазой (точнее это суперпозиция сигналов соответствующих частот входящих в указанную субполосу, каждый из которых обладает своей фазой), то требование (6) позволяет так же сохранять фазовые соотношения между субполосами, что позволяет нам говорить не только о равенстве энергий исходного и восстановленного сигналов (с точностью до множителя), но и о равенстве их фаз, что однозначно определяет исполнение (1).

Стоит отметить, что когда речь идет о сжатии с потерями, то равенство исходного и восстановленного сигналов (2) принято понимать как приближенное (даже в перцептуальном смысле мы можем говорить только о «прозрачности» – т.е. о неотличимости сигналов на слух)

$$X_{вост} \approx X_{исх}, \tag{11}$$

поэтому тот факт, что в общем случае на реальных аудиоданных точного выполнения (6) достичь не удается, говорит о том, что (6) более уместно записать как

$$\frac{z_i}{\tilde{z}_i} = a \pm \varepsilon, \ \ где\ \varepsilon \gg a. \tag{12}$$

что не идет в разрез с общепринятыми в науке определениями.

Другими словами, т.к. речь идет о сжатии с потерями, то выполнение (6) в смысле выполнения (12) является достаточным, а поиск точного решения для (6) представляется избыточным и противоречащим самой идее сжатия с потерями, т.к. смысл сжатия с потерями состоит в передаче более важных с психоакустической точки зрения компонент сигнала (временных или спектральных) и отбрасывании менее важных в соответствии с требуемым субъективным качеством восстановленного сигнала, что подразумевает возможность достижения только приближенного равенства исходного и восстановленного сигналов в смысле (11). Вместе с тем, нельзя не отметить тот факт, что ни один из описанных в литературе современных методов сжатия не выполняет (6) в своей

работе (даже в смысле (12)), а следовательно, и (2) (даже в смысле (11)) также не выполняется.

Предлагаемая теория сжатия аудиоданных с потерями призвана устранить указанные несовершенства современных методов сжатия аудиоданных с потерями и позволить достигать (2), в смысле (11). В дальнейшем, для упрощения вычислений, решения будут приведены в смысле (6) и (2), т.к. это не влияет на результат вычислений.

Таким образом, на основании вышеизложенного доказательства сформулируем второй Принцип теории сжатия аудиоданных с потерями.

Принцип 2. На всех этапах процедуры сжатия-восстановления аудиосигнала необходимо выполнение (6), т.к. это в свою очередь является необходимым и достаточным условием выполнения (2).

Теорема доказана.

3.2. Уменьшение битовых представлений аудиоданных с соблюдением перцептуального равенства исходного и восстановленного аудиосигналов. Стоит отметить, что уменьшение битовых представлений аудиоданных (аудиосигналов) возможно исключительно за счет максимально возможного снижения энтропии сжатой (обработанной) последовательности и последующего применение статистического сжатия (например, арифметического кодирования).

Снижения энтропии сжатого аудиофайла возможно за счет квантования по уровню отсчетов исходного сигнала или его спектральных компонент (в том числе и квантования к нулю).

3.2.1. Разделение на тональные и шумоподобные компоненты. С точки зрения психоакустики удаление или обнуление произвольных отсчетов сигнала или его спектральных компонент совершенно неверно. Исключением являются спектральные компоненты, лежащие ниже абсолютного порога слышимости, которые не воспринимаются человеком в принципе ни при каких условиях – поэтому их можно сразу отнести к шумоподобным, и это не скажется на достижении перцептуального равенства исходного и восстановленного аудиосигналов, особенно учитывая тот факт, что энергия подобных компонент крайне незначительна и подобные компоненты есть в каждой субполосе, и их энергия в достаточной мере уравновешивает друг друга.

Для всех остальных спектральных компонент звукового сигнала, допустимо только квантовать отсчеты сигнала или спектральные компоненты в пределах, которые будут незаметны для

слушателя восстановленного аудиосигнала, базируясь на «эффекте маскировки».

Во всех современных методах сжатия аудиоданных с потерями для использования данного эффекта строят так называемые кривые маскировки, позволяющие определить, маскируется ли данный отсчет или спектральный коэффициент соседними или нет и насколько сильно возможно его проквантовать.

Однако, в реальной практике современных аудиокодеков в случае недостатка бюджета бит в процессе квантования по уровню при достижении кривыми маскировки уровня, который не позволяет присвоить квантованному значению никакого уровня, кроме нуля, возникает необходимость зануления части спектральных компонент.

Это является важным нарушением принципов психоакустики и подлежит устранению.

Кроме того, важным является тот факт, что принятие решения об отбрасывании (обнулении) текущего компонента осуществляется на основе эвристически полученных коэффициентов (используемых при построении кривых маскировки), значения которых вызывают существенные разночтения (проще говоря – они у каждой психоакустической модели свои, и их значения различаются значительно, например, у психоакустических моделей MPEG AAC и Dolby AC-4 обе из которых являются современными и эффективными решениями), что позволяет делать вывод о не точном соответствии психоакустической теории.

Прежде чем говорить о способах преодоления зануления спектральных компонент на низких битрейтах, поговорим о корректной с точки зрения психоакустики маскировке спектральных компонент.

Если в процессе проведения процедуры сжатия разделять субполосы в соответствии с критическими полосами слуха, то решение задачи маскировки достаточно очевидно – любой спектральный компонент в пределах субполосы маскируется спектральными компонентами с большей энергией.

Это обусловлено самим определением критической полосы слуха, т.к. в ней кривая маскировки любого маскера линейна (находится на одном уровне, без убывания), а убывание возникает за пределами критической полосы слуха (собственно, так и возникло понятие «критической полосы слуха» – ширины спектра, где кривая маскировки имеет линейный вид). Наличие линейной части в кривой маскирования представляется крайне важным, т.к. полное описание эффекта маскировки в критической полосе слуха позволяет построить

перцептуально обоснованную теорию сжатия аудиоданных с потерями.

Таким образом, каждый компонент спектра в субполосе маскируется главным маскером, и все, что лежит ниже кривой маскирования главного маскера, можно рассматривать как шумоподобные компоненты.

Важным моментом является то, что при таком подходе не происходит зануления спектральных компонент, т.к. все шумоподобные компоненты будут закодированы одним значением, вычисляемым как

$$Z_n = \frac{\sum f_{nz}}{N}, \tag{13}$$

где N – количество шумоподобных компонент в субполосе, f_{nz} – шумоподобные спектральные компоненты в субполосе

Данный подход представляется обоснованным, т.к. согласно, например, [5] в пределах критической полосы слух реагирует лишь на общую энергию шума, не учитывая его распределение.

Таким образом, представление спектральных компонент с низкой энергией в пределах критической полосы слуха как шумоподобных и дальнейшая замена их на среднее значение их энергии полностью согласуется с принципами психоакустики. На этой основе сформулируем Принцип 3 теории сжатия аудиоданных с потерями.

