Speech Task Overview 2

Болотов Дмитрий

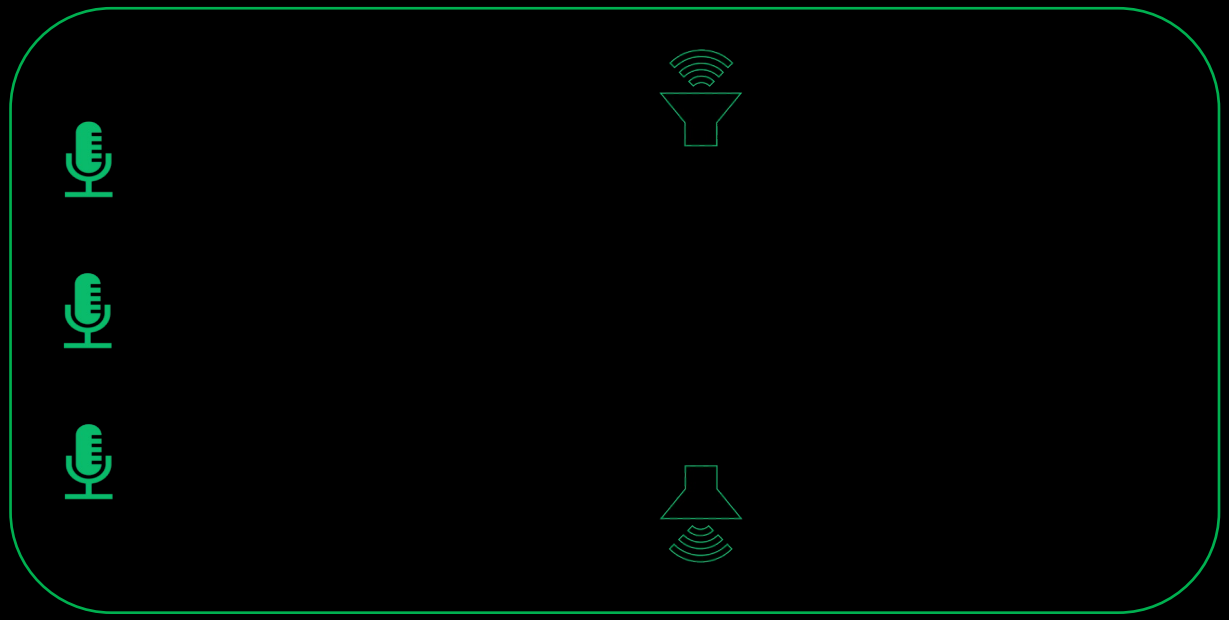
План:

- Voice assistant pipeline
- User perceived latency
  - Keyword detection
  - Endpoint detection


