Обзор технологий передачи аудио потока по низкоскоростным и нестабильным каналам связи

А. Ю. Дорогов
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
Vaksa2006@yandex.ru

А. Л. Лившиц
ПАО «Интелтех»
alexeylivshits@gmail.com

Ю. В. Савенкова
ПАО «Интелтех»
poli-teh@list.ru

Аннотация. **В статье выполнено исследование интеллектуальных средств кодирования и восстановления звуковой информации с улучшенными показателями качества в условиях использования нестабильных низкоскоростных радиоканалов. Представлены существующие технологии, используемые для кодирования и декодирования аудиоинформации, оценивается их применимость к низкоскоростным и нестабильным каналам связи. Рассматриваются технологии восстановления и повышения качества аудиоинформации при возникновении потерь при передаче.**

Ключевые слова: аудиоинформация, аудиокодеки, низкоскоростные и нестабильные каналы связи, нейронные сети, телекоммуникационные сервисы, восстановление аудиоинформации

I. Введение

В современном мире передача аудиоинформации становится все более востребованным способом коммуникации. Видео чаты приобрели популярность в связи с переходом большого количества людей на удаленную работу. При этом и в видео чате основным источником, переносящим информацию, является аудио-сигнал.

Для кодирования аудио-информации в цифровых сетях используются устройства или программы, называемые кодеками. Кодек, или, другими словами, кодировщик, – это программное либо аппаратное средство для кодирования и декодирования информации по определенному алгоритму (в данном случае – аудио-информации).

Кодирование, или сжатие аудиосигналов, может быть двух видов: с потерями информации и без потерь. Для каждого вида кодирования существуют свои виды аудио-кодеков.

При прохождении сигнала через кодек, канал связи и декодер, кодеки без потери качества обеспечивают на выходе декодера сигнал, полностью идентичный сигналу на входе кодера.

Кодеки с потерями на выходе декодера выдают сигнал, отличный от оригинального, но достаточный для восприятия человеком. В таких кодеках при кодировании аудио-сигнала звук модифицируется, из него могут вырезаться неслышимые человеческому уху частоты, фоновые шумы и другие компоненты сигнала, слабо влиявшие на восприятие аудио-информации. Более того, некоторые кодеки при декодировании добавляют особые шумы, для более приятного восприятия аудио-сигнала человеком. Кодеки с потерями позволяют существенно снизить объем передаваемых данных, поэтому это направление активно развивается в последнее время.

При пакетной передаче аудиоинформации из-за большой нагрузки на сетевую инфраструктуру часть пакетов может задерживаться в очередях маршрутизаторах или повреждаться в результате искажений в линиях. Поврежденные пакеты отправляются повторно или передаются другим маршрутом, что увеличивает их время прохождения по сети. Все это влечет за собой проблему прихода пакетов на оконечное устройство в последовательности, отличной от начальной. При перегрузке сети часто возникает ситуация «неактуальности пакета, когда пакет приходит на оконечное устройство уже после того момента, когда он должен был быть использован. Проблему иллюстрирует схема прохождения пакетов через сеть, показанная на рис. 1.

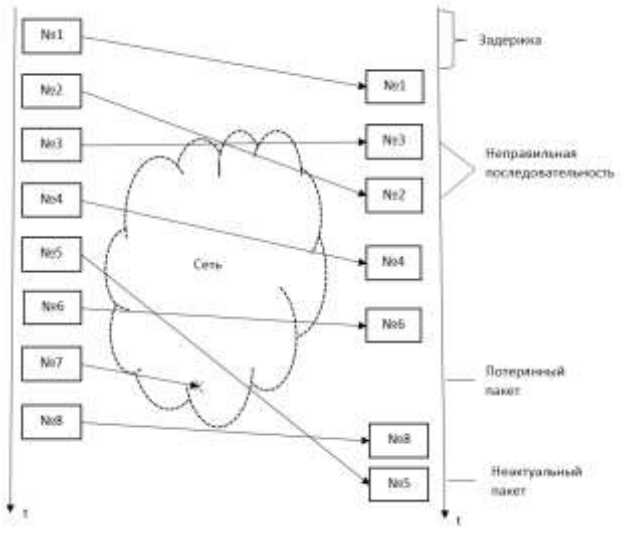

Рис. 1. Потери пакетов по сети

На данной схеме показано, что пакет №3 пришел на оконечное устройство с задержкой после пакета №4, пакет №7 был отброшен одним из маршрутизаторов по пути следования, а пакет №5 пришел с такой большой

Работа выполнена в рамках гранта РНФ (заявка №22-21-20094)

задержкой, что он стал неактуальным и оконечному устройству пришлось его отбросить.

