Как мы в VK Звонках работаем над качеством звука

Алексей Шпагин

ВКонтакте

VideoTech 2022

Алексей Шпагин

- Руководитель команды бэкенда ВК Звонки
- 10 лет работы в VoIP телефонии и видеозвонках
- Бэкграунд - разработчик C++
- В руководстве командами 4 года

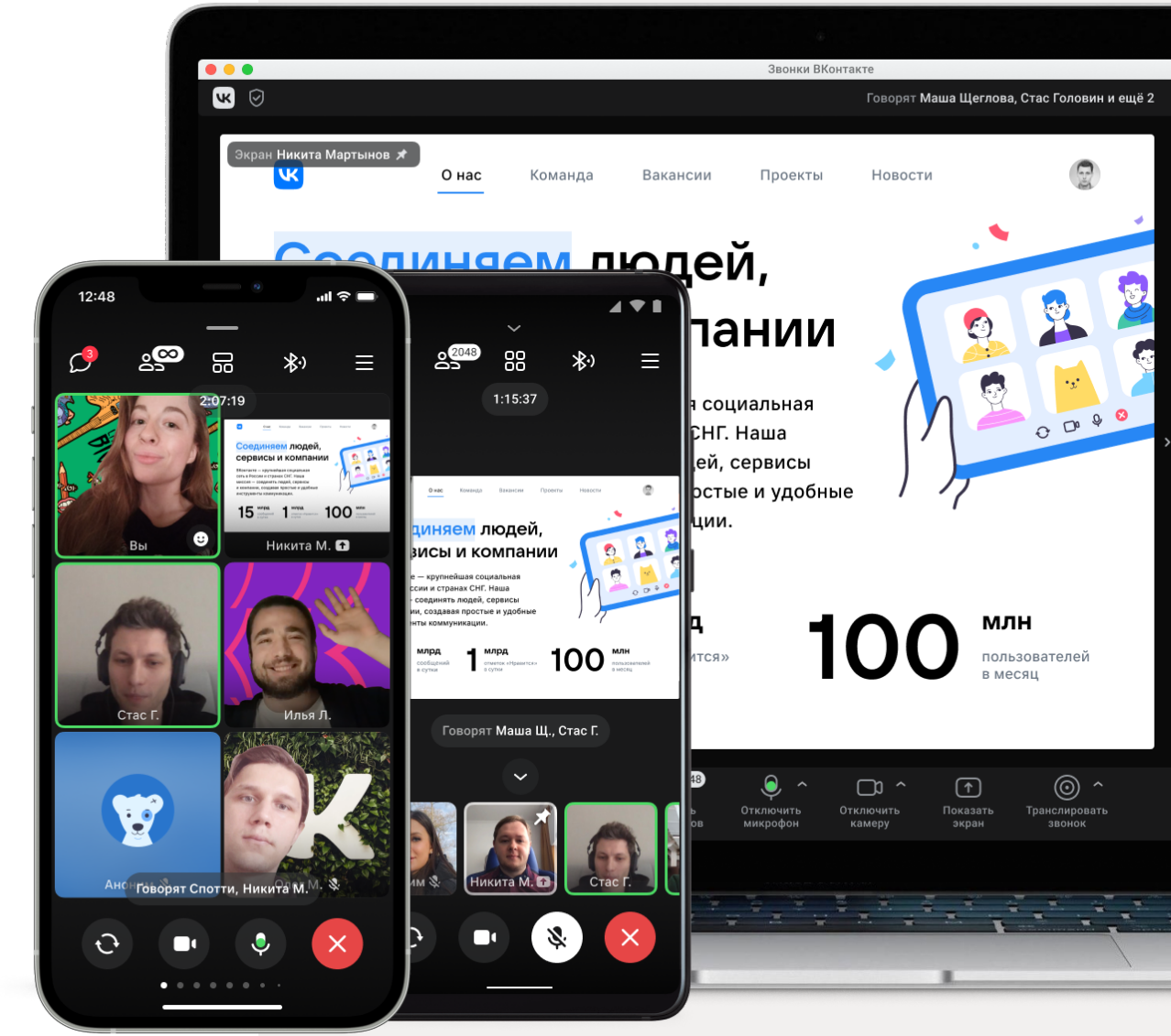
VK Звонки

Бесплатные звонки без ограничений по времени и количеству участников.

Для работы и учёбы
Демонстрация экрана в 4K, трансляция, планирование и запись.

Управление звонками
Зал ожидания, управление микрофонами, функция «Поднять руку» и другие возможности модерации.

Технологичность
Интеллектуальное шумоподавление, собственная AR-технология замены фона.



20 млн

пользователей общаются в VK Звонках ежемесячно

Источник: ВКонтакте, апрель 2022 г.

VK Звонки

6 млн

звонков в день

20 млн

пользователей в месяц

15 тыс.

одновременных звонков

Содержание

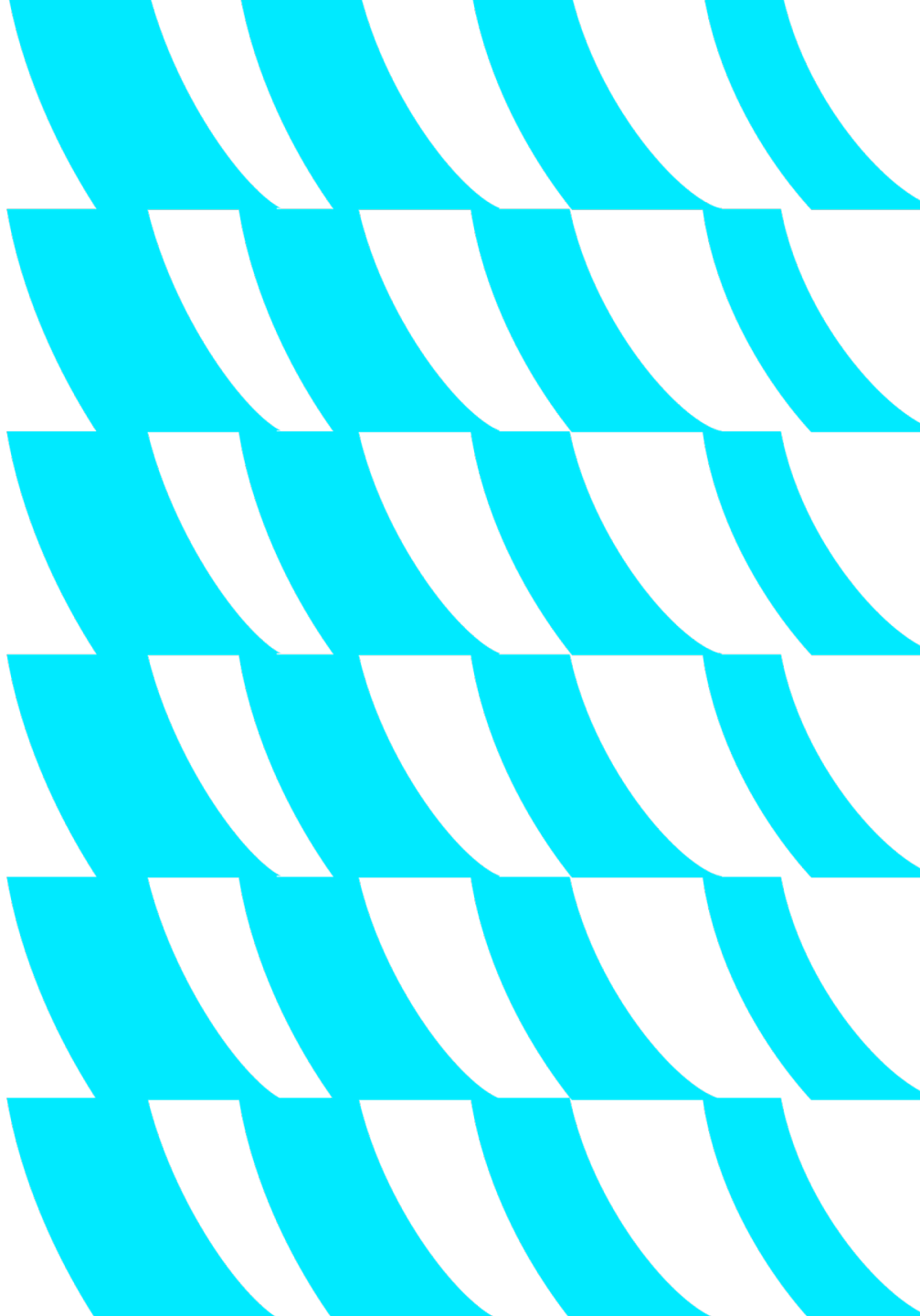
Из чего складывается качество звука?