Sber 75″
miniLED
Салют ТВ

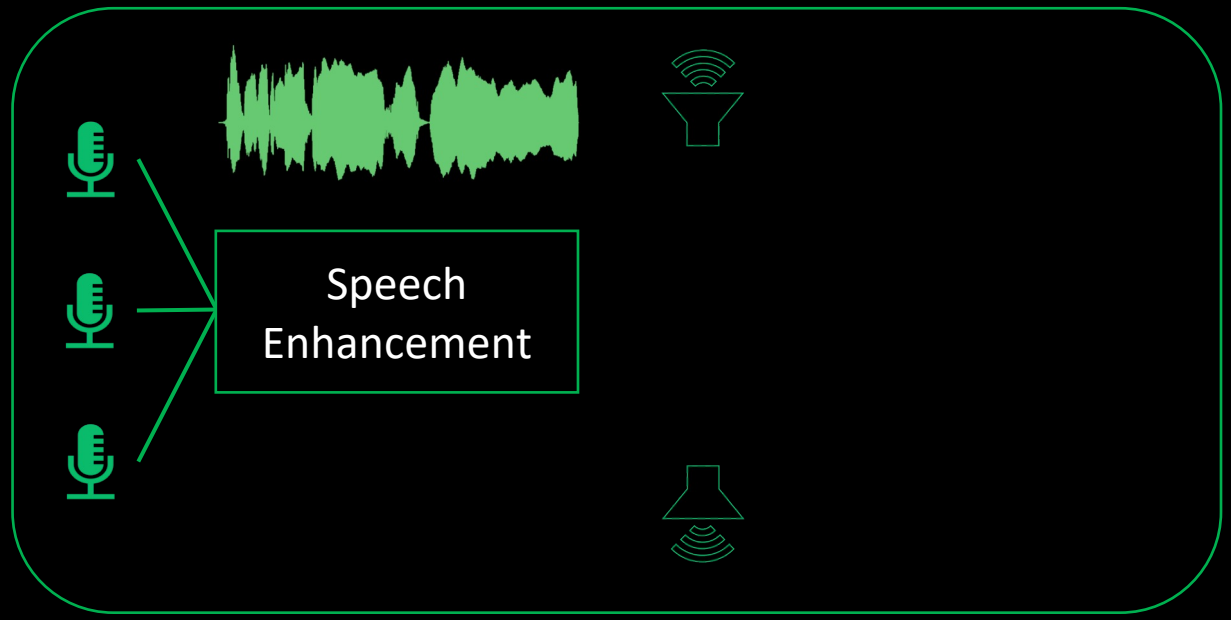

18:41

Voice assistant pipeline

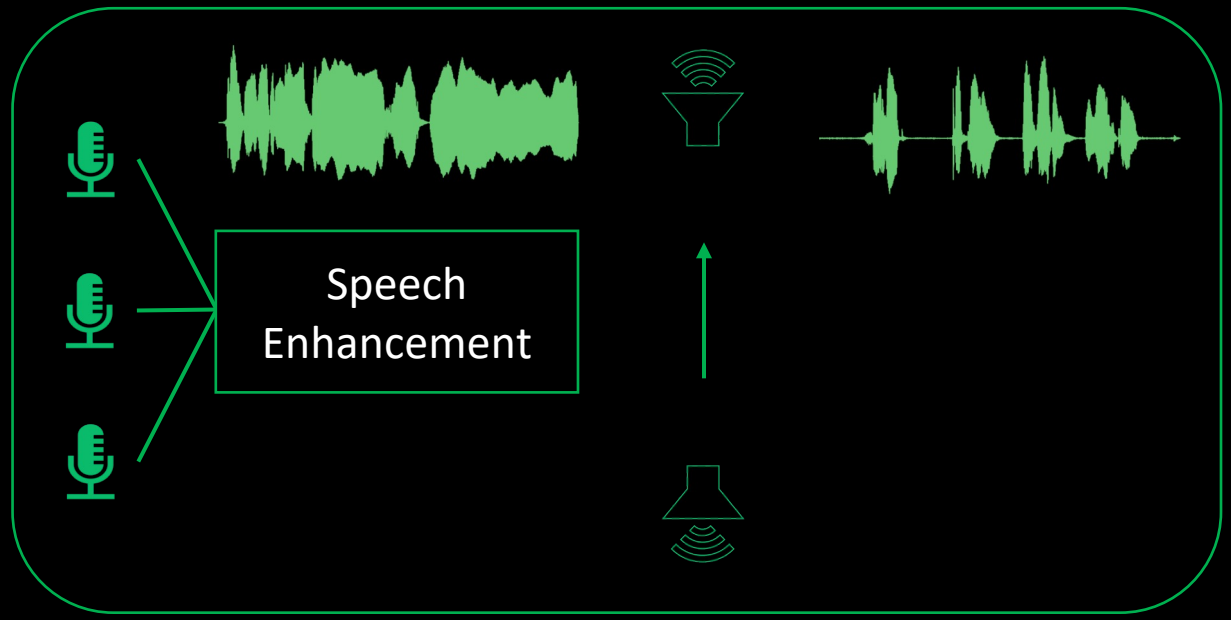


Voice assistant pipeline

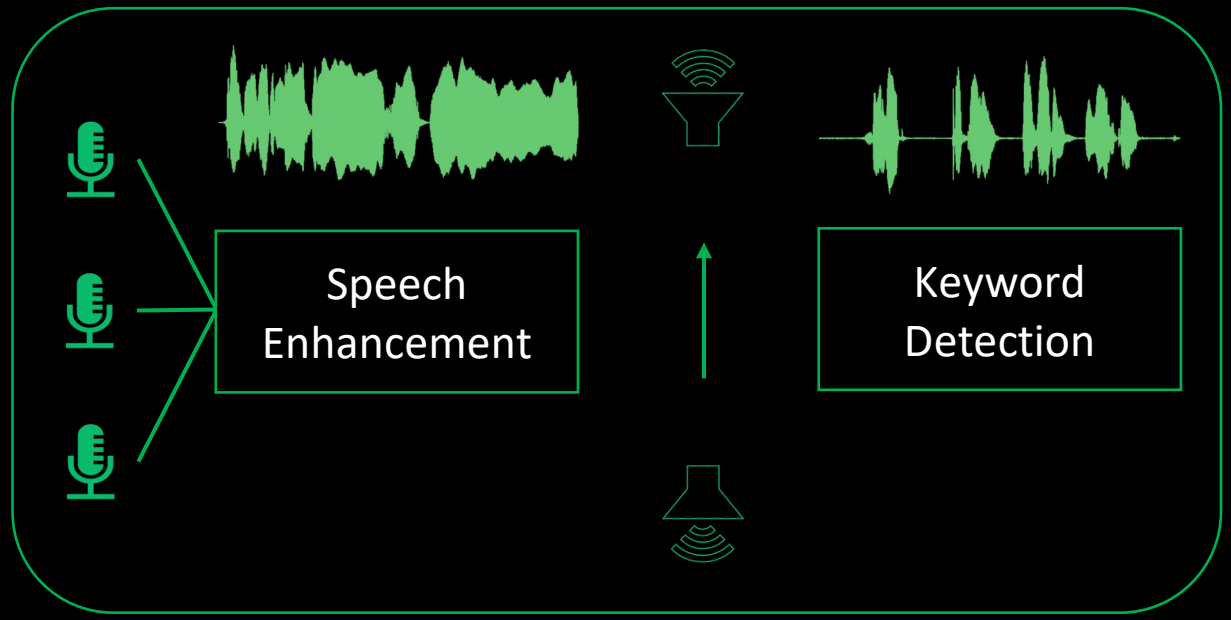
Voice assistant pipeline



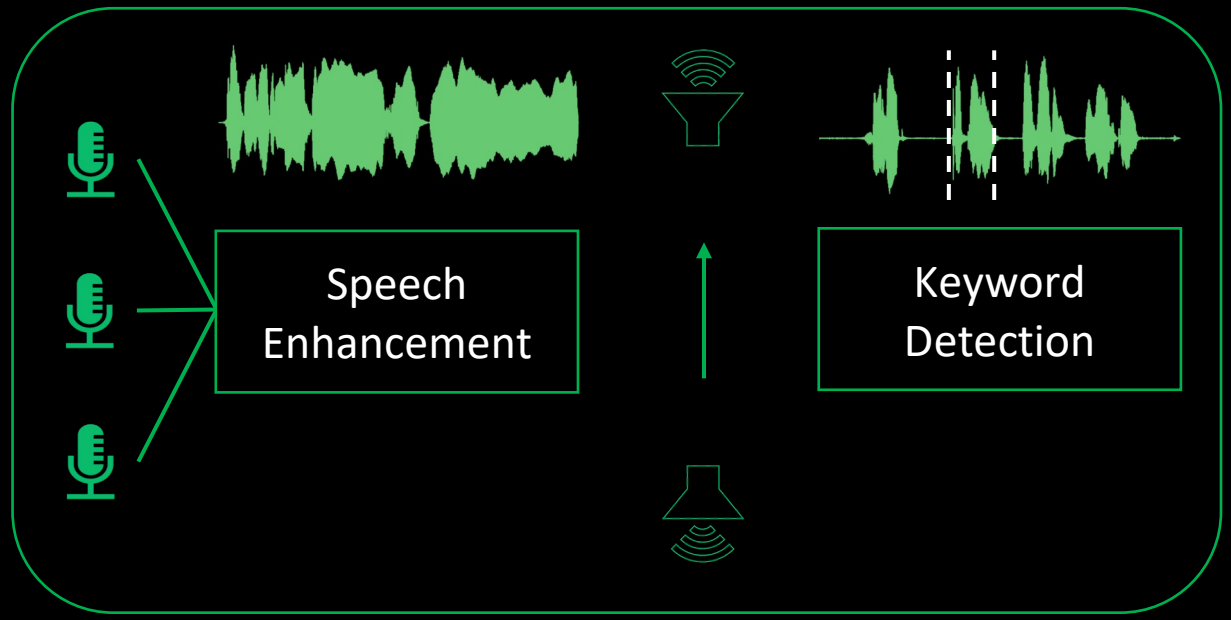
Voice assistant pipeline