Из вышесказанного можно сделать вывод, что существует два вида потерь аудиоинформации:

- потери при кодировании;
- потери при передаче.

Первый вид связан с намеренным использованием кодеков с потерями, при этом качество восстановленного сигнала ухудшается, но при этом уменьшается объем передаваемых данных. Потери при передаче возникают в сети ненамеренно и носят случайный характер. В этом случае на стороне декодера нужно применять технологии по восстановлению потерянных частей аудиоданных.

В нормальном режиме передаче потери обычно составляют не более 3 %. Современные декодеры пытаются восстановить недостающие пакеты основываясь на различных принципах. Некоторые кодеки заведомо передают излишнюю информацию о «соседних» пакетах, чтобы декодер мог восстановить потерянный пакет. В некоторых декодерах при декодировании недостающие пакеты заменяются на «средний» сигнал предыдущего и последующего декодированного пакета. Существуют и более сложные способы «компенсации» потерянных пакетов, но все они справляются только с небольшими потерями.

При передаче по низкоскоростным и нестабильным каналам связи данные кодеки практически не применимы, т. к. в таких каналах скорость передачи крайне мала, что влечёт за собой невозможность избыточного кодирования сигнала. Нестабильность каналов связи влечет за собой потерю пакетов на уровне 10–30 %, что приводит к необходимости использования специальных технологий восстановления аудиоинформации. В настоящей работе приведен краткий обзор наиболее перспективных технологий, применяемых для кодирования речевой информации при большом уровне потерь.

II. Кодеки для нестабильных каналах связи

Рассмотрим некоторые варианты кодеков, которые применялись и применяются сейчас для решения задач передачи аудио-информации по низкоскоростным и нестабильным каналам связи

A. Opus

Аудио-кодек Opus разработанный сообществом Internet Engineering Task Force (IETF) основан на принципе сжатия с потерями, т. е. выходные данные (аудио сигнал) отличаются от исходных. Протокол Opus является открытым форматом, стандартизованным в RFC 6716 [1].

Аудио-кодек поддерживают частоты дискретизации от 8 до 48 кГц и битрейт (количество бит, используемых для передачи/обработки данных в единицу времени) от 6 до 510 кбит/с. Основным преимуществом данного кодека является минимальная задержка кодирования, которая составляет от 2.5 до 60 мс. Кодек является отказоустойчивым и при потере одного или нескольких блоков возможно восстановление сигнала. Также одной из ключевых характеристик кодека является его способность при декодировании использовать частоту дискретизации, отличную от частоты дискретизации, использовавшейся при кодировании.

Кодек содержит два программных модуля – для компрессии аудио-данных высокого качества (музыка) и для компрессии голоса.

Для компрессии аудио-данных используется модуль CELT. Его алгоритм схож с принципом работы наиболее популярных кодеков с потерями, использующих принцип «оптимизации» звука. Оптимизация заключается в том, что из сигнала удаляются составляющие, которые не существенны для слухового восприятия человеком аудио-сигнала. Разложение сигнала на составляющие реализуется с помощью дискретного косинусного преобразования.

Для кодирования голоса используется модуль SILK. При использовании данного модуля аудио-кодек анализирует входной аудио-сигнал на предмет наличия человеческой речи. После обнаружения речи голосовые составляющие отделяются от прочих звуков. Далее кодек выполняет анализ частотной характеристики звука, понижая уровень дискретизации для данных, содержащих голосовую информацию, то есть речь. Затем анализирует присутствующие шумы и оптимизирует сигнал для определенного битрейта.

Используя речевые кадры, модуль предсказания частоты аудио-сигнала вносит изменения в последующие кадры. Далее следует важный этап обработки звука — устранение искажений, возникающих при недостаточно высоком битрейте. После этого используется модуль фильтрации шума квантования, который снижает шумы внутри рабочей полосы сигнала.

Кодек поддерживает передачу как моно так и стереосигнала, используя технологию постоянного и переменного битрейта, а также поддерживает компрессию до 255 каналов, причем количество каналов можно менять «на лету», без пере инициализации. Opus широко используется в популярных приложениях для голосовых чатов, а также для воспроизведения любого видео на Youtube.