Видеозвонки — замена живым встречам

- При встрече оффлайн все отлично друг друга видят и слышат
- Система видеозвонков вносит искажения в передачу звука
- Если искажения сильные, общаться не комфортно и встреча перестает быть похожа на встречу живьем

Требования к передаче звука в системе видеозвонков

В порядке убывания критичности



Непрерывность звукового потока

Минимальная latency

input

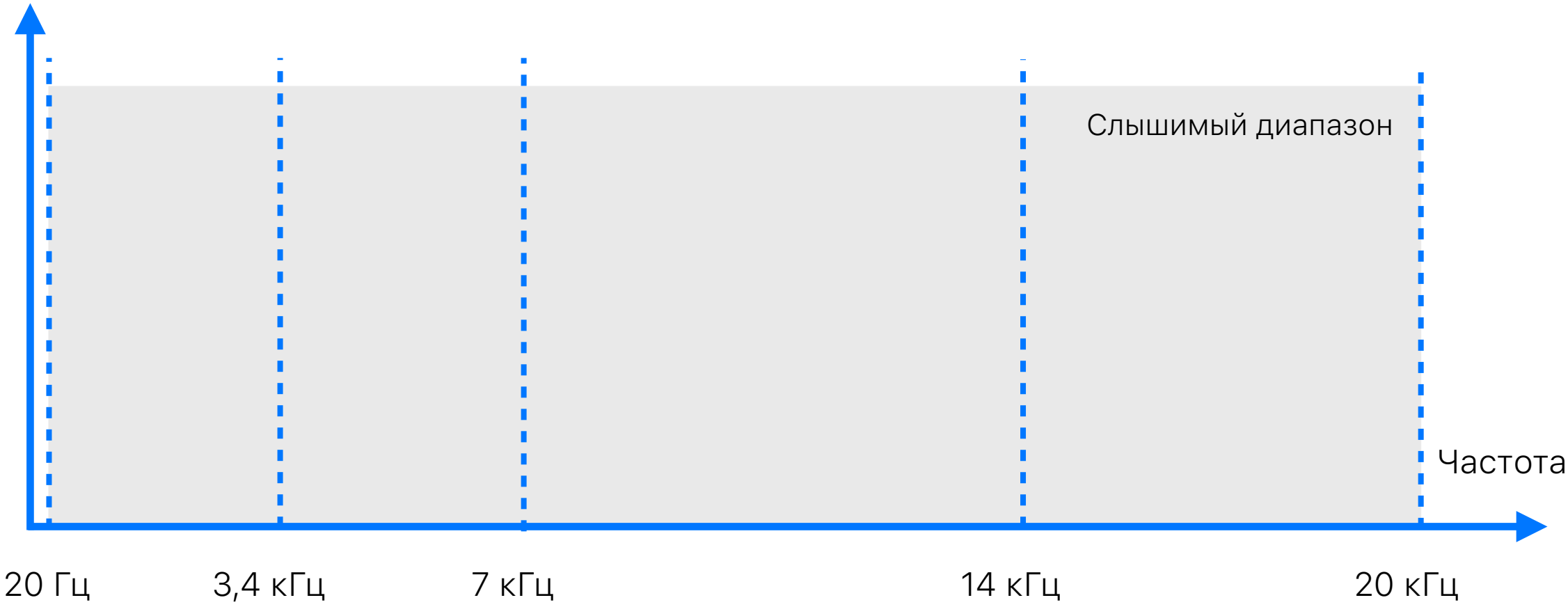
output

latency

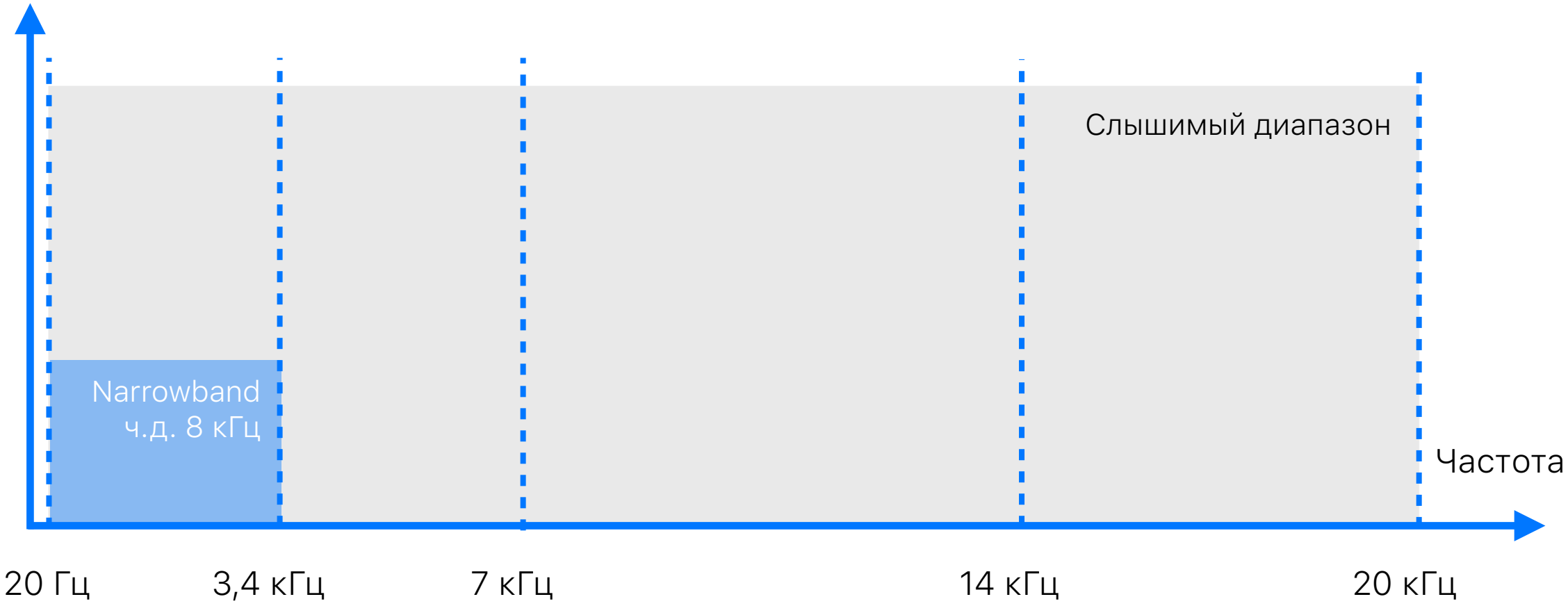
Отсутствие артефактов в звуке

- Постоянный треск
- Периодические щелчки в произвольное время
- Клиппинг

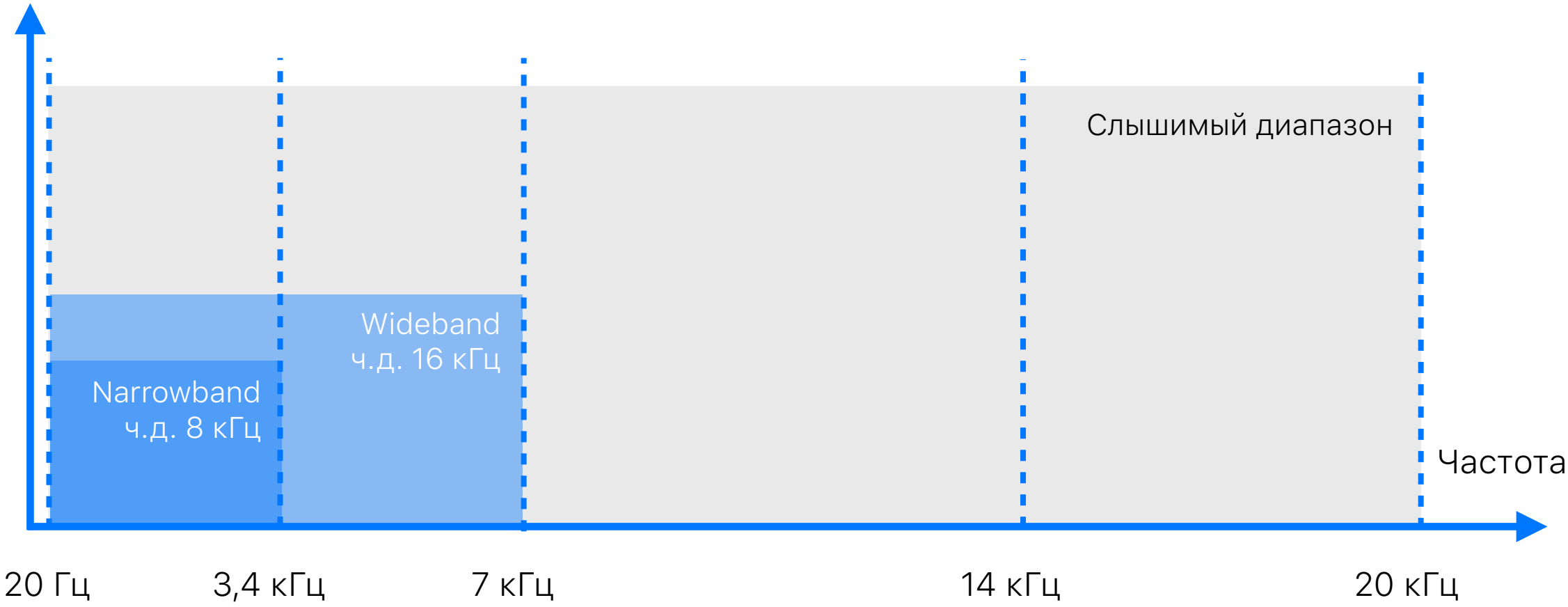
Достаточный частотный диапазон



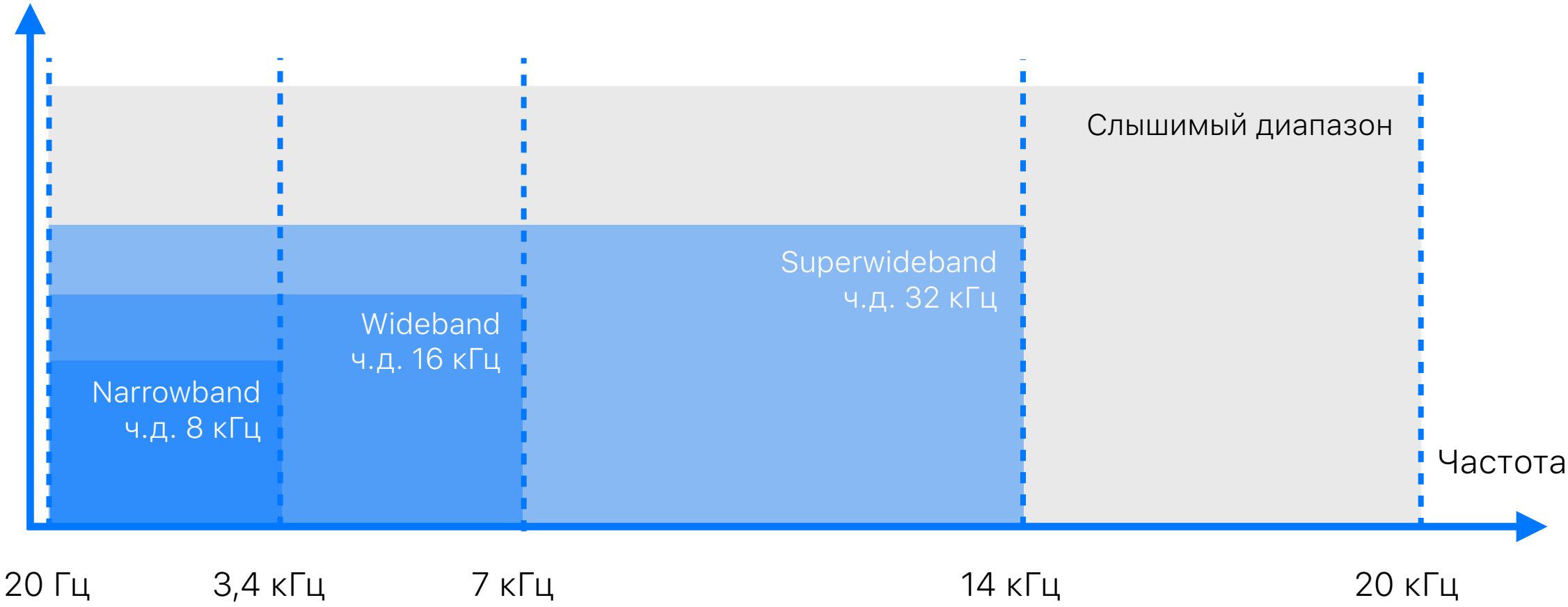
Достаточный частотный диапазон