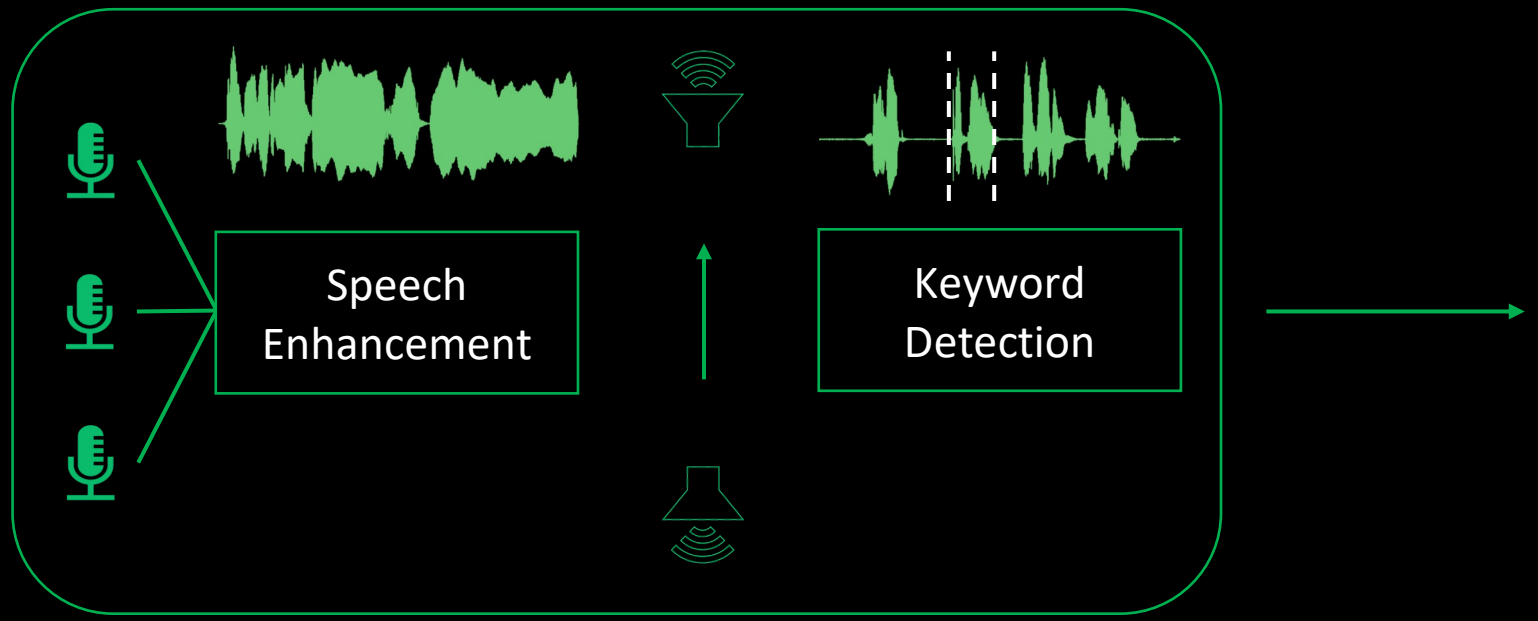
Voice assistant pipeline



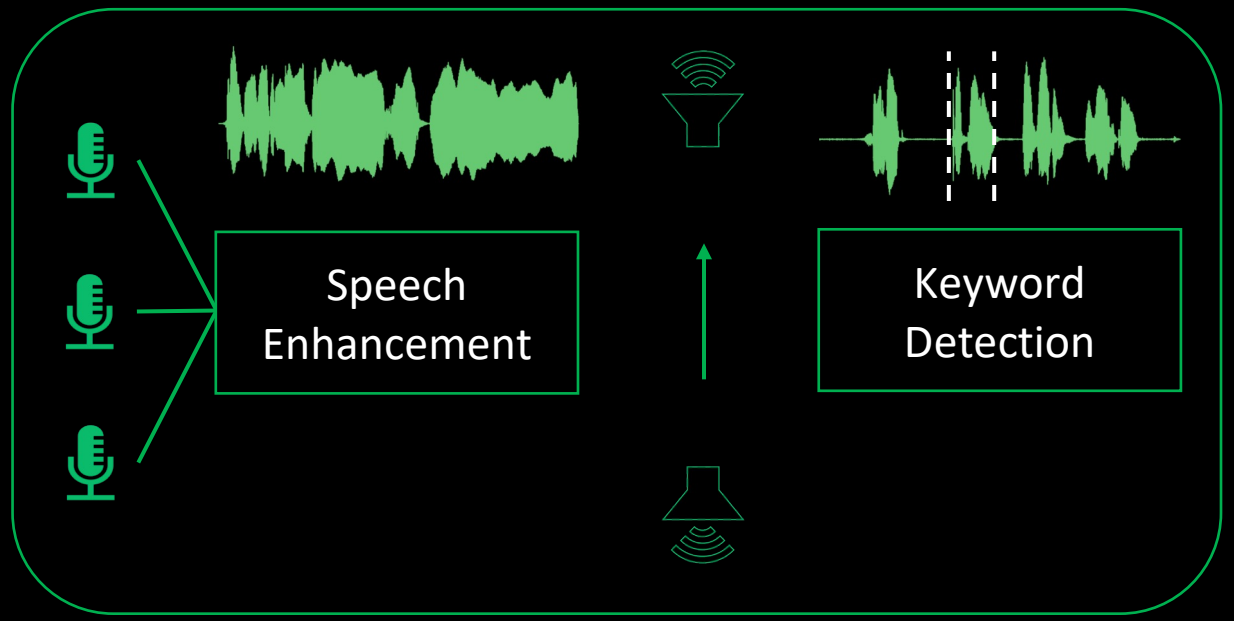
Voice assistant pipeline



Voice assistant pipeline

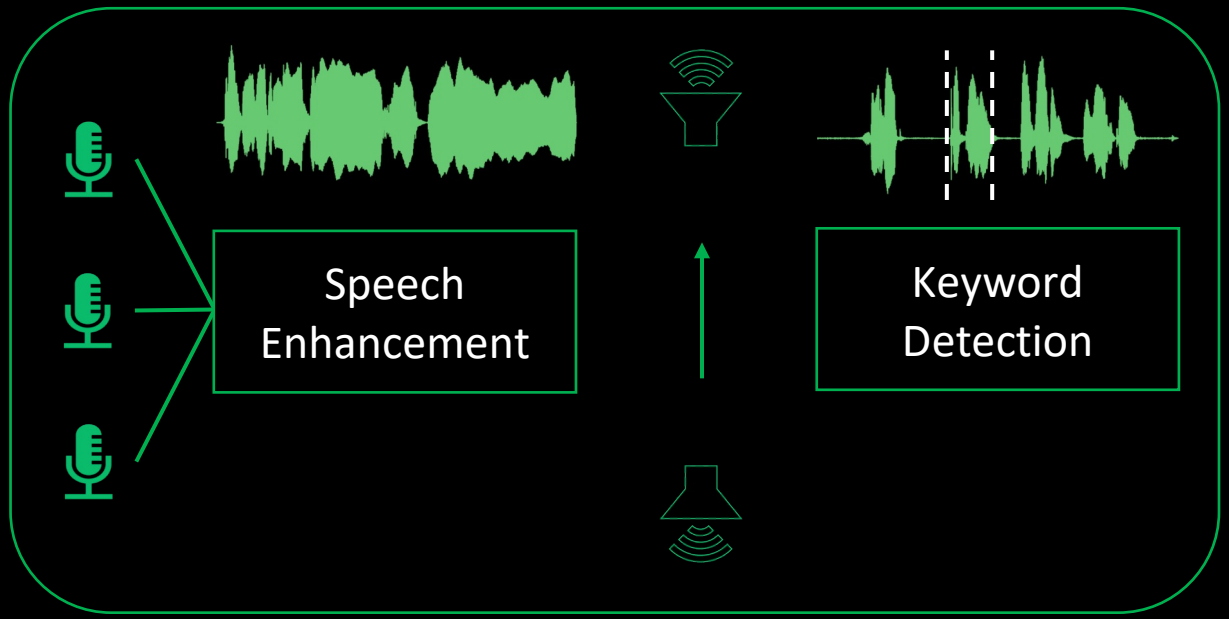


Voice assistant pipeline



Keyword
Validation

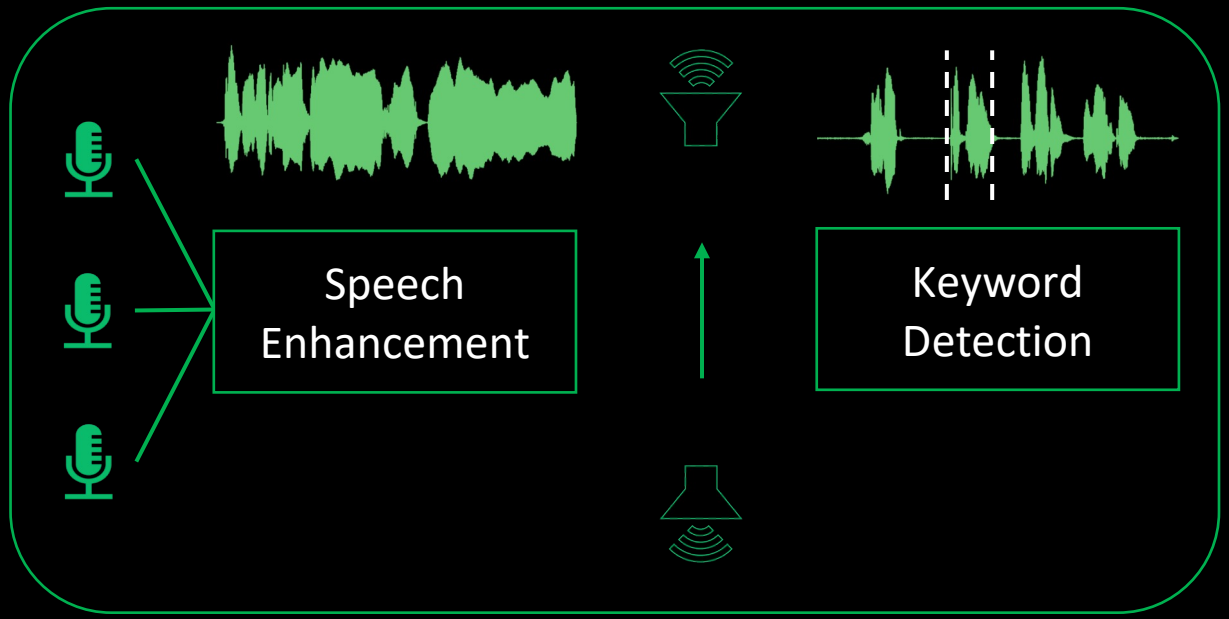
Voice assistant pipeline