B. Enhanced Voice Services

В области подвижной телефонной связи наиболее часто используется кодек Enhanced Voice Services. Кодек был разработан совместно производителями чипсетов и производителями мобильных телефонов [2]. Данный кодек использует раздельные режимы сжатия для звука и голоса в зависимости от содержимого аудио-сигнала. Кодек обеспечивает полосу пропускания звука до 20кГц, имеет механизм улучшенного маскирования потери пакетов, а также обладает высокой устойчивостью к задержке джиттера. Дополнительной возможностью данного кодека является реализация технологии исправления ошибок по каналу. При передаче в каждый пакет добавляются биты, несущие информацию о предыдущих кадрах, что дает возможность частичного восстановления пакета при его повреждении или в случаях неактуальности пакета по времени. За счёт этого кодек EVS обеспечивает восстановление до 15 % потерянных пакетов при незначительном снижении качества восприятия. EVS кодек применяется в телефонии, телеконференциях и в потоковом аудио. К отличительным особенностям кодека можно отнести:

- переменный битрейт. Битрейт можно переключать каждые 20 мс на стороне источника;

- детектор голоса;
- переключение способов сжатия речи и аудио-сигнала;
- частичное восстановление пакетов при потере;
- маскирование ошибок.

C. Lyra

Аудио-кодек Lyra разработан компанией Google. Этот аудио-кодек ориентирован на достижение максимального качества при передаче речи по низкоскоростным каналам связи.[3] Lyra представляет собой один из первых кодеков, которые используют машинное обучение для восстановления речевого сигнала.

Кодек включает в себя кодировщик и декодировщик. Алгоритм работы кодировщика сводится к извлечению параметров голосовых данных каждые 40 миллисекунд, их сжатию и передаче получателю по сети. Для передачи данных достаточно иметь канал связи со скоростью 3 килобита в секунду. Извлекаемые звуковые параметры включают в себя логарифмические мел-спектрограммы, учитывающие характеристики энергии речи в различных частотных диапазонах с учётом модели человеческого слухового восприятия.

В декодировщике используется генеративная модель, которая на основе переданных звуковых параметров воссоздаёт речевой сигнал. Для снижения сложности вычислений применена модель на основе рекуррентной нейронной сети, представляющей собой вариант модели синтеза речи WaveRNN [4], в котором используется более низкая частота выборок, но генерируется параллельно сразу несколько сигналов в разном диапазоне частот. Полученные сигналы затем суммируются для получения выходного сигнала, соответствующего заданной частоте дискретизации.

Lyra использует обученную нейросеть. Данная нейросеть способна воссоздавать речевой аудио-сигнал, используя звуковые параметры, полученные на входе. Модель генерации речи Lyra обучалась на тысячах часов звуковых данных, взятых из различных открытых аудио-библиотек более чем 70 мировых языков. Исследования показали, что аудио-кодек Lyra сжимает и передает речь на битрейте 3 Кбит/с с таким же уровнем качества, как это делает кодек Opus на 8 Кбит/с. По заявлениям разработчиков для ускорения работы кодека применены специализированные процессорные инструкции, доступные в 64-разрядных процессорах ARM. В итоге, несмотря на применение затратного по времени машинного обучения, кодек Lyra может использоваться для кодирования и декодирования речи в реальном режиме времени на смартфонах среднего ценового диапазона, демонстрируя задержку передачи сигнала на уровне 90 миллисекунд.

D. Codec2

Codec 2 является речевым кодеком с открытым исходным кодом [5]. На данный момент он является единственным свободно распространяемым кодеком, способным работать на скоростях ниже 5000 бит/с.

Codec 2 использует гармоническое кодирование речи. Он разделяет речь на сегменты по 10−30 мс, которые называются кадрами. Каждый кадр затем анализируется на предмет фундаментального уровня (pitch) и количества гармоник, которые вписываются в полосу пропускания 4 кГц. Далее для каждой гармоники в диапазоне 4 кГц передаются значения амплитуды и фазы. При таком кодировании качество речи на выходе получается достаточно хорошим, непереданные части речи (в частности согласные) хорошо моделируются набором гармоник со случайными фазами. При уменьшении полосы пропускания падает качество восприятия речи, поэтому при использовании кодека приходится искать «золотую середину» между полосой пропускания и комфортным восприятием речи.