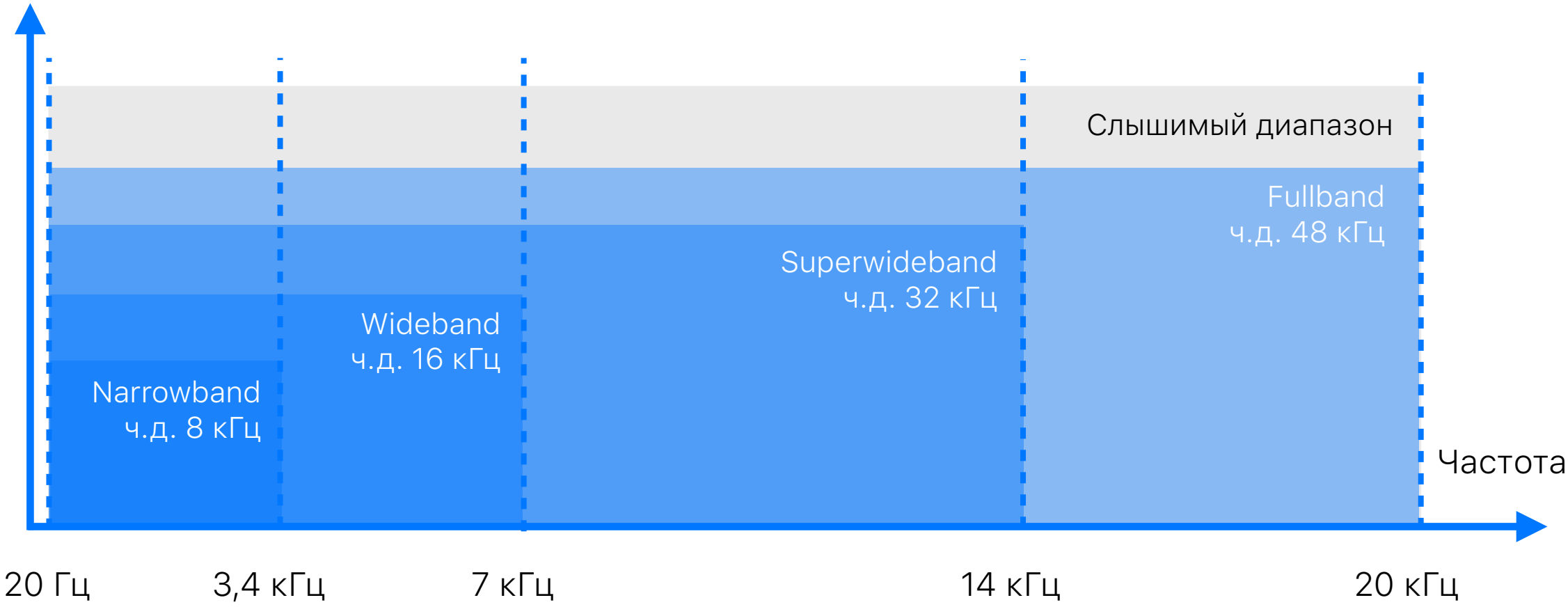
Достаточный частотный диапазон



Достаточный частотный диапазон



Достаточный частотный диапазон



Достаточный уровень громкости

Требования к передаче звука для комфортного общения

1. Непрерывность звукового потока
2. Минимальная latency
3. Отсутствие артефактов в звуке
4. Достаточный частотный диапазон
5. Достаточный уровень громкости

Измерение качества звука

Что же будем оценивать?

Не звук абстрактно
и не качество
звукозаписи — а речь:

- Степень искажений
- Разборчивость
- Комфортность восприятия

MOS - Mean Opinion Score

MOS	Качество	Усилия при прослушивании
5	Отличное	Нет
4	Хорошее	Не значительные
3	Среднее	Средние
2	Низкое	Значительные
1	Неприемлемое	Сверх возможного