Принцип 3. В целях выполнения (6) и, как следствие, (2), необходимо не допускать зануления спектральных компонент с малой энергией, а вместо этого, в случае нехватки бюджета бит для кодирования всех спектральных компонент с необходимой с точки зрения психоакустики точностью (т.е. после преодоления психоакустического предела сжатия), кодировать их как шумоподобные, т.е. одним значением (средним значением их энергии), вычисляемой при помощи (13) с сохранением знака.

Остается показать, каким образом осуществить указанную операцию, чтобы не был нарушен принцип (6).

С точки зрения психоакустической теории, указанные спектральные компоненты с малой энергией допустимо отнести к шумоподобным, и, как следствие, закодировать только средним значением, если их энергия ниже кривой маскирования главного маскера.

Уровень кривой маскирования главного маскера (M_i) в децибелах по отношению к значению самого маскера можно вычислить по следующей формуле (14), позволяющей рассчитать уровень маскирования шума тоном [1]:

$$M_i = -1,525 - 0,175 * z_i - 0,5, \quad (14)$$

где z_i – номер текущей субполосы.

Как уже отмечалось ранее, в качестве дополнительного шага можно сразу отнести к шумоподобным спектральные компоненты, лежащие ниже абсолютного порога слышимости, как не воспринимаемые человеческим слухом.

В общих чертах, в соответствии с предлагаемой теорией сжатия с потерями, процедура выделения шумоподобных спектральных компонент должна выглядеть следующим образом:

В случае нехватки бюджета бит выполняется следующее:

1. Ранжирование спектральных компонент в субполосе по модулю.
2. Помечание части спектральных компонент, лежащих ниже кривой маскирования главного маскера M_i (в соответствии с (14)) как шумоподобных, начиная с наименьшего, до тех пор, пока битовая скорость потока не будет соблюдена.
3. Замена помеченных шумоподобных спектральных компонент на среднее значение их энергии и (если это предусматривает метод сжатия) квантование по уровню полученного значения.

Следовательно, т.к. после применения процедуры выделения шумоподобных спектральных компонент во всех субполосах будет выполнено (6), то это позволяет говорить о выполнении (2).

Количество шумоподобных спектральных компонент определяется целевым битрейтом, т.е. тем бюджетом бит, которым обладает энкодер: чем ниже битрейт, тем их берем больше, т.к. это очевидно позволяет уменьшить расход бит за счет передачи не точных значений конкретных спектральных коэффициентов, а только одного значения Z_n для всех. Напомним, что современные аудиокодеки эти компоненты просто зануляют.

Кстати говоря, степень сжатия при предлагаемом подходе уменьшиться незначительно, т.к. нужно дополнительно передавать только значение Z_n для каждой субполосы. При этом перерасход бит компенсируется за счет более высокого субъективного качества получаемого сигнала, что позволит, в свою очередь, применить

квантование по уровню спектральных компонент на меньшее число уровней квантования.

Важно, что, т.к. имеет место замена спектральных компонент на среднее значение их энергии, то эта процедура не влияет на корректность процедуры сжатия в смысле выполнения (6).

Однако, (6) будет необратимо нарушено, если в шумоподобные компоненты попадут маскеры, т.е. спектральные компоненты, лежащие выше кривой маскирования главного маскера и не относящиеся к шуму. Это грубое нарушение теории и его необходимо избегать.

Для реализации указанной процедуры прореживания предлагается использовать итерационную процедуру:

Вектор Z_i, отсортированный по возрастанию, будем называть Z_{i_copm}

$$Z_{i\_copm} = \left(f_{copm1}, f_{copm2}, \ldots f_{copmL}\right), \tag{15}$$

где L – количество f в субполосе.

Достаточно очевидно, что

$$z_i = z_{i\_copm} = \sum\nolimits_{j=1}^{L} f_j^2. \tag{16}$$

Для определения шумоподобных компонент в соответствии с (14) вычисляем реальное значение кривой маскирования с использованием обратной формулы для вычисления децибелов:

$$\tilde{M}_i = f_{max}^2 * \sqrt[20]{M_i}, \tag{17}$$

Следовательно, для выполнения (6) выполняем итерации по j пока

$$f_{copm_{ij}}^2 \le \widetilde{M}_i, \tag{18}$$

Если

$$f_{copm_{ij}}^2 > \widetilde{M}_i, \tag{19}$$

то $j = j - 1$, т.е. отмена пометки текущего спектрального коэффициента как шумоподобного и конец итерационной процедуры.

Следовательно, после (19) получаем вектор Z_{copm}

$$Z_{copm} = \left(f_{copmk}, f_{copmk+1}, \ldots, f_{copmL} \right), \tag{20}$$

где $k = j$, т.е. последнему успешному шагу итерационной процедуры.

Проводим обратную сортировку, т.е. возвращаем f на исходные места и завершаем итерационную процедуру.

Далее, как уже говорилось, применяем (13) и проводим замену всех шумоподобных спектральных компонент на среднее значение их энергии (Zn).

После выполнения процедуры получаем вектор Z_i с энергией, равной исходному значению в строгом соответствии с (6), т.к. замена значений энергии каждого из шумоподобных спектральных компонент на среднее значение их энергии не вносит изменения в энергию в субполосе.

Еще раз отметим тот факт, что указанную процедуру кодирования шумоподобных компонент стоит проводить только в случае нехватки бюджета бит (когда современные кодеки начитают занулять спектральные компоненты).

Это обусловлено тем, что хотя с точки зрения психоакустики маскирование шума тоном и реакция слуха человека на общую энергию шума и не вызывают сомнений, но сама структура музыкального (звукового) сигнала не всегда включает реальную шумовую составляющую, для которой было рассчитано соотношение (14), например, когда в сигнале присутствует только один чистый тон, то никакого реального шума в нем нет и Принцип 3 неприменим. При этом достаточно очевидно, что для кодирования такого простого сигнала и не может потребоваться Принцип 3 – бюджет бит позволяет ограничиться остальными Принципами.

3.2.2. Квантование по уровню. Если проводится квантование по уровню Z_n, то необходимо руководствоваться тем же Принципом, как и при квантовании маскеров (тональных компонент):

Принцип 4. При квантовании по уровню спектральных компонент аудиосигнала необходимо, чтобы погрешность квантования не превышала едва заметных изменений звука в смысле изложенном Цвайкером [33] (это значение лежит в диапазоне 1-2 дБ в зависимости от энергии спектрального компонента) и чтобы квантованная последовательность соответствовала (6). Метод квантования при этом

может быть любым – это не влияет на предлагаемую ТСАП, главное – это выполнение (6).

В статье применен метод квантования по уровню с адаптивным шагом квантования, позволяющий естественным образом контролировать погрешность квантования на каждом уровне и держать ее в диапазоне 1-2 дБ, что гарантирует выполнение (6). Разумеется, что значения шагов квантования необходимо так же передавать на приемную сторону. Детали метода квантования по уровню не имеют важного значения для предлагаемой Теории, поэтому выходят за рамки настоящей статьи.