III. Технологии восстановления аудио-данных

Для передачи речи по низкоскоростным каналам связи со скоростью ниже 3кбит/с подходят только несколько кодеков, среди которых MELPe, Lyra, Codec2.

Поскольку практически достигнут предел сжатия речи при кодировании, единственным решением улучшения качества становится восстановление речи на выходе декодера, при этом неважно по каким причинам произошло искажение – потеря пакетов, низкая скорость канала или заведомо плохое качество компрессии аудио-сигнала. Для восстановления аудио потока в рассмотренных технологиях используется два подхода.

A. PLC

PLC (Packet Loss Concealment) – технология, широко используемая в аудио-кодеках для восстановления потерянных пакетов. Суть технологии заключается в хранении нескольких отсчетов звука до текущего момента времени, а также задержки звукового сигнала на выходе, и соответственно хранение нескольких отсчетов после текущего момента времени. При потере текущего отсчета он заменяется на отсчет, сгенерированный на основе предыдущих и последующих отсчетов (рис. 2).

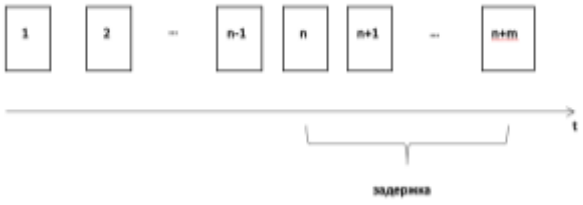

Рис. 2. Буфер хранения истории отсчетов

В стандартном режиме буфер хранит 48,75 мс (390 образцов) до текущего и 30 образцов после текущего момента времени. 30 отсчетов вносят задержку выходного сигнала на 3.75 мс. Общий алгоритм работы показан на рис. 3.

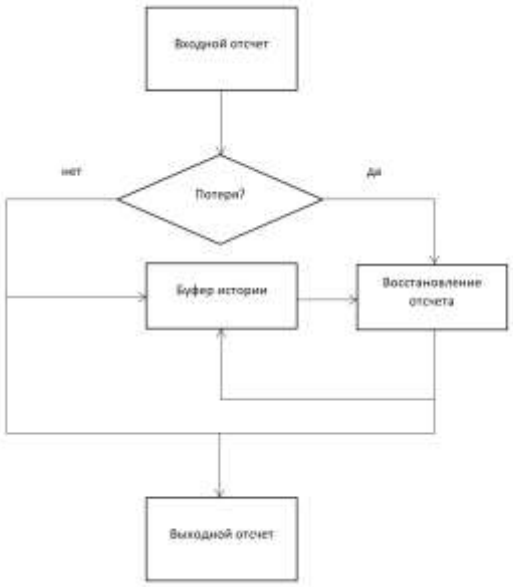

Рис. 3. Алгоритм генерации потерянных отсчетов

На вход поступает очередной аудио-отсчет. Если отсчет не пустой, т. е. пакет не был потерян, то он добавляется в буфер истории, а из буфера удаляется один «самый старый» отсчет. В этом случае тот же отсчет подается на выход. Если же отсчет был пустой, т. е. пакет потерян, то генерируется новый отсчет на основе буфера истории, и он добавляется в буфер истории с удалением «самого старого» и подается на выход.

Данный метод успешно справляется с потерями единичных отсчетов, разнесенных по времени. К минусам можно отнести заметное ухудшение качества звука при потерях нескольких отсчетов подряд. При таких потерях звук становится менее разборчивым и возможно кратковременное полное пропадание аудио информации.

B. Neural Network Loss Concealment

При восстановлении сигнала с использованием нейросетей используется следующий алгоритм. При получении каждого отсчета звука выделяются его основные характеристики. Далее значения каждого отсчета передаются на вход нейросети. При отсутствии одного или нескольких отсчетов в сигнале, основываясь на знаниях, полученных при обучении и на характеристиках предыдущих и последующих отсчетов, нейросеть способна предсказать характеристики отсутствующего отсчета, и на их основе может быть сгенерирован новый отсчет.