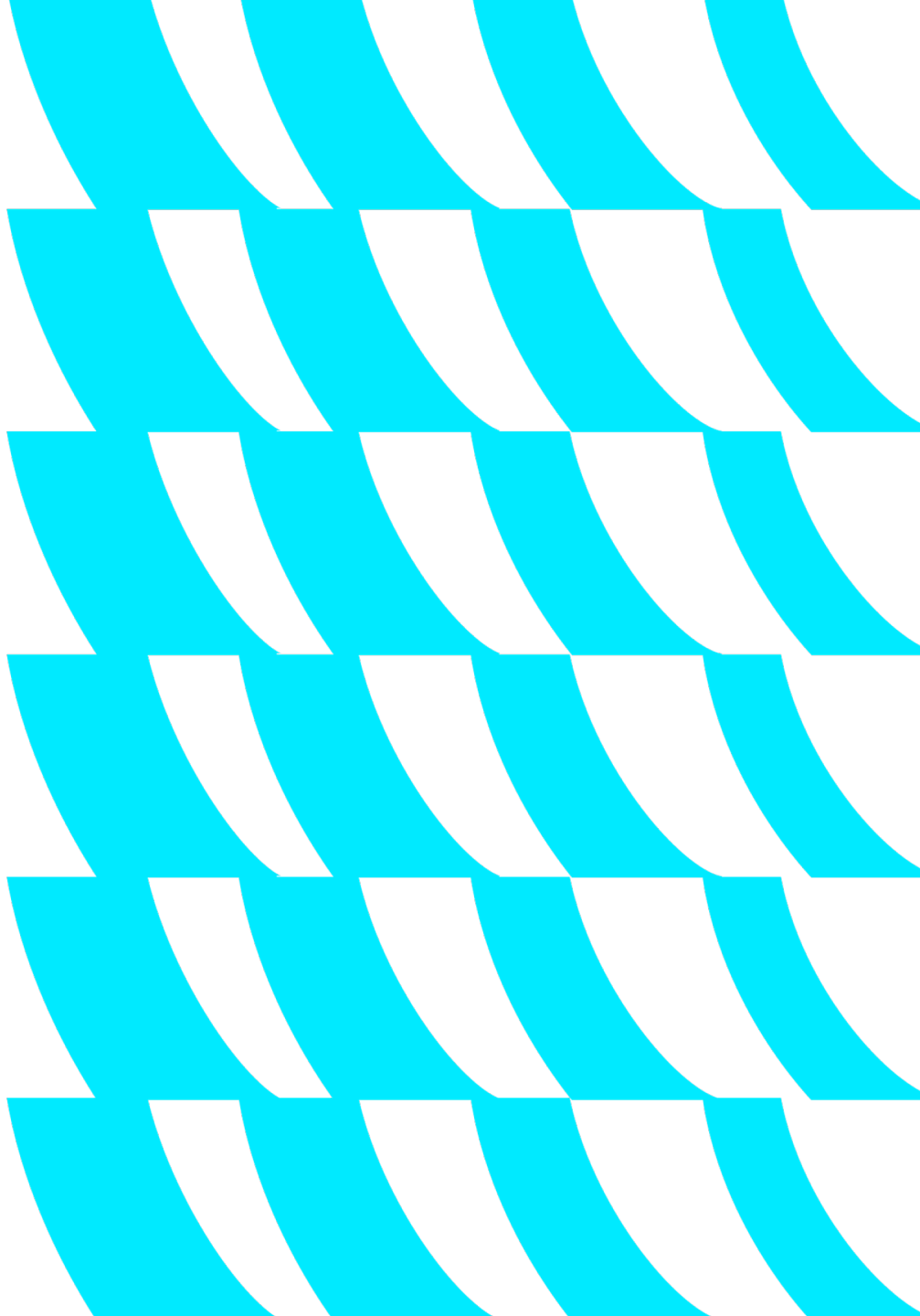
MOS равный 5 — не достижим

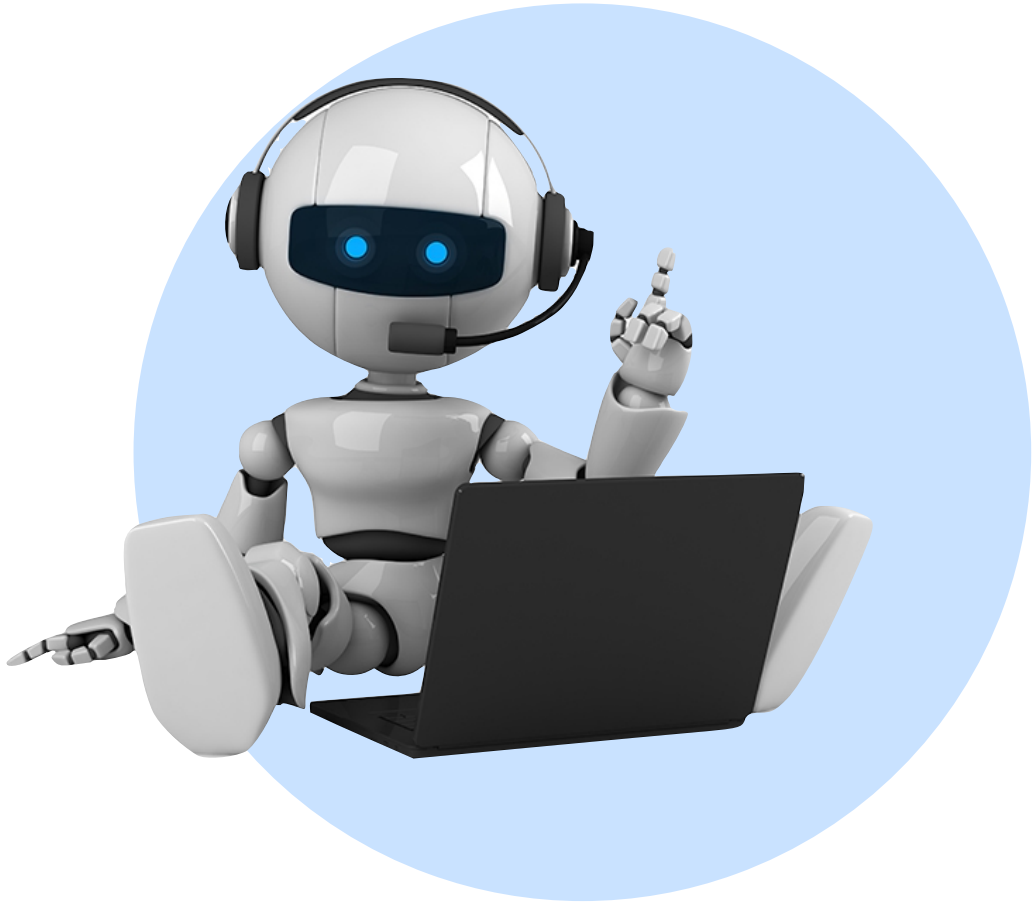
- Звук в системах видеосвязи жмется кодеком
- Кодек жмет с потерей данных
- Как следствие MOS зависит от кодека
- И не бывает максимальным

Кодек	Max. MOS
G.711	4,4
G.722	4,5
G.729	3,92
OPUS*	4,5

* У OPUS несколько режимов работы, MOS приведен для 48 кГц, 32 kbit/s

Подходы к измерению качества речи



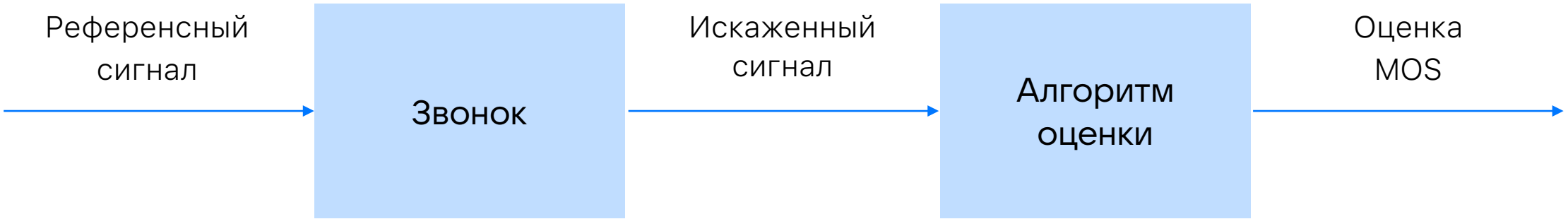


Ориентируемся на восприятие речи человеком

- Алгоритмы измерения качества речи предсказывают как бы живой человек оценил заданный звуковой фрагмент
- Оценка MOS показывает то, как человек воспринимает фрагмент речи