Keyword Validation

Intended Query Detection

Voice assistant pipeline

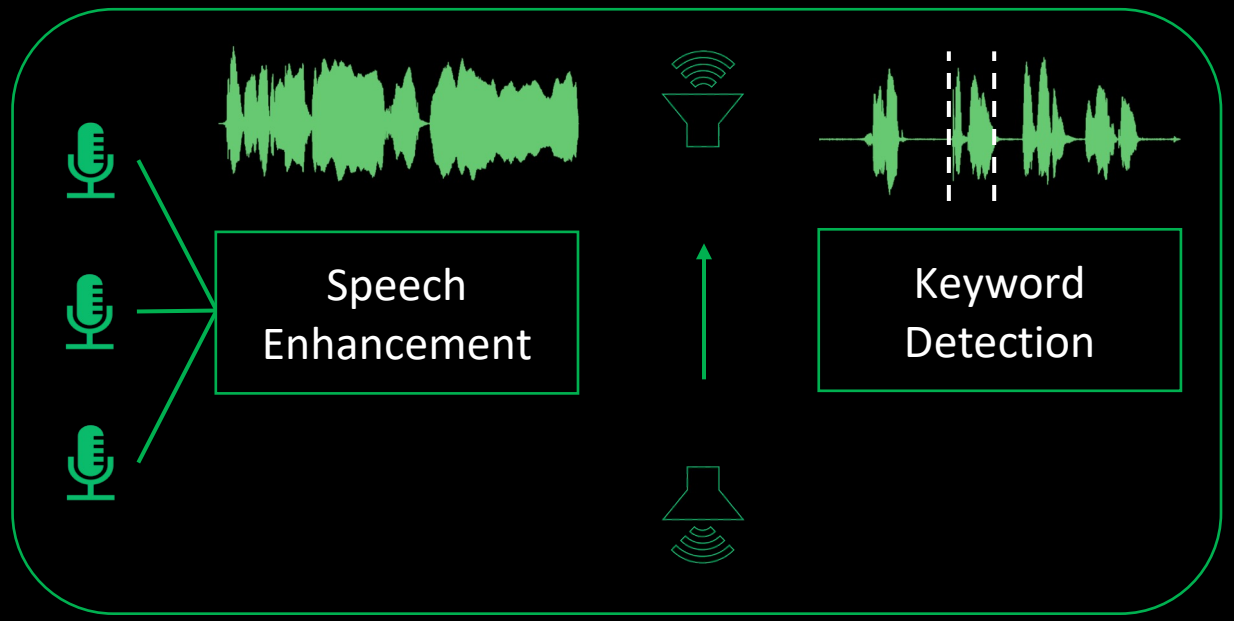


Keyword Validation

Intended Query Detection

Speech Recognition

Voice assistant pipeline



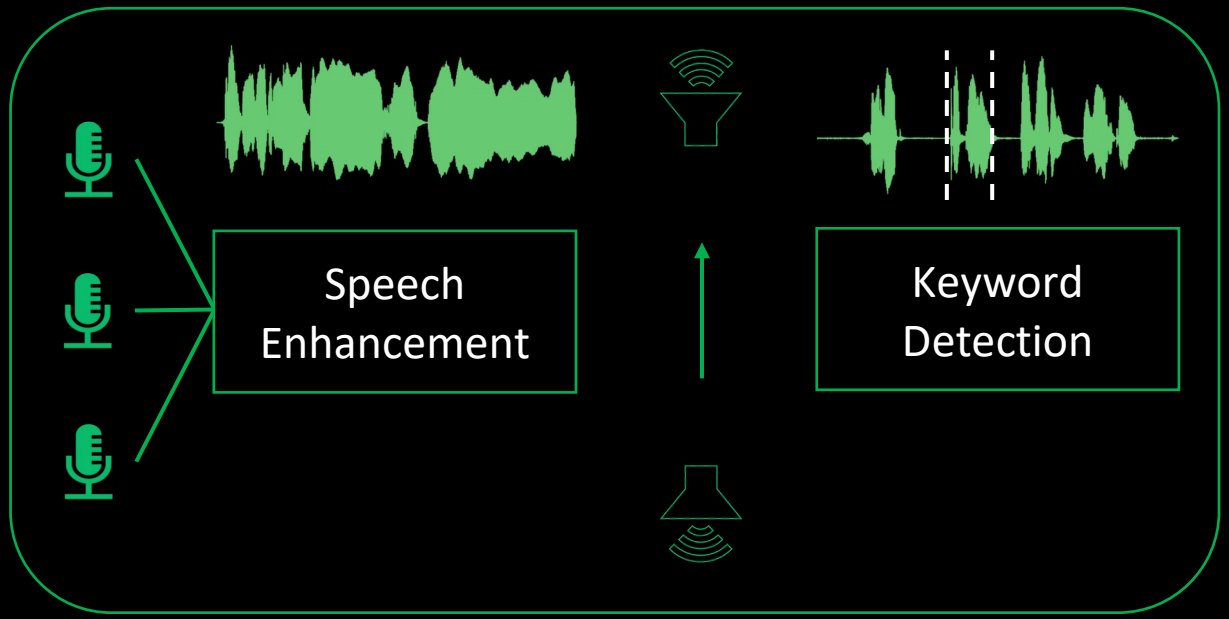
Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Voice assistant pipeline