В качестве примера нейронных сетей можно рассмотреть технологию WaveNet, которая была применена для восстановления сигнала, сжатого при помощи Codec2. [6]. Для улучшения речи при декодировании была применена генеративную модель глубокого обучения. WaveNet использует нейросеть, которая способна обучаться в процессе ее использования. Общий алгоритм восстановления аудиосигнала показан на рис. 4. [7]

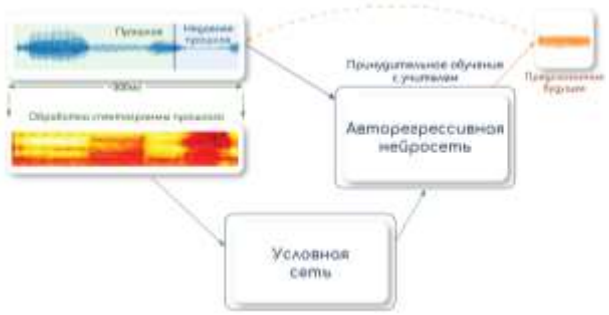

Рис. 4. Алгоритм восстановления WaveNet

Каждые 300мс на вход условной сети подаётся спектрограмма предыдущего аудио-сигнала. На выходе эта сеть выдает информацию, описывающую признаки данного сигнала. Эта информация поступает на вход авторегрессионной нейросети, в которой происходит предсказание следующего отсчета. Особенностью этого подхода является то, что в авторегрессионную сеть на входе подается ее же выходной сигнал, что позволяет при потере отсчетов не только продолжить воспроизведение речи, но и сделать переход от синтетической речи к реальной более плавным и при этом избежать заметного шума.

Технология WaveNet показала, что при скорости канала менее 2400 бит/с на выходе можно достичь качества речевого сигнала, приемлемого для восприятия и выделения необходимой информации.

IV. Заключение

В данной статье рассмотрены причины возникновения потерь при передаче аудиосигналов через цифровые сети. Рассмотрены различные способы кодирования и декодирования аудиоинформации, а также способы ее восстановления при потерях.

В данной статье рассмотрено два кодека, дающих возможность использования нейросетей для улучшения качества выходного аудио-сигнала – это кодеки Lyra и Codec2. По сравнению с Lyra у Codec2 есть большое преимущество – он является кодеком с открытым кодом, при этом показатели качества речи остаются на том же уровне, что и Lyra без использования механизма восстановления звука.

Были рассмотрены два основных способа восстановления аудиоинформации, а именно технология PLC и технология WaveNet, которая использует нейросети и машинное обучение.

Нейросети показали хорошее качество восстановленной речи при потерях 15–30 %, что особенно актуально на низкоскоростных и нестабильных сетях связи. Большим преимуществом использования нейросетевых технологий также является возможность восстанавливать целые блоки (несколько пакетов расположенных рядом) потерянных речевых данных.

Нейросетевые технологии показали заметное увеличение качества выходного сигнала и возможность уменьшения полосы пропускания при поддержании аудио сигнала надлежащего качества.

Из проведённого обзора следует вывод, что при передаче аудиоинформации по нестабильным низкоскоростным каналам нейросетевые методы восстановления сигнала являются в настоящее время наиболее перспективными. Кодеки, на основе других методов восстановления сигнала применимы для нестабильных каналов связи, но на низкоскоростных каналах показывают свою неэффективность.

Список литературы

[1] Internet Engineering Task Force (IETF). Definition of the Opus Audio Codec. Mozilla Corporation. September 2012

[2] Nokia 3GPP Enhanced Voice Services (EVS) codec 2012

[3] W. Bastiaan Kleijn, Andrew Storus, Michael Chinen, Tom Denton, Felicia S. C. Lim, Alejandro Luebs, Jan Skoglund, Hengchin Yeh “Generative speech coding with predictive variance regularization” ICASSP 2021 – 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 18 February 2021

[4] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron van den Oord, Sander Dieleman, Koray Kavukcuoglu “Efficient Neural Audio Synthesis”, 25Jun 2018

[5] David Rowe, South Australia “Codec 2 – Open source at 2400 bit/s below”

[6] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu “Wavenet: a generative model for raw audio” Computer Science”, 2 September 2016

[7] Florian Stimberg, Alex Narest, Alessio Bazzica, Lennart Kolmodin, Pablo Barrera Gonzalez,Olga Sharonova, Henrik Lundin, and Thomas C. Walters “WaveNetEQ – Packet Loss Concealment with WaveRNN”, 54th Asilomar Conference on Signals, Systems and Computers, November 2020