Методика оценки качества речи на основе референса

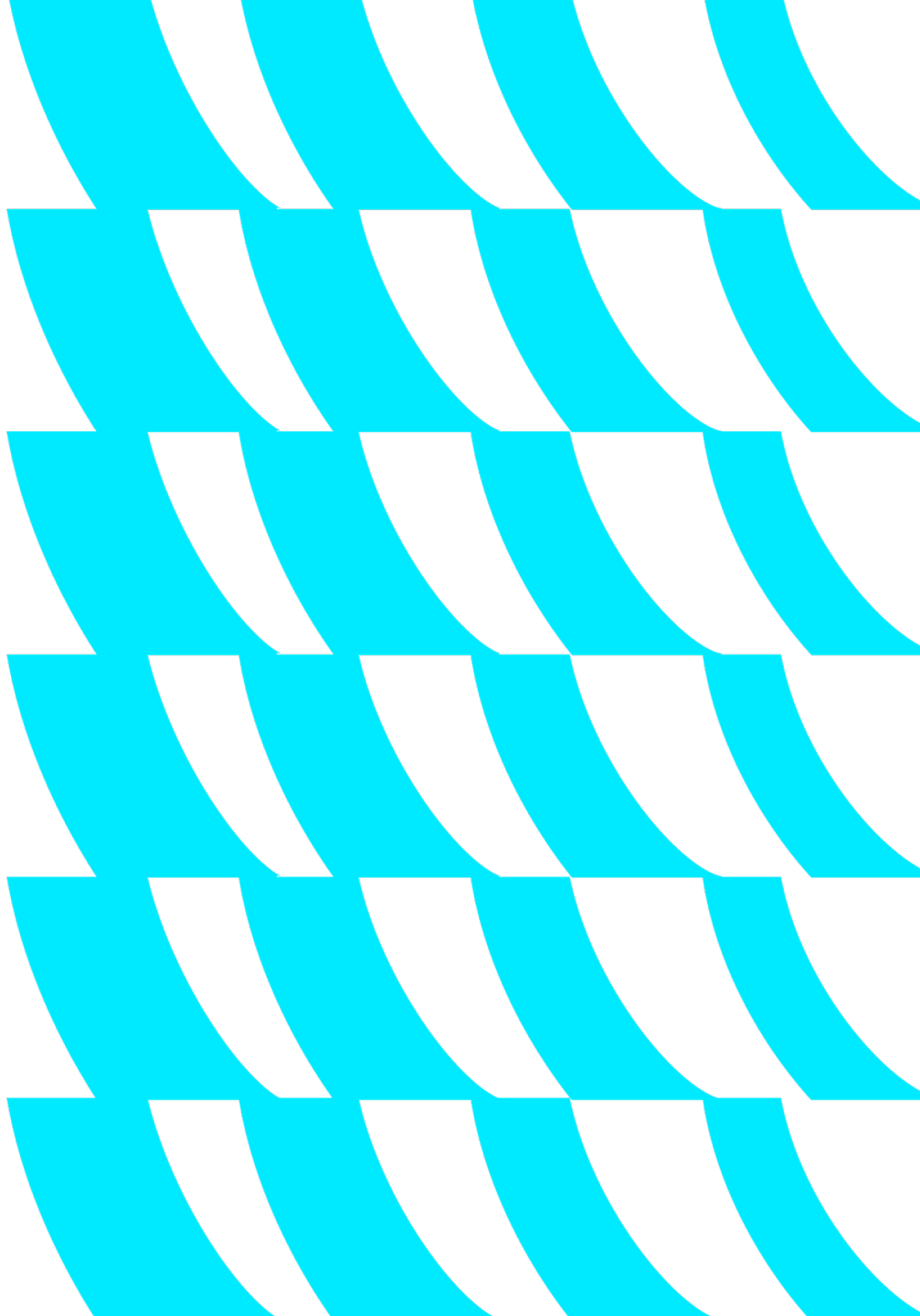
Безреференсная методика оценки качества речи



Итого про оценку качества речи

1. Измеряем качество голоса, речи, а не звука вообще
2. Используем метрику MOS со значениями от 1 до 5.
3. MOS не бывает равен 5
4. Методики оценки пытаются предсказать реакцию человека на звуковой фрагмент
5. Бывают подходы основанные на референсе, а бывают безреференсные

Методики и алгоритмы измерения качества речи



PESQ — Perceptual Evaluation of Speech Quality

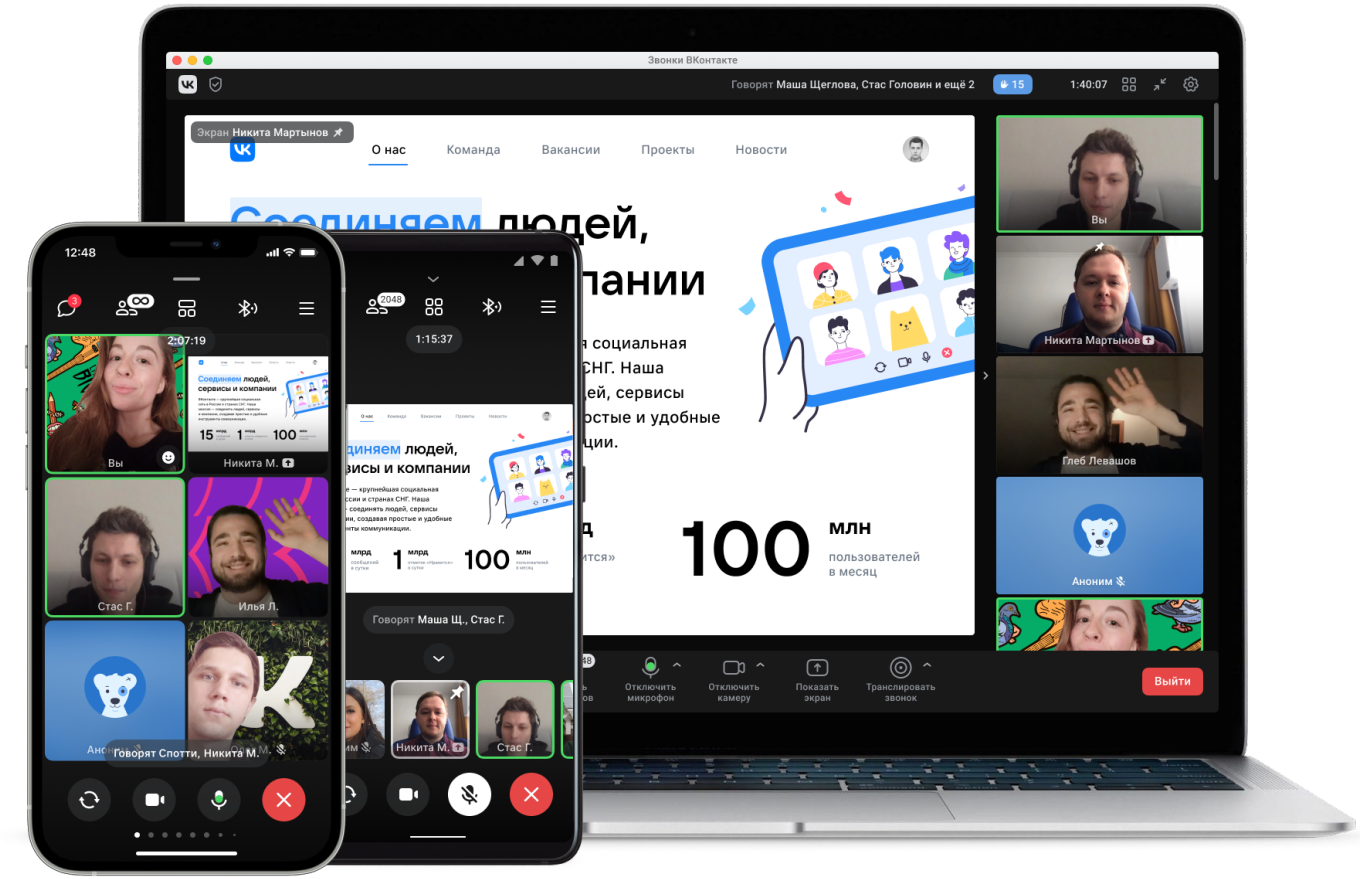
- VoIP
- 2001 год
- 8 кГц и 16 кГц
- PESQ score от -0.5 до 4.5, а не MOS
- Коммерческое решение

POLQA — Perceptual Objective Listening Quality Analysis

- Первая редакция 2011 г.
- Третья, современная редакция 2018 г.
- Частоты дискретизации до 48 кГц
- MOS
- Коммерческое решение

ViSQOL - Virtual Speech Quality Objective Listener

- Open source тул от Google
- Частота дискретизации до 48 кГц для музыки и 16 кГц для голоса
- MOS



AQuA — Audio Quality Analyzer alternative for POLQA and ViSQOL

- Коммерческое решение
- В сравнении с POLQA показала себя хуже на наших тестах
- MOS