Keyword Validation

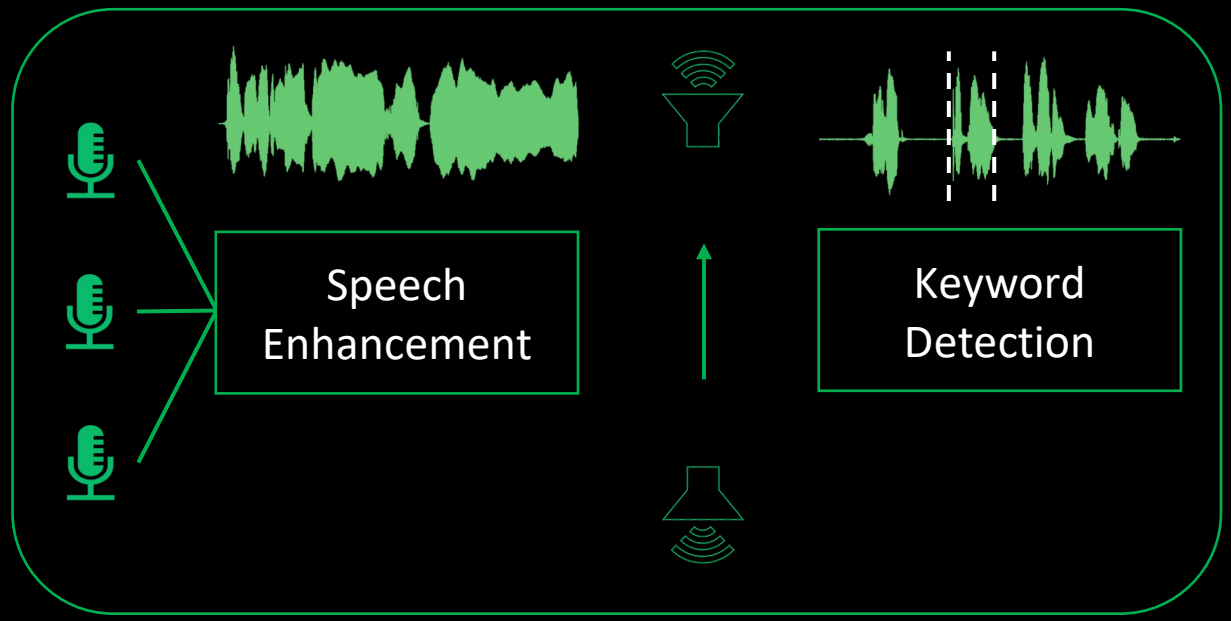
Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