3.3. Общая схема Аудиокодека на основе ТСАП. Аудиокодек на основе перцептуального равенства исходного и восстановленного звукового сигнала представляет собой программно-алгоритмическую реализацию указанной ТСАП и использует все 4 указанных Принципа в явном виде. Разработанный аудиокодек в настоящее время работает на битовых скоростях 32, 64, 128, 195 кбит/с (для моносигнала), в моно, стерео и многоканальных режимах. Сжатие аудиосигнала с потерями на любой битовой скорости может осуществляться как независимо для каждого канала, так и с учетом общего битрейта в многоканальном сигнале, что позволяет более эффективно использовать доступный бюджет бит и кодировать более сложные каналы большим числом бит.

Общая схема работы предлагаемого метода сжатия аудиоданных с потерями (аудиокодека) на основе ТСАП представлена на рисунке 1.

Легко видеть, что в структуре кодека отсутствует модуль психоакустической модели. Как было показано выше, аудиокодеки, построенные на основе ТСАП, не нуждаются в данном модуле, т.к. перцептуальное равенство исходного и восстановленного сигналов обеспечено выполнением (12) и Принципов ТСАП, которые в свою очередь обеспечивают выполнение психоакустических основ восприятия звука человеком и позволяют выполнить (12).

Указанное обстоятельство существенно снижает вычислительную сложность разработанного кодека по сравнению с современными кодеками (такими как AAC) и при этом (как будет показано далее) позволяют получить более высокое качество восстановленного сигнала. Все это является следствием применения ТСАП.

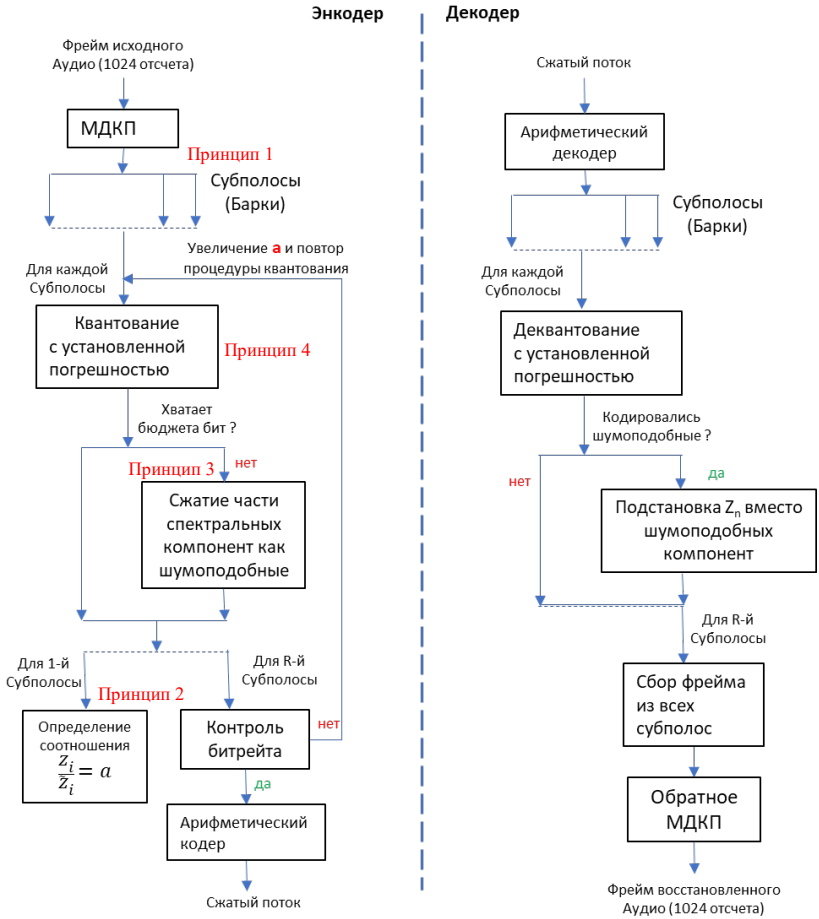


Рис. 1. Схема работы аудиокодека

Несложно заметить, что Все Принципы ТСАП и (12) нашли свое отражение на схеме:

– Переход в спектральную область и субполосное разложение в соответствии с Барк шкалой (Принцип 1).

– Квантование по уровню спектральных компонент в соответствии с (12), т.е. чтобы общая энергия квантованной последовательности в субполосе отличалась от исходного значения энергии в субполосе на множитель «a» (Принцип 4).

– Кодирование шумоподобных спектральных компонент средним значением их энергии (Принцип 3).

– Контроль выполнения соотношения (12) для каждой из субполос. Стоит отметить, что Ɛ выбирается в соответствии с битрейтом и должно быть существенно меньше «a». Разумно определить Ɛ как долю от «a» (Принцип 2).

Точная оценка вычислительной сложности разработанного аудиокодека будет выполнена в ходе дальнейших исследований, но даже приблизительная оценка этапов сжатия показывает потенциальное снижение вычислительной сложности по сравнению с современными аудиокодеками, использующими полноценную «психоакустическую модель».

Отметим, что кодек несимметричен, т.к. на стороне энкодера возможно несколько итераций по подбору параметра «a» и возможен поиск шумоподобных компонент сигнала.

4. Вычислительные эксперименты по сравнению объективного качества сжатых аудиофайлов. Для подтверждения работоспособности предложенного в работе Аудиокодека на основе перцептуального равенства исходного и восстановленного звукового сигнала были проведены вычислительные эксперименты по сравнению результатов сжатия с потерями исходного звукового сигнала:

– по технологии Advanced Audio Coding (AAC);

– на основе предлагаемого Аудиокодека на основе перцептуального равенства исходного и восстановленного звукового сигнала.

4.1. Алгоритм вычислительного эксперимента. В качестве исходных сигналов использовались музыкальные файлы различных жанров, направлений и времени издания (около 1000 файлов), подобранные таким образом, чтобы максимально полно представить всю генеральную совокупность музыкальных произведений, записанных и используемых в настоящее время.

В качестве сравнения использовался стандартизованный формат сжатия аудиоданных с потерями (по Рекомендации Международного союза электросвязи ITU-R BS.1196-8 [18], более известный как Advanced Audio Coding (AAC MPEG-4), который в настоящее время является стандартом в отрасли как де-юре, так и де-факто.

Сравнение полученных результатов с результатами стандартизированного кодека Advanced Audio Coding (AAC) осуществлялось в смысле объективного качества

(по стандартизованной мере PEAQ) и субъективного качества восстановленного сигнала, оцениваемого на слух группой экспертов.