NISQA — Speech Quality and Naturalness Assessment

- Безреференсная система оценки качества голоса
- Open source
- MOS

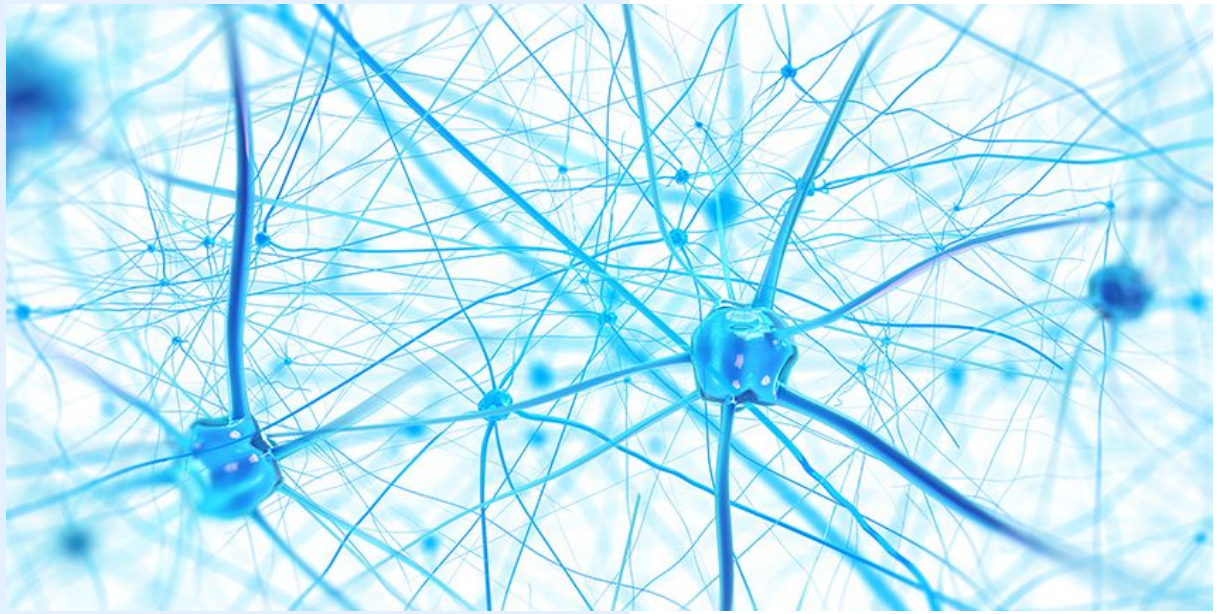
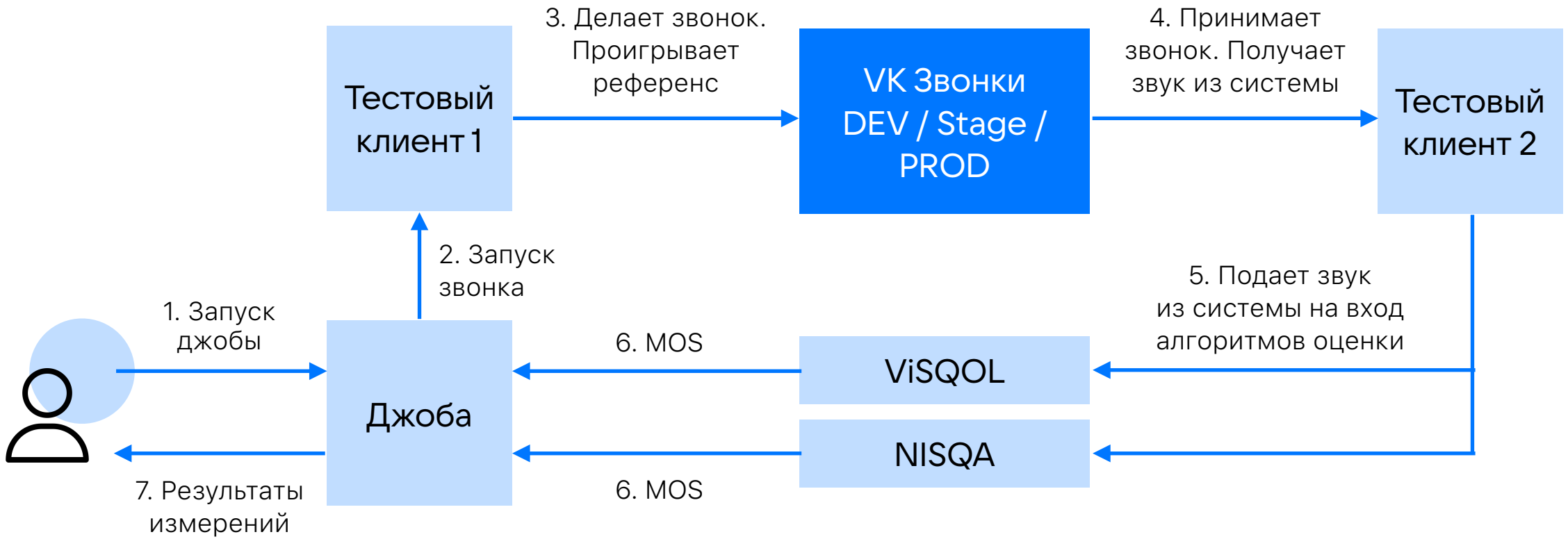
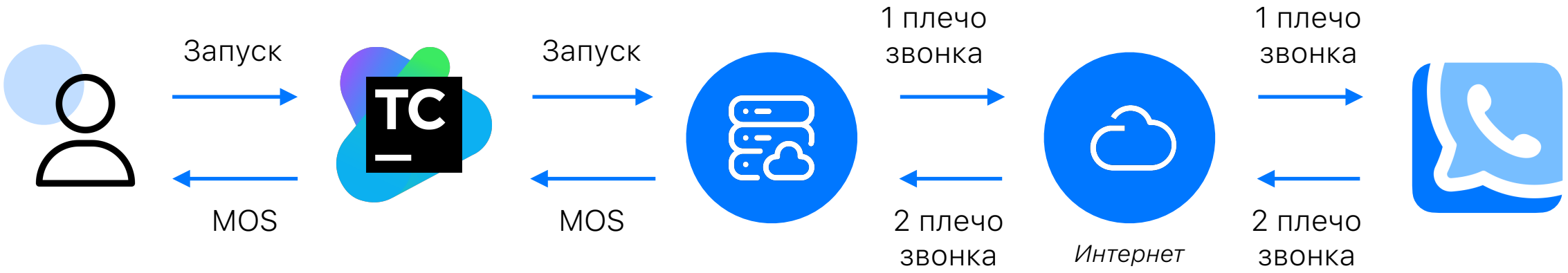