Voice assistant pipeline



Keyword Validation

Intended Query Detection

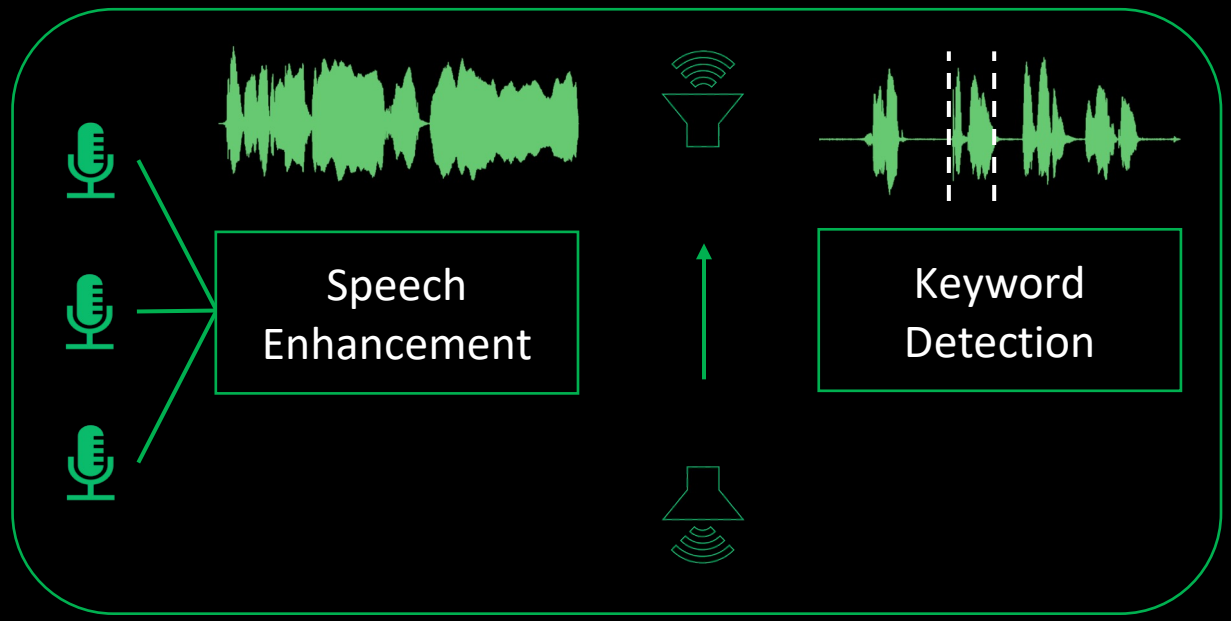
Speech Recognition

Endpoint Detection

Whisper Detection

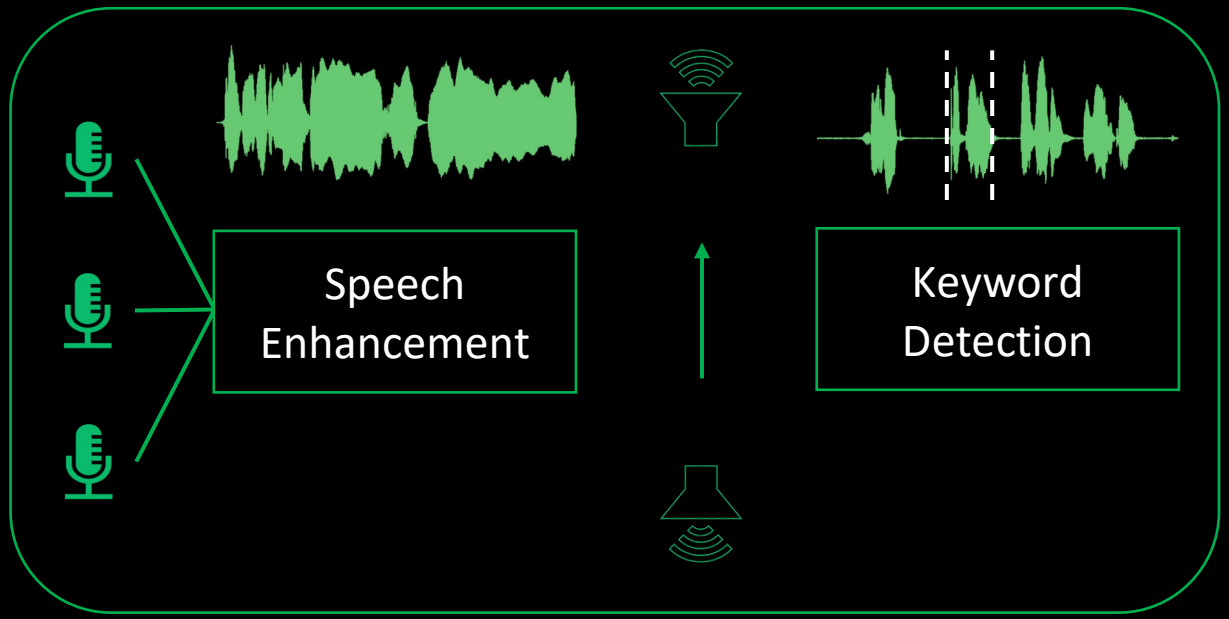
Emotion Detection

Voice assistant pipeline



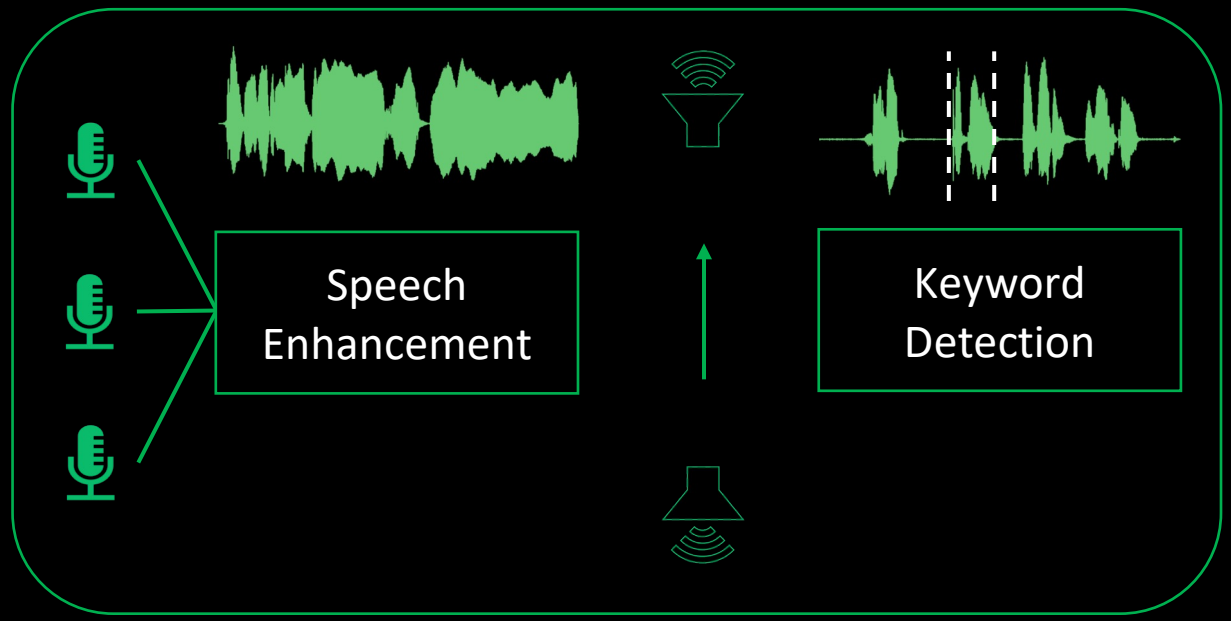
- Keyword Validation
- Intended Query Detection
- Speech Recognition
- Endpoint Detection
- Whisper Detection
- Emotion Detection
- Speaker Identification

Voice assistant pipeline



- Keyword Validation
- Intended Query Detection
- Speech Recognition
- Endpoint Detection
- Whisper Detection
- Emotion Detection
- Speaker Identification
- Speech Separation

Voice assistant pipeline



Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

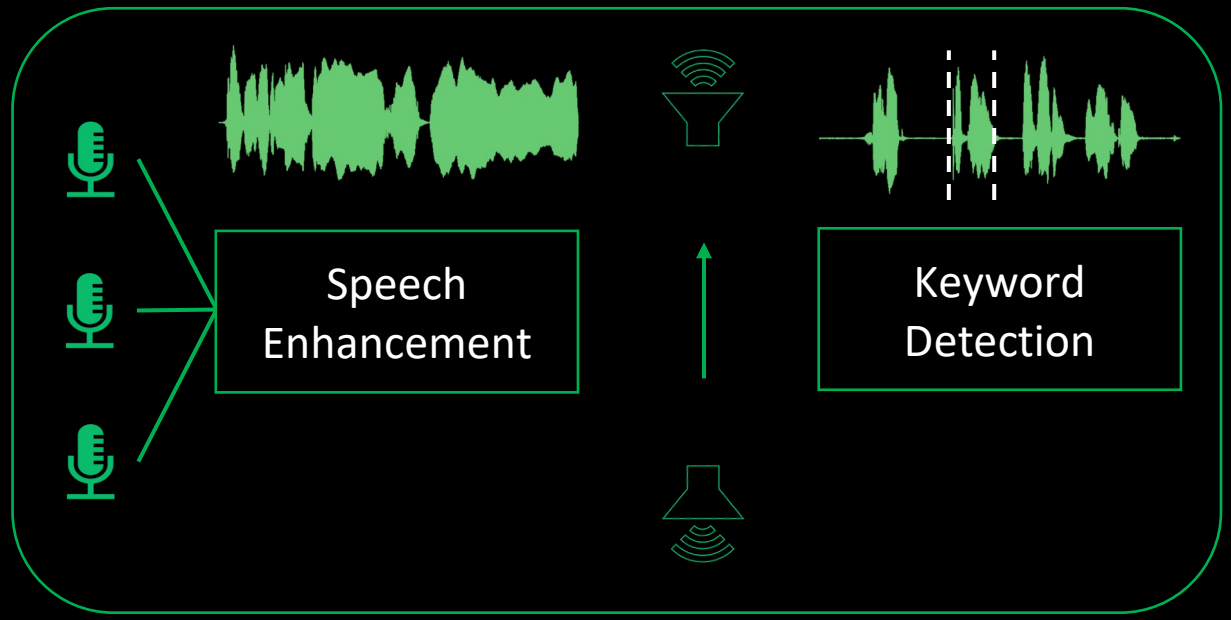
Emotion Detection

Speaker Identification

Speech Separation

салют какая погода в москве

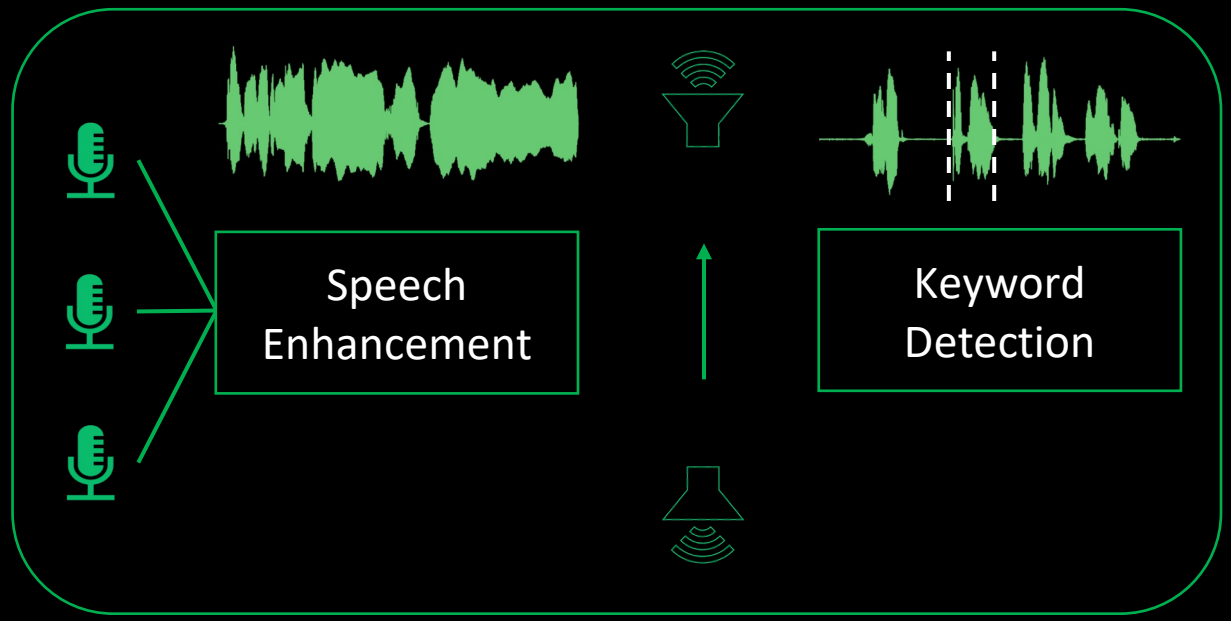
Voice assistant pipeline



- Keyword Validation
- Intended Query Detection
- Speech Recognition
- Endpoint Detection
- Whisper Detection
- Emotion Detection
- Speaker Identification
- Speech Separation
- Text Normalization