Мера PEAQ (Objective Measurements of Perceived Audio Quality) стандартизована Международным союзом электросвязи и является общепризнанной на сегодняшний день как мера оценки объективного качества звукового сигнала. PEAQ измеряется в условных единицах, которые соответствуют MOS (Mean opinion score – усредненное мнение экспертов) и лежит в диапазоне от 5,0 – качество оригинала, до 1,0 – качество ужасно (невозможно или очень трудно узнать оригинальный сигнал) – Рекомендация Международного союза электросвязи ITU-R BS.1284-2 [34].

Необходимо отметить, что в реальной практике для оценки объективного качества часто используется шкала ODG (Objective Difference Grade) [6].

В таблице 1 приведены эквивалентные значения качества по шкалам MOS и ODG, а также описание этого качества, как это приводится в Рекомендациях Международного союза электросвязи [34] и Рекомендации ITU-R BS.1387-2 [6]).

Таблица 1. Эквивалентные значения качества по шкале MOS и ODG

Тип вносимых искажений	**MOS (ITU)**	**ODG**
Незаметные искажения	5,0	0,0
Заметные, но не раздражающие	4,0	−1,0
Немного раздражающие	3,0	−2,0
Раздражающие	2,0	−3,0
Крайне раздражающие	1,0	−4,0

4.2. Анализ метода AAC на соответствие ТСАП. Прежде чем привести результаты вычислительных экспериментов, проведем теоретический анализ технологии AAC на соответствие ТСАП.

Как несложно заметить из [18] и краткого описания, приведенного в разделе 2.2, в целом AAC в достаточной степени соответствует принципам психоакустики, что вполне ожидаемо для лучшего современного коммерческого аудиокодека.

Однако, есть ряд тонких моментов, в которых AAC нарушает ТСАП:

1. Отсутствует контроль выполнения основного соотношения (12), что не позволяет говорить о гарантированном выполнении (2), подробно указанные соотношения рассмотрены в разделе 3.1.

2. Начиная с битрейта 64 кбит/с для каждого канала, метод AAC начинает занулять часть верхних частот, чтобы сэкономить

бюджет бит для нижних и средних частот. Это очевидное нарушение (2), т.к. часть спектра сигнала просто отсутствует, и если для средних битрейтов (например, 64 кбит/с) зануляются только частоты выше 18кГц, (что, разумеется, тоже нарушение, но разницу услышат только молодые люди), то для 32 кбит/с зануляются частоты выше 14кГц, что будет слышно практически всем. Отметим, что аудиокодек на основе ТСАП не имеет указанного недостатка – он всегда кодирует полный спектр сигнала.

3. Несмотря на то, что общая схема AAC в целом соответствует психоакустическим принципам, проблемы начинаются после преодоления психоакустического порога сжатия, т.е. когда имеющегося бюджета бит начинает не хватать для кодирования всех спектральных компонент в соответствии с порогами маскирования. В этот момент авторы AAC вынуждены идти на осознанное нарушение психоакустических принципов и понижать пороги маскирования. Часто это приводит к тому, что многие спектральные компоненты не могут быть проквантованы ни к какому другому уровню кроме нулевого и поэтому зануляются, что является грубым нарушением Принципа 3 ТСАП (подробное описание приведено в разделе 3.2.1).

Таким образом, несложно заключить, что, несмотря на свою очевидную эффективность, метод AAC нарушает ТСАП и не может гарантировать перцептуального равенства исходного и восстановленного сигнала.

4.3. Вычислительный эксперимент. Проведенный вычислительный эксперимент подтвердил справедливость теоретических выводов, приведенных выше.

Сжатие аудиоданных осуществлялось с битовой скоростью 390 кбит/с для стереосигнала или 195 кбит/с для каждого канала кодируемого независимо и 64 кбит/с для стереосигнала или 32 кбит/с для каждого канала кодируемого независимо.

В качестве примера приведем результаты 10 специально выбранных тестовых файлов, которые были определены группой MPEG как рекомендуемые для тестирования аудиокодеков.

4.3.1. Вычислительный эксперимент на битовой скорости 390 кбит/с. В Таблице 2 представлены результаты ODG для каждого сжатого звукового файла для битовой скорости 390 кбит/с, сравнение с результатами AAC на равном битрейте, разность между AAC и предлагаемым методом сжатия с потерями в абсолютных значениях и процентах, а также усредненная оценка по всем файлам.

Таблица 2. Результаты сопоставительного анализа результатов сжатия аудиофайлов при помощи кодека AAC MPEG-4 и предлагаемого аудиокодека по шкале ODG для битрейта 390кбит/с

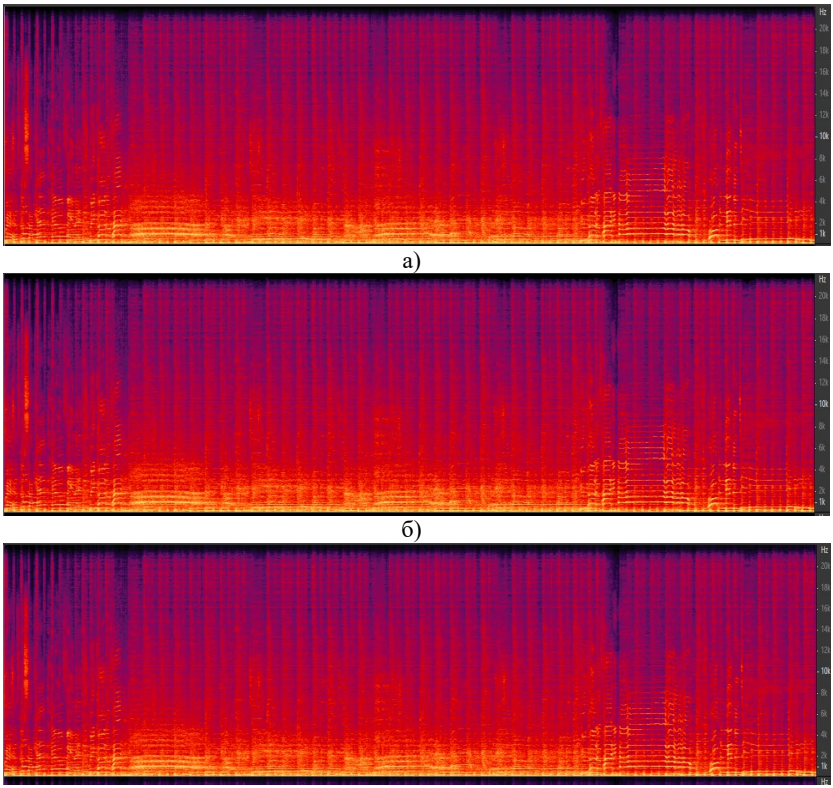
№	Имя файла	AAC, [ODG]	Предлагаемый метод, [ODG]	Разность (Предлагаемый – AAC), [ODG]	Разность, [%]
1	Adel	-0,076	-0,034	0,042	55,26
2	Avril_Lavil	-0,062	-0,024	0,038	61,29
3	Elvis_Prestly	-0,124	-0,12	0,004	3,23
4	Evanescence	-0,064	-0,06	0,004	6,25
5	Hotel_California	-0,148	-0,146	0,002	1,35
6	Jump_in_my_car	-0,118	-0,094	0,024	20,34
7	NewYorkCity	-0,074	-0,065	0,009	12,16
8	Poison	-0,069	-0,041	0,028	40,58
9	SlavesOfFear	-0,054	-0,026	0,028	51,85
10	You_are_yong	-0,075	-0,073	0,002	2,67
Среднее по всем:		**-0,086**	**-0,068**	**0,018**	**25,50**

Несложно заметить, что, по усредненной оценке, преимущество предлагаемого аудиокодека составило 25%, а для ряда файлов превысило 50%, при этом важно помнить, что речь идет о высоких битрейтах, где у современных методов сжатия аудиоданных с потерями нет проблем с бюджетом бит для кодирования сжатого звукового потока.