Схема оценки качества голоса в VK Звонках



Стенд для измерения качества голоса в VK Звонках



Итого про методики оценки качества речи

Мы пользуемся в VK Звонках

ViSQOL
Open source аналог POLQA

NISQA
Безреференсная методика.
Open source

Мы не пользуемся

POLQA
Коммерческое решение

AQuA
Перспективное решение, но пока проигрывает POLQA

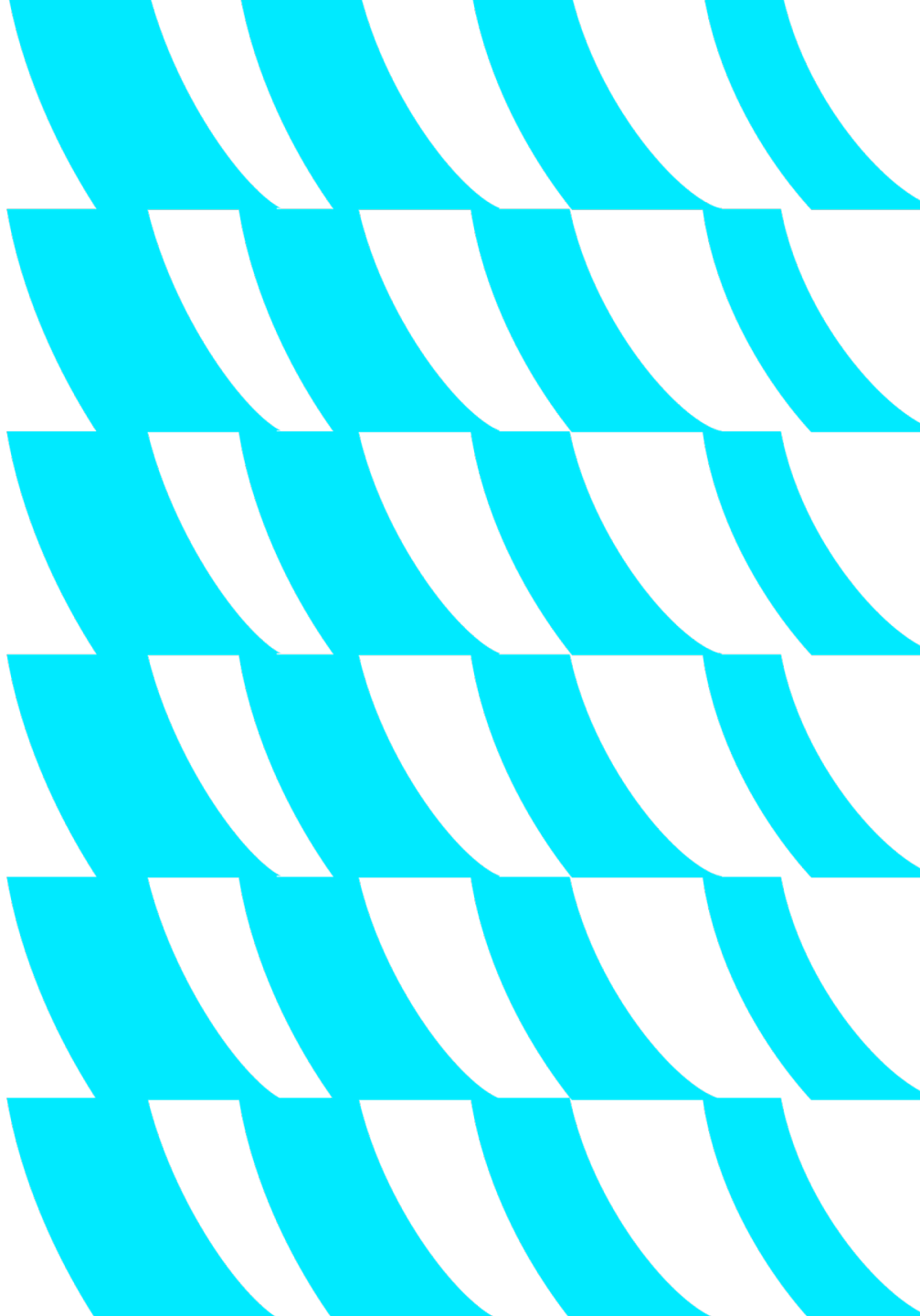
PESQ
Устаревшее коммерческое решение

Примеры проблем при передаче звука, которые мы решали

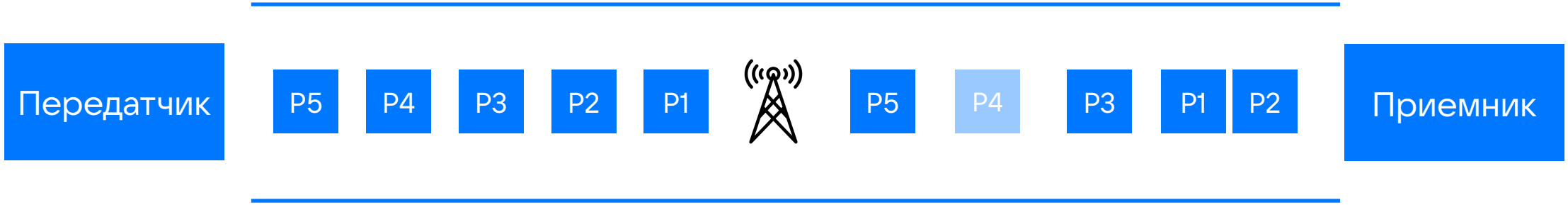
Виды проблем при передаче голоса через Интернет

- Проблемы обусловленные особенностью передачи данных по TCP/IP сетям
- Проблемы акустического характера

Проблемы обусловленные особенностью передачи данных по сети



Принцип передачи голоса по сети



- Пакеты могут теряться (packet loss)
- Пакеты могут задерживаться на небольшое переменное время, «дрожать» (Jitter)
- Пакеты могут задерживаться на константное время (delay)
- Пакеты могут меняться местами (reordering)

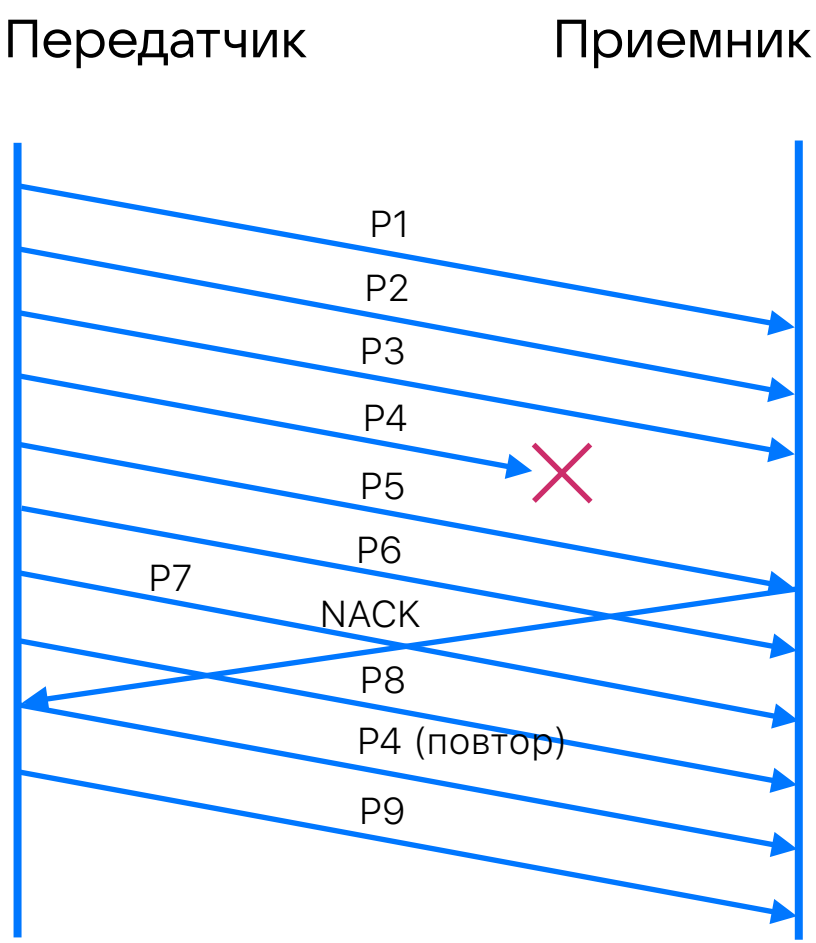
Jitter Buffer

- Компенсирует «дрожание»
- Выравнивает трафик
- Может «подождать» недошедший вовремя пакет
- Больше Jitter Buffer - больше latency

PLC — Packet Loss Concealment

- Восстанавливает фрагмент звука
- Если потеря пакетов большая, то голос становиться металлическим
- Бывают отдельные алгоритмы, бывают встроенные в аудио кодек
- В кодеке Opus - встроенные PLC

NACK — Negative Acknowledgment



- Позволяет перезапросить потерянный пакет
- Хорошо работает на коротких RTT
- При длинных RTT не имеет смысла

FEC — Forward Error Correction

- Кодек добавляет в битстрим дополнительную информацию, которая позволяет восстановить данные
- Хорошо работает на больших RTT, когда перезапрашивать пакет с помощью NACK долго
- Минус - увеличивает битрейт

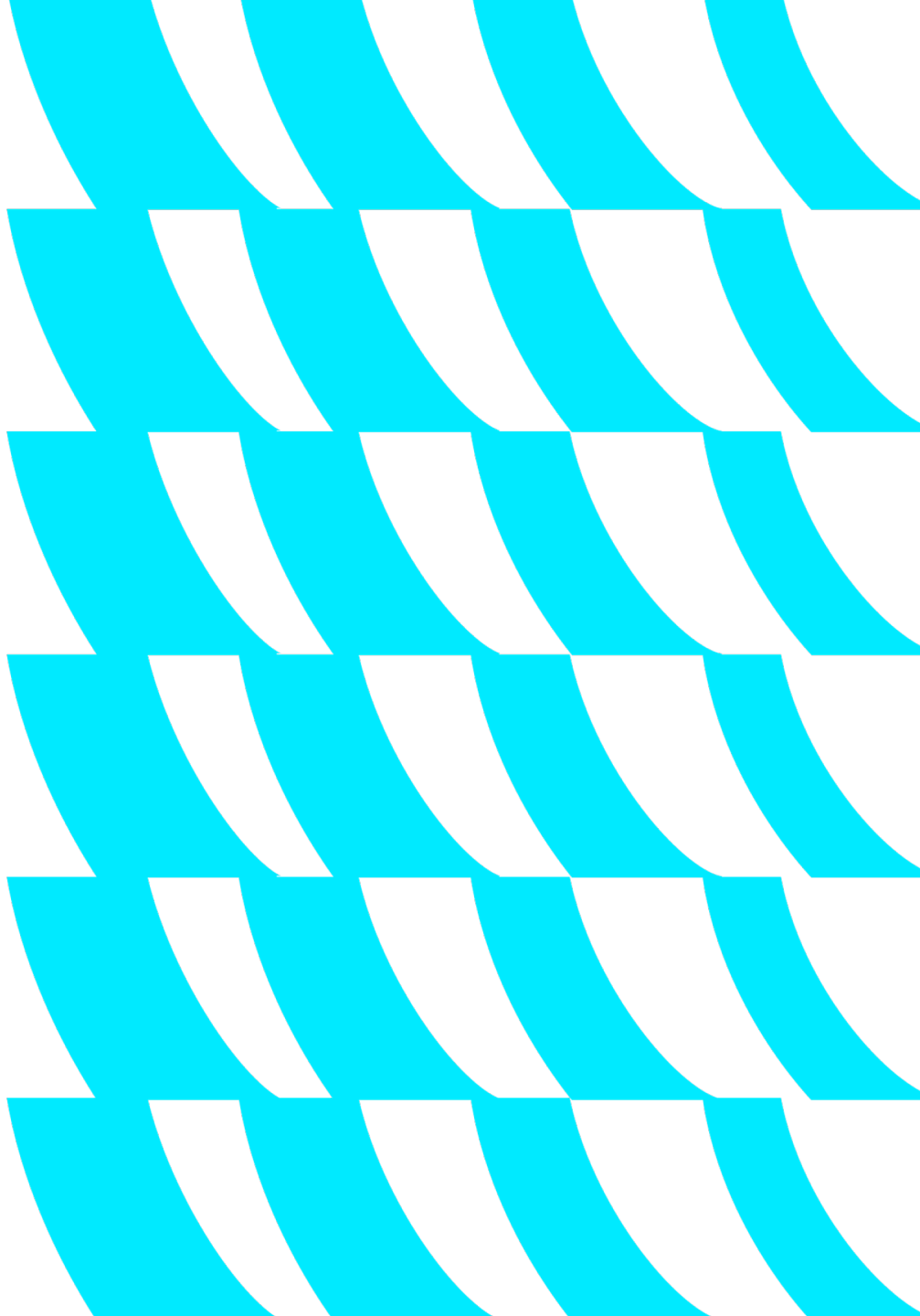
Условия	**Средний MOS**
loss = 5%, FEC = false	3,94
loss = 5%, FEC = **true**	**4,19**