салют какая погода в москве

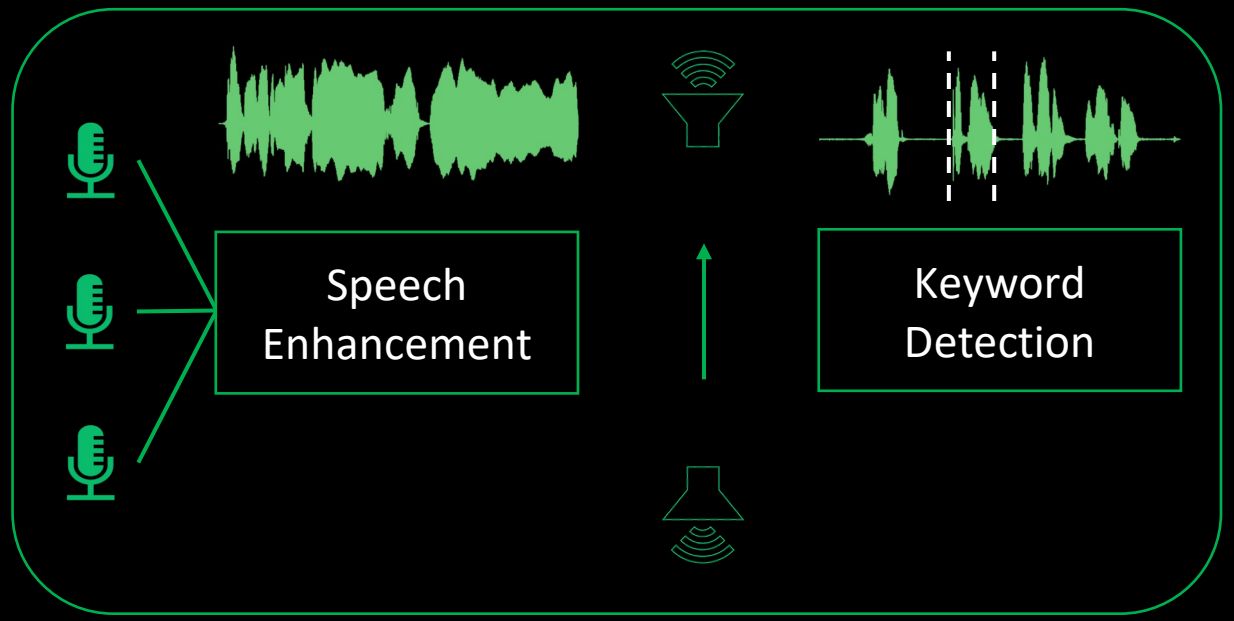
Voice assistant pipeline



- Keyword Validation
- Intended Query Detection
- Speech Recognition
- Endpoint Detection
- Whisper Detection
- Emotion Detection
- Speaker Identification
- Speech Separation
- Text Normalization