Таким образом, доказано, что применение Аудиокодека на основе перцептуального равенства исходного и восстановленного звукового сигнала позволяет превзойти результаты современных аудиокодеков даже на высоких битрейтах, что в современной науке считается малодостижимой задачей, т.к. при наличии достаточного бюджета бит AAC нарушает ТСАП только в части отсутствия контроля выполнения (12), а остальных нарушений в явном виде нет, но даже этого вполне достаточно, чтобы предложенный аудиокодек превзошел AAC.

На рисунке 2 представлены спектрограммы исходного аудиофайла и восстановленных файлов, сжатых при помощи кодека AAC и предлагаемого аудиокодека.

Стоит отметить, что на подобном высоком битрейте отличия восстановленного аудиосигнала от оригинала должны быть крайне незначительны – тогда можно говорить о корректной работе аудиокодека. В нашем случае отличия составляют 0,068 ODG, что крайне мало, и увидеть такое отличие на спектрограмме крайне сложно.



а)

б)

в)

Рис. 2. Сравнение спектрограмм: а) Оригинал, б) AAC, в) Предлагаемый метод (файл №1)

Приведенные спектрограммы зрительно практически неотличимы друг от друга, что подтверждает корректность работы предлагаемого кодека и достижения им качества оригинала.

Таким образом, для корректного сравнения результатов предложенного аудиокодека и формата Advanced Audio Coding (AAC MPEG-4) необходим инструмент более точный, чем зрительное сравнение спектрограмм – именно поэтому мы опираемся на результат объективного оценивания по мере PEAQ, которая показывает явное превосходство предложенного аудиокодека.

Также отметим, что оценка субъективного качества восстановленного звукового сигнала по шкале MOS носила в данном исследовании вспомогательный характер и была направлена на

контроль ошибок со стороны объективной меры оценки качества восстановленного звукового сигнала PEAQ и контроля качества воспроизводимого сигнала на слух.

Для битрейта 390 кбит/с MOS был оценен всеми экспертами как близкий к оригиналу, но вместе с тем большее число экспертов признали результат, полученный на основе предлагаемой теории, более насыщенным и близким к оригиналу.

4.3.2. Вычислительный эксперимент на битовой скорости 64 кбит/с. В Таблице 3 представлены результаты ODG для каждого сжатого звукового файла для битовой скорости 64 кбит/с, сравнение с результатами AAC на равном битрейте, разность между AAC и предлагаемым методом сжатия с потерями в абсолютных значениях и процентах, а также усредненная оценка по всем файлам.

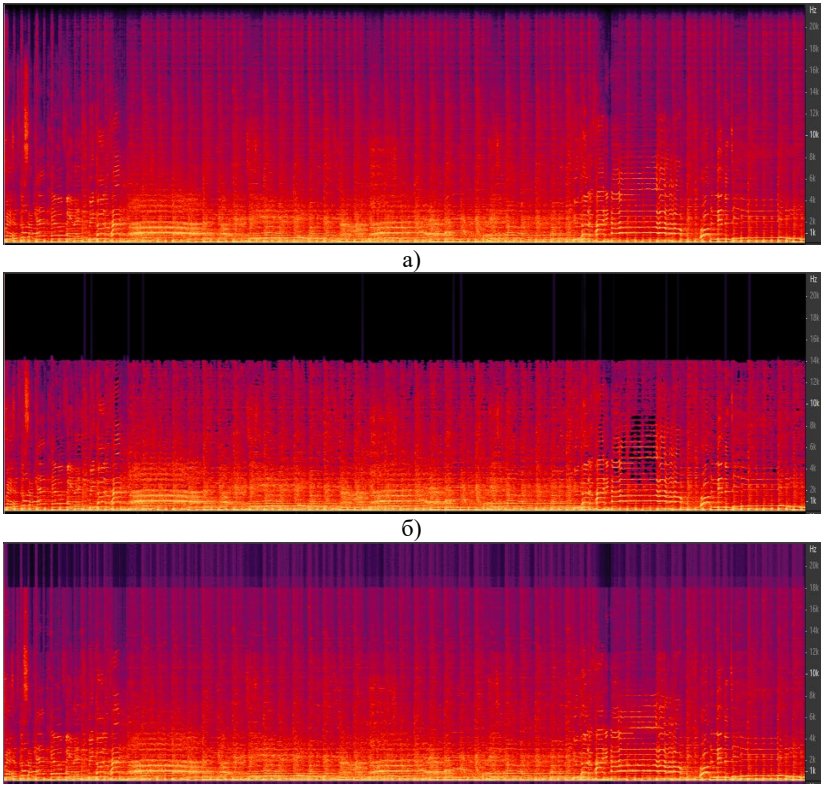
Таблица 3. Результаты сопоставительного анализа результатов сжатия аудиофайлов при помощи кодека AAC MPEG-4 и предлагаемого аудиокодека по шкале ODG для битрейта 64кбит/с

№	Имя файла	AAC, [ODG]	Предлагаемый метод, [ODG]	Разность (Предлагаемый – AAC), [ODG]	Разность, [%]
1	Adel	-3,127	-1,215	1,912	61,14%
2	Avril_Lavil	-2,585	-0,949	1,636	63,29%
3	Elvis_Prestly	-3,139	-2,080	1,059	33,74%
4	Evanescence	-2,895	-1,114	1,781	61,52%
5	Hotel_California	-3,137	-1,726	1,411	44,98%
6	Jump_in_my_car	-3,019	-1,410	1,609	53,30%
7	NewYorkCity	-2,831	-1,260	1,571	55,49%
8	Poison	-2,934	-0,973	1,961	66,84%
9	SlavesOfFear	-2,500	-0,837	1,663	66,52%
10	You_are_yong	-2,838	-1,525	1,313	46,26%
Среднее по всем:		**-2,901**	**-1,309**	**1,592**	**54,87%**

Несложно заметить, что, на низком битрейте, преимущество предлагаемого аудиокодека составило уже 55%, а для ряда файлов превысило 65%. Это значительно превосходит результат на высоких битрейтах и является следствием нарушений ТСАП, о которых говорилось выше.