RED — Redundancy (RTP Payload for Redundant Audio Data)

- Избыточность на уровне RTP пакетов
- В один RTP пакет помещается 2 или более аудио-кадров
- Разница с FEC в том, что избыточная информация не в битстриме, а объединяется несколько битстримов

Условия	**Средний MOS**
loss = 10%, delay = 200 ms, RED = false	3,60
loss = 10%, delay = 200 ms, RED = **true**	**4,14**

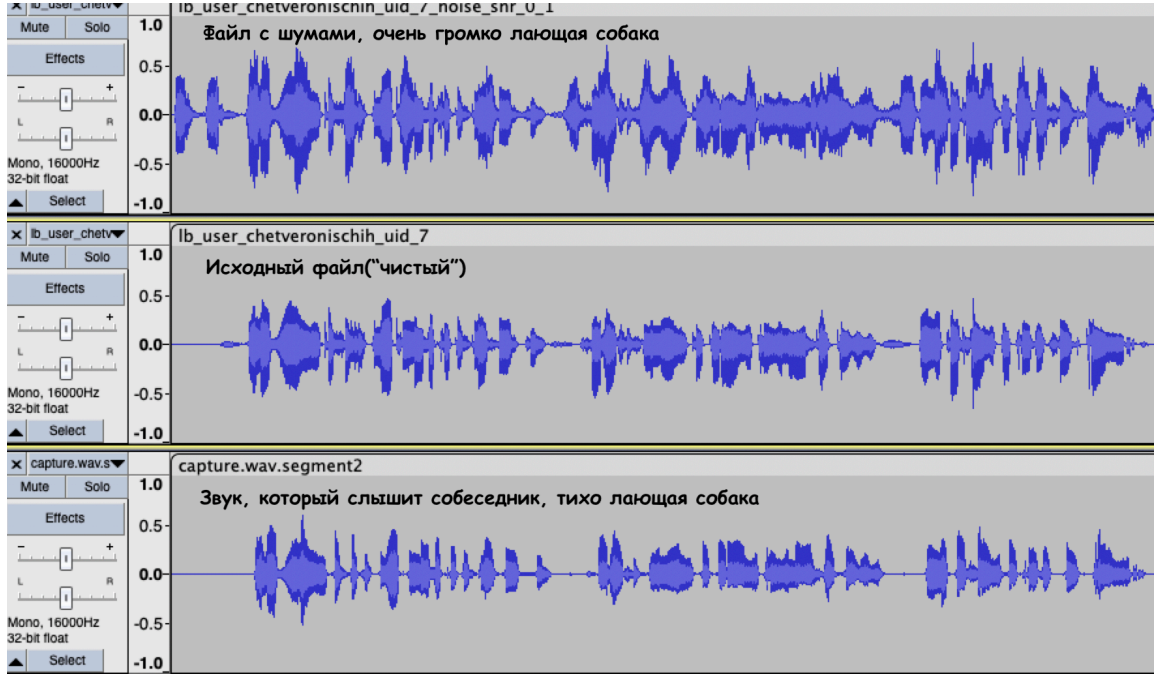
Проблемы акустического характера



Шумоподавление

- Слабое - WebRTC
- Сильное - собственное решение на базе ML

Пример оценки результата шумоподавления



Условия	Средний MOS
Референс без шумов	3,21
Референс + шумы без шумодава	~ 1,00
Референс + шумы + шумоподавление	2,19

VAD — Voice Activity Detection

- Вставляет тишину там, где нет голоса
- Собственное решение на базе ML

Условия	Средний MOS
VAD выключен	3,81
VAD включен	**3,67**

SNR — Signal to Noise Ratio

- Определяет присутствует ли шум в звуке от участника звонка
- Если есть шум, то включаем шумодав, если нет, то не включаем, чтобы сохранить качество голоса
- Учимся оценивать качество работы SNR

Условия	**Средний MOS**
VAD - включен SNR - выключен	3,67
VAD и SNR - оба включены	**3,46**

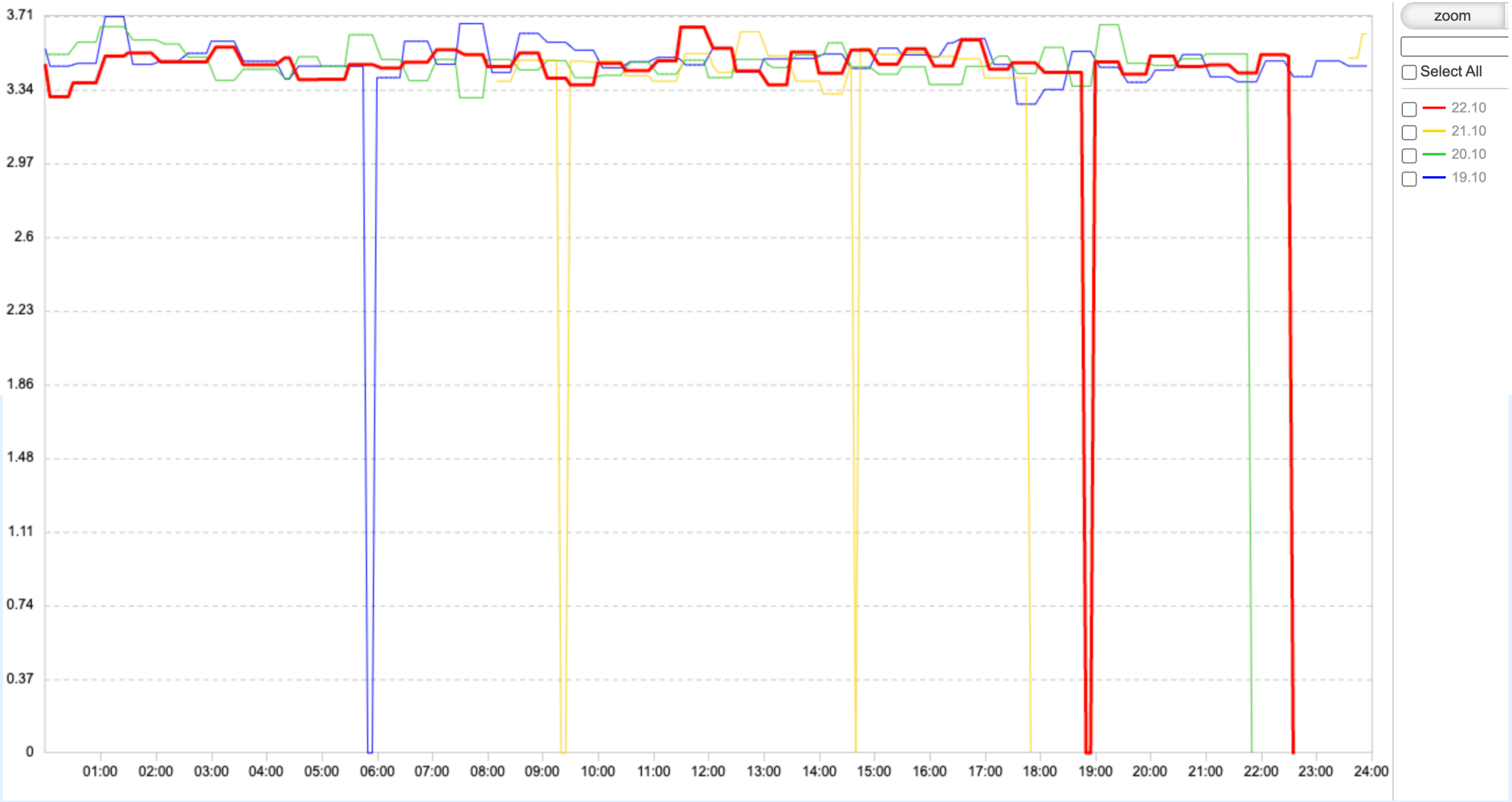
Проблемы, с которыми мы сталкиваемся

1. Шумоподавление
2. VAD
3. SNR - Signal to Noise Ratio
4. Jitter Buffer
5. PLC - Packet Loss Concealment
6. NACK - Negative Acknowledgment
7. FEC - Forward Error Correction
8. RED - Redundancy

Применение инструментов измерения качества голоса

Варианты использования джобы для оценки качества

- Запуск на девелоперском окружении при разработки фичи
- Запуск на регрессионном окружении при выпуске релиза
- Сравнение разных версий продукта
- Выявления эффекта от новой фичи
- Мониторинг продакшн окружения



Мониторинг качества голоса на продакшене

Будем ВКонтакте!