Салют какая погода в Москве

Voice assistant pipeline



Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

Emotion Detection

Speaker Identification

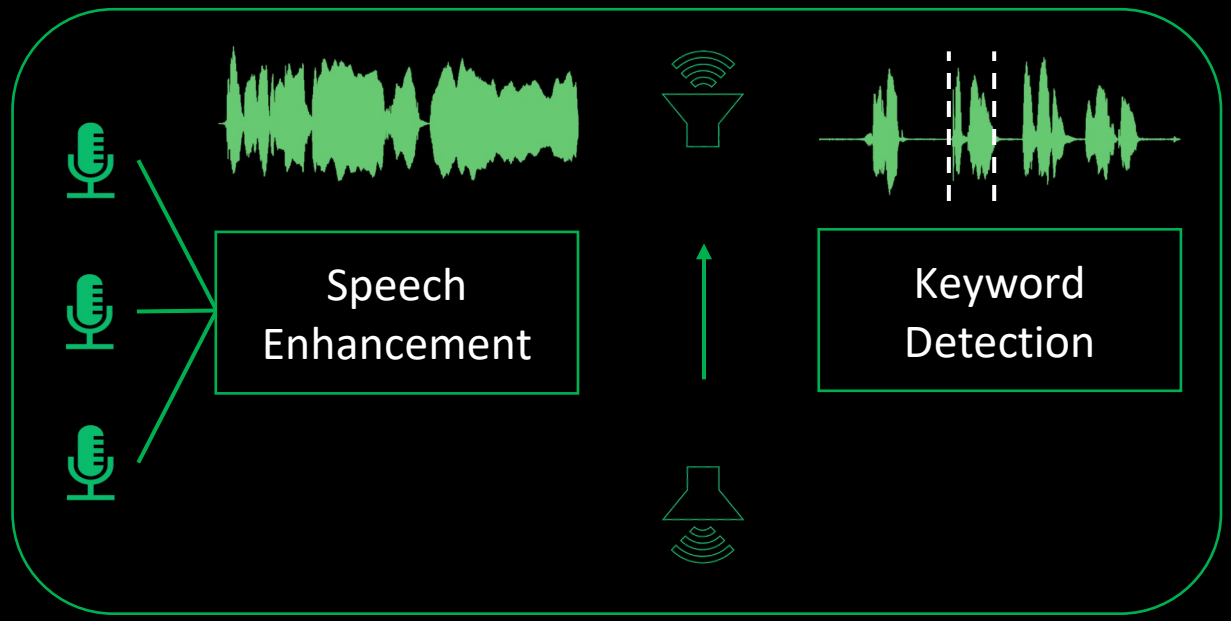
Speech Separation

Text Normalization

Punctuation Prediction

Салют какая погода в Москве

Voice assistant pipeline



Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

Emotion Detection

Speaker Identification

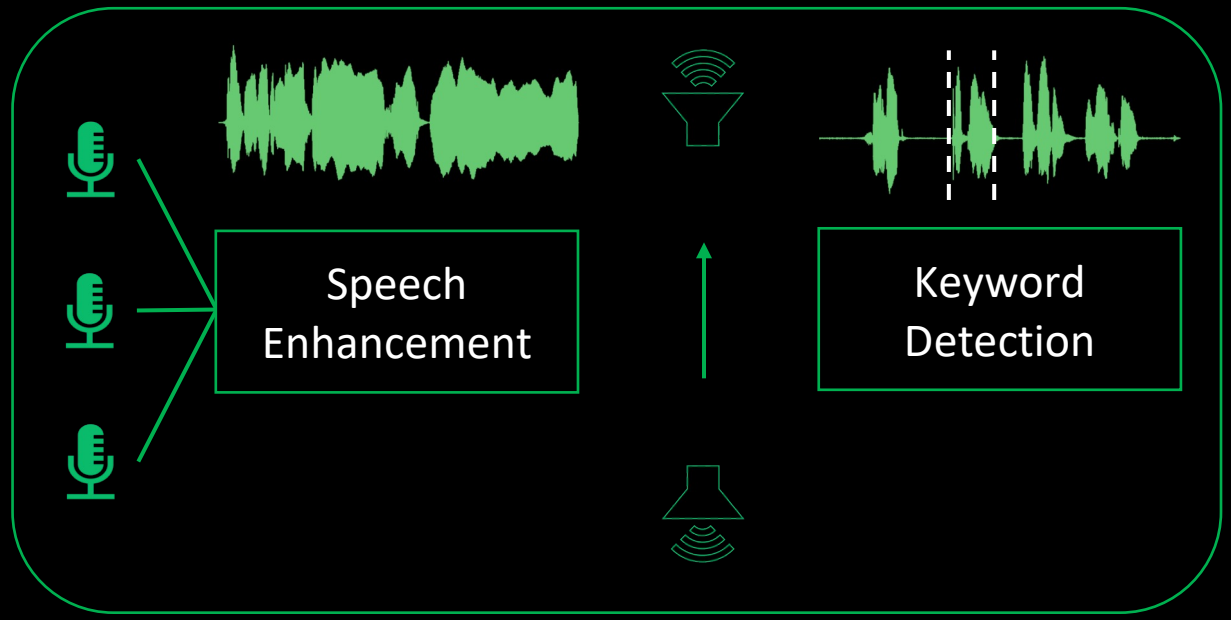
Speech Separation

Text Normalization

Punctuation Prediction

Салют, какая погода в Москве?

Voice assistant pipeline



Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

Emotion Detection

Speaker Identification

Speech Separation

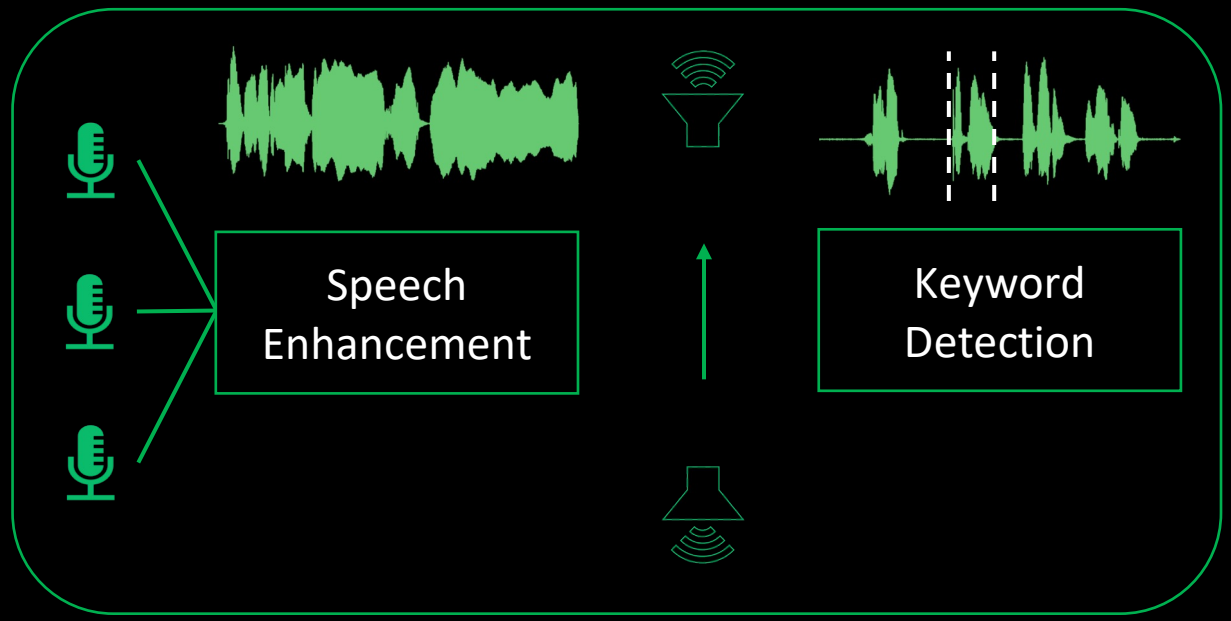
Text Normalization

Punctuation Prediction

Салют, какая погода в Москве?

Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Voice assistant pipeline



Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

Emotion Detection

Speaker Identification

Speech Separation

Text Normalization

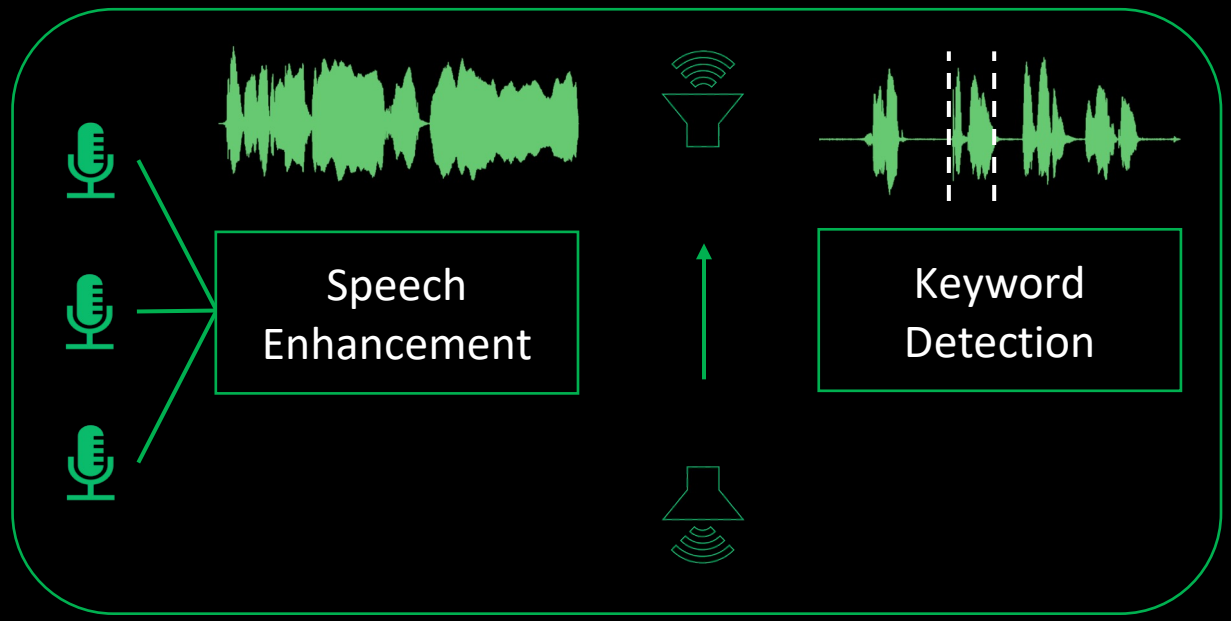
Punctuation Prediction

Салют, какая погода в Москве?

Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Language Generation

Voice assistant pipeline



Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

Emotion