Указанный результат представляется крайне важным, т.к. повышение перцептуального качества звуковых сигналов на низких и ультранизких битрейтах является важнейшей задачей, стоящей перед современным кодированием аудиоданных с потерями.

На рисунке 3 представлены спектрограммы исходного аудиофайла и восстановленных файлов, сжатых при помощи кодека AAC и предлагаемого аудиокодека.



а)

б)

в)

Рис. 3. Сравнение спектрограмм: а) Оригинал, б) AAC, в) Предлагаемый метод (файл №1)

Отметим, что на низких битовых скоростях искажения на спектрограмме метода AAC легко заметить невооруженным глазом – это полностью черные области, относящиеся к зануленным спектральным компонентам, чего, как показано в ТСАП, быть не должно.

Таким образом, как и было показано теоретически, после преодоления перцептуального предела сжатия, качество современных коммерческих кодеков быстро падает и на битрейте 64кбит/с

становиться уже достаточно низким. Причины подобных проблем подробно описаны в разделе 4.2.

Для битрейта 64 кбит/с субъективная оценка на слух исходного и полученных при помощи предлагаемого аудиокодека аудиосигналов, показала, что большинство экспертов не могут с уверенностью определить оригинал в слепом безреференсном тесте, а в референсном – MOS был оценен экспертами достаточно высоко, как не отличающийся от оригинала и не имеющий существенных искажений.

В свою очередь, результаты формата Advanced Audio Coding (AAC) на данном битрейте уже вызвали нарекания экспертов и, в первую очередь, вызвало недовольство сужение полосы пропускания, т.е. обрезание верхних частот, которое является важной частью алгоритма AAC для низких битрейтов, что обусловлено нехваткой имеющегося бюджета битов для кодирования всей слышимой полосы звукового сигнала.

5. Заключение. Доказано, что предложенный в статье Аудиокодек на основе перцептуального равенства исходного и восстановленного звукового сигнала, базирующийся на новых положениях ТСАП, позволяет достичь перцептуального равенства исходного и восстановленного аудиосигнала, что невозможно для современных методов сжатия аудиоданных с потерями. Кроме того, эти положения ТСАП позволяют создавать понятные и теоретически обоснованные методы сжатия аудиоданных с потерями, соответствующие (6) и как следствие (2). В качестве примера было проведено сжатие аудиоданных с потерями, в ходе которого Аудиокодек на основе перцептуального равенства исходного и восстановленного звукового сигнала показал объективное качество восстановленного сигнала на 25% выше (для битрейта 390 кбит/с, стерео) и 55% выше (для битрейта 64 кбит/с, стерео) по сравнению с наиболее прогрессивным на сегодняшний день форматом AAC MPEG-4.

Стоит отметить, что прогресса в объективном качестве восстановленного аудиосигнала на указанных высоких битовых скоростях не наблюдалось последние 10 лет и высокие битрейты считаются наиболее сложными для улучшения результатов сжатия аудиоданных с потерями. Как показано в статье, современные форматы AAC HE (v1/2) не могут помочь в повышении качества на высоких битрейтах, т.к. используют тот же алгоритм перцептуального сжатия, как и AAC MPEG-4.

Проведенные эксперименты подтверждают тот факт, что метод сжатия аудиоданных, построенный на основе развитой теории сжатия аудиоданных с потерями, позволяет получать более высокие, в смысле объективного качества (по стандартизованной мере PEAQ), результаты в сравнении с лидирующим на сегодняшний день на рынке форматом AAC даже на высоких битрейтах, где качество восстановленных аудиосигналов приближается к качеству оригинала.

Литература

1. Ковалгин Ю.А., Вологдин Э.И. Цифровое кодирование звуковых сигналов // М.: КОРОНА-принт, 2015. 240 с.
2. Журавлёва Л.В., Шишурин А.И. Сравнительный анализ аудиоформатов // Технологии инженерных и информационных систем. 2022. № 2. С. 67–78.
3. Каргин Р.И., Стаценко Л.Г. Форматы сжатия аудиоданных. Анализ и сравнение // Известия СПбГЭТУ ЛЭТИ. 2019. № 9. С. 31–37.
4. Koops H.V., Micchi G., Quinton E. Robust lossy audio compression identification. 2024. arxiv preprint arxiv:2407.21545.
5. Ковалгин Ю.А., Фадеева Д.Р. Исследование психоакустических моделей кодеков с компрессией цифровых аудиоданных // Современная наука: актуальные проблемы теории и практики. Серия: Естественные и технические науки. 2016. № 7. С. 29–38.
6. Официальный сайт ITU. Method for objective measurements of perceived audio quality. Recommendation ITU-R BS.1387-2 (05/2023). URL: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1387-2-202305-I!!PDF-E.pdf (дата обращения: 05.01.2025).
7. Cormen T., Leiserson C., Rivest R., Stein C. Introduction to Algorithms 4th Edition. Cambridge, Massachusetts: The MIT Press, 2022. 1312 p.
8. Чугунова О.В., Буслова В.Е. Архивации данных методами Шеннон-Фано и Хаффмана // Актуальные проблемы науки и техники: Сборник статей по материалам международной научно-практической конференции. Уфа: Изд. НИЦ Вестник науки, 2020. С. 58–65.
9. Сергеев И.С., Балакирев Н.Е. Сравнение алгоритмов сжатия звуковой информации алгоритмом Хаффмана и арифметическим кодированием // Наукосфера. 2022. № 8-2. С. 31–35.
10. Официальная страница проекта FLAC. URL: https://sourceforge.net/projects/flac/ (дата обращения: 05.01.2025).
11. Salomon D. Data compression: the complete reference 4th Edition // London: Springer-Verlag. 2007. 1117 c.
12. Официальный сайт Monkey’s Audio (Спецификация Monkey’s Audio). URL: https://www.monkeysaudio.com/index.html (дата обращения: 05.01.2025).
13. Официальный сайт проекта ALAC (Спецификация Apple Lossless Audio Codec). URL: https://macosforge.github.io/alac/ (дата обращения: 05.01.2025).
14. Официальный сайт Microsoft (Windows Media Player). URL: https://apps.microsoft.com/detail/9WZDNCRFJ3PT?hl=en-us&gl=US (дата обращения: 05.01.2025).
15. Официальный сайт Xiph.Org фонда (Спецификация Vorbis I Xiph.Org). URL: https://xiph.org/vorbis/doc/Vorbis_I_spec.html (дата обращения: 05.01.2025).
16. Официальный сайт Opus Interactive Audio Codec. URL: https://opus-codec.org/ (дата обращения: 05.01.2025).

17. Петровский Ал.А., Петровский А.А. Масштабируемые аудиоречевые кодеры на основе адаптивного частотно-временного анализа звуковых сигналов // Труды СПИИРАН. 2017. № 1(50). С. 55–92. DOI: 10.15622/sp.50.3.
18. Официальный сайт ITU. Audio coding for digital broadcasting. Recommendation ITU-R BS.1196-8 (10/2019). URL: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1196-8-201910-I!!PDF-E.pdf (дата обращения: 05.01.2025).
19. Jenrungrot T., Chinen M., Kleijn W.B., Skoglund J., Borsos Z., Zeghidour N., Tagliasacchi M. LMcodec: a Low Bitrate Speech Codec With Causal Transformer Models // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095442.
20. Shi H., Shimada K., Hirano M., Shibuya T., Koyama Y., Zhong Z., Takahashi S., Kawahara T., Mitsufuji Y. Diffusion-Based Speech Enhancement with Joint Generative and Predictive Decoders // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. pp. 12951–12955. DOI: 10.1109/ICASSP48485.2024.10448429.
21. Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis // Proceedings of the 34th Conference on Neural Information Processing Systems (NIPS). 2020. vol. 33. pp. 17022–17033.
22. Kaneko T, Tanaka K., Kameoka H., Seki S. Istftnet: Fast and Lightweight Mel-spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform. 2022. arxiv preprint arXiv:2203.02395v1.
23. Subramani K., Valin J.-M., Isik U., Smaragdis P., Krishnaswamy A. End-to-end LPCNet: A Neural Vocoder With Fully-Differentiable LPC Estimation // Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). 2022. pp. 818–822.
24. Valin J.-M., Skoglund J. LPCNet: Improving Neural Speech Synthesis Through Linear Prediction // Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. pp. 5891–5895. DOI: 10.1109/ICASSP.2019.8682804.
25. Valin J.-M., Isik U., Smaragdis P., Krishnaswamy A. Neural Speech Synthesis on a Shoestring: Improving the Efficiency of LPCNet // Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 8437–8441.
26. Valin J.-M., Buthe J., Mustafa A. Low-Bitrate Redundancy Coding of Speech Using a Rate-distortion-optimized Variational Autoencoder // Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096528.
27. Zeghidour N., Luebs A., Omran A., Skoglund J., Tagliasacchi M. SoundStream: An End-to-End Neural Audio Codec // Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. vol. 30. pp. 495–507.
28. Du Z., Zhang S., Hu K., Zheng S. FunCodec: A Fundamental, Reproducible and Integrable Open-Source Toolkit for Neural Speech Codec. 2023. arxiv preprint arXiv:2309.07405v1.
29. Defossez A., Copet J., Synnaeve G., Adi Y. Funcodec: High Fidelity Neural Audio Compression. 2022. arxiv preprint arXiv:2210.13438v1.
30. Демо-сайт проекта High Fidelity Neural Audio Compression (EnCodec). URL: https://ai.honu.io/papers/encodec/samples.html (дата обращения: 18.10.2024).
31. Yin D., Luo C., Xiong Z., Zeng W. PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network. 2019. arxiv preprint arXiv:1911.04697v1.
32. Рогозинский Г.Г. Перцепционное сжатие звука с использованием вейвлетных пакетов // Диссертация СПбГУКиТ. 2010.

33. Zwicker E., Fastl H. Psychoacoustics: Facts and Models // Springer-Verlag, Berlin Heidelberg. 1990.
34. Официальный сайт ITU. General methods for the subjective assessment of sound quality. Recommendation ITU-R BS. 1284-2 (01/2019). URL: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1284-2-201901-I!!PDF-E.pdf (дата обращения: 05.01.2025).

Чижов Илья Игоревич — канд. техн. наук, ведущий инженер ключевых проектов, руководитель команды речевого и аудиокодирования, Российский исследовательский институт Huawei. Область научных интересов: сжатие аудиоданных, цифровая обработка сигналов, машинное/глубокое обучение. Число научных публикаций — 22. aproximation18@yandex.ru; улица Крылатская, 17/2, 121614, Москва, Россия; р.т.: +7(495)234-0686.

DOI 10.15622/ia.24.2.3

I. CHIZHOV

AN AUDIOCODEC BASED ON THE PERCEPTUAL EQUALITY BETWEEN THE ORIGINAL AND RESTORED AUDIO SIGNALS

Chizhov I. **An AudioCodec Based on the Perceptual Equality between the Original and Restored Audio Signals.**

Abstract. A method for lossy audio data compression (AudioCodec) is presented. It allows for improving objective quality of the restored audio signal by 25% at a bitrate of 390 kbps and 55% at a bitrate of 64 kbps compared to the AAC MPEG-4 format. The proposed method of audio data compression is based on an advanced theory of lossy audio data compression (TLAC), which is also introduced in the article. The improvement in the objective quality of the reconstructed audio signal (according to the standardized PEAQ measure) is achieved because the TLAC overcomes issues in modern lossy audio data compression methods related to the use of psychoacoustic principles of human sound perception, including after overcoming the "psychoacoustic compression limit" of the audio signal (i.e. the moment in perceptual coding when the available bit budget is insufficient to encode all spectral components with the accuracy required from a psychoacoustic perspective). This allows for achieving perceptual equality between the original and reconstructed audio signals. As an analysis of the state of the art, solutions for both lossless and lossy audio data compression, as well as those using artificial intelligence, are considered. In all modern lossy audio data compression methods, the procedure for selecting the spectral components to be preserved, as well as the permissible quantization error, is carried out through a series of highly complex procedures collectively referred to as the "psychoacoustic model of the lossy audio compression method". In a strict sense, perceptual equality between the spectra of the original and restored signals has not been proven by any research group and, therefore, cannot be guaranteed by them. Independent experts regularly publish tests demonstrating that modern audio codecs have issues with certain audio signals. The article proposes an AudioCodec based on the perceptual equality between the original and restored audio signals, which is based on the new ideas of the theory of lossy audio compression (TLAC). These ideas guarantee the achievement of perceptual equality between the original and restored audio signals at different bitrates, therefore, the AudioCodec built on its basis is free from the above-mentioned issues and, as a result, significantly outperforms modern AudioCodecs in terms of the objective quality of the restored audio signal, as measured by PEAQ.

Keywords: audio data compression, psychoacoustic model, spectrum, subband division, perceptual equality of the spectra.

References

1. Kovalgin Yu.A., Vologdin E.I. Cifrovoe kodirovanie zvukovyh signalov [Digital coding of sound signals]. M.: KORONA-print, 2015. 240 p. (In Russ.).
2. Zhuravleva L.V., Shishurin A.I. [Comparative analysis of audio formats]. Tehnologii inzhenernyh i informacionnyh sistem – Technologies of engineering and information systems. 2022. no. 2. pp. 67–78. (In Russ.).
3. Kargin R.I., Statsenko L.G. [Audio data compression formats. Analysis and comparison]. Izvestija SPbGJeTU LJeTI – Bulletin of ETU LETI. 2019. no. 9. pp. 31–37.
4. Koops H.V., Micchi G., Quinton E. Robust lossy audio compression identification. 2024. arxiv preprint arxiv:2407.21545.

5. Kovalgin Yu.A., Fadeeva D.R. [Study of psychoacoustic models of codecs with digital audio data compression]. Sovremennaja nauka: aktual'nye problemy teorii i praktiki. Serija: Estestvennye i tehnicheskie nauki – Modern science: current problems of theory and practice. Series: Natural and Technical Sciences. 2016. no. 7. pp. 29–38. (In Russ.).
6. Official website of the ITU. Method for objective measurements of perceived audio quality. Recommendation ITU-R BS.1387-2 (05/2023). Available at: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1387-2-202305-I!!PDF-E.pdf. (accessed 05.01.2025).
7. Cormen T., Leiserson C., Rivest R., Stein C. Introduction to Algorithms 4th Edition. Cambridge, Massachusetts: The MIT Press, 2022. 1312 p.
8. Chugunova O.V., Buslova V.E. [Data archiving using Shannon-Fano and Huffman methods] Aktual'nye problemy nauki i tehniki: Sbornik statej po materialam mezhdunarodnoj nauchno-prakticheskoj konferencii [Actual problems of science and technology: Collection of articles based on the materials of the international scientific and practical conference]. Ufa: Izd. NIC Vestnik nauki, 2020. pp. 58–65. (In Russ.).
9. Sergeev I.S., Balakirev N.E. [Comparison of audio compression algorithms using the Huffman algorithm and arithmetic coding]. Naukosfera – Naukosphere. 2022. no. 8-2. pp. 31–35. (In Russ.).
10. Official page of the FLAC project. Available at: https://sourceforge.net/projects/flac/. (accessed 05.01.2025).
11. Salomon D. Data compression: the complete reference 4th Edition. London: Springer-Verlag. 2007. 1117 c.
12. Official website of Monkey's Audio (Monkey's Audio Specification). Available at: https://www.monkeysaudio.com/index.html. (accessed 05.01.2025).
13. Official website of the ALAC project (Apple Lossless Audio Codec Specification). Available at: https://macosforge.github.io/alac/. (accessed 05.01.2025).
14. Official website of Microsoft (Windows Media Player). Available at: https://apps.microsoft.com/detail/9WZDNCRFJ3PT?hl=en-us&gl=US. (accessed 05.01.2025).
15. Official website of the Xiph.Org Foundation (Vorbis I Specification Xiph.Org). Available at: https://xiph.org/vorbis/doc/Vorbis_I_spec.html. (accessed 05.01.2025).
16. Official website of Opus Interactive Audio Codec. Available at: https://opus-codec.org/. (accessed 05.01.2025).
17. Petrovsky A.A., Petrovsky A.A. [A scalable speech and audio coders based on adaptive time-frequency signal analysis. Trudy SPIIRAN – SPIIRAS Proceedings. 2017. no. 1(50). pp. 55–92. DOI: 10.15622/sp.50.3. (In Russ.).
18. Official website of ITU. Audio coding for digital broadcasting. Recommendation ITU-R BS.1196-8 (10/2019). Available at: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1196-8-201910-I!!PDF-E.pdf. (accessed 05.01.2025).
19. Jenrungrot T., Chinen M., Kleijn W.B., Skoglund J., Borsos Z., Zeghidour N., Tagliasacchi M. LMcodec: a Low Bitrate Speech Codec With Causal Transformer Models. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095442.
20. Shi H., Shimada K., Hirano M., Shibuya T., Koyama Y., Zhong Z., Takahashi S., Kawahara T., Mitsufuji Y. Diffusion-Based Speech Enhancement with Joint Generative and Predictive Decoders. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2024. pp. 12951–12955. DOI: 10.1109/ICASSP48485.2024.10448429.

21. Kong J., Kim J., Bae J. HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. Proceedings of the 34th Conference on Neural Information Processing Systems (NIPS). 2020. vol. 33. pp. 17022–17033.
22. Kaneko T, Tanaka K., Kameoka H., Seki S. Istftnet: Fast and Lightweight Mel-spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform. 2022. arxiv preprint arXiv:2203.02395v1.
23. Subramani K., Valin J.-M., Isik U., Smaragdis P., Krishnaswamy A. End-to-end LPCNet: A Neural Vocoder With Fully-Differentiable LPC Estimation. Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). 2022. pp. 818–822.
24. Valin J.-M., Skoglund J. LPCNet: Improving Neural Speech Synthesis Through Linear Prediction. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. pp. 5891–5895. DOI: 10.1109/ICASSP.2019.8682804.
25. Valin J.-M., Isik U., Smaragdis P., Krishnaswamy A. Neural Speech Synthesis on a Shoestring: Improving the Efficiency of LPCNet. Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. pp. 8437–8441.
26. Valin J.-M., Buthe J., Mustafa A. Low-Bitrate Redundancy Coding of Speech Using a Rate-distortion-optimized Variational Autoencoder. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10096528.
27. Zeghidour N., Luebs A., Omran A., Skoglund J., Tagliasacchi M. SoundStream: An End-to-End Neural Audio Codec. Proceedings of the IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021. vol. 30. pp. 495–507.
28. Du Z., Zhang S., Hu K., Zheng S. FunCodec: A Fundamental, Reproducible and Integrable Open-Source Toolkit for Neural Speech Codec. 2023. arxiv preprint arXiv:2309.07405v1.
29. Defossez A., Copet J., Synnaeve G., Adi Y. Funcodec: High Fidelity Neural Audio Compression. 2022. arxiv preprint arXiv:2210.13438v1.
30. Demo site of the High Fidelity Neural Audio Compression (EnCodec). Available at: https://ai.honu.io/papers/encodec/samples.html. (accessed 18.10.2024).
31. Yin D., Luo C., Xiong Z., Zeng W. PHASEN: A Phase-and-Harmonics-Aware Speech Enhancement Network. 2019. arxiv preprint arXiv:1911.04697v1.
32. Rogozinsky G.G. [Perceptual sound compression using wavelet packets]. Dissertation of St. Petersburg State University of Culture and Technology. 2010. (In Russ.).
33. Zwicker E., Fastl H. Psychoacoustics: Facts and Models. Springer-Verlag, Berlin Heidelberg. 1990.
34. Official website of ITU. General methods for the subjective assessment of sound quality. Recommendation ITU-R BS. 1284-2 (01/2019). Available at: https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1284-2-201901-I!!PDF-E.pdf. (accessed 05.01.2025).

Chizhov Ilya — Ph.D., Principal engineer, team lead of AI speech/audio coding team, Huawei Russian Research Institute. Research interests: audio compression, digital signal processing, ML/DL, AI-based speech enhancement, personalized noise reduction, text-to-speech. The number of publications — 22. aproximation18@yandex.ru; 17/2, Krylatskaya St., 121614, Moscow, Russia; office phone: +7(495)234-0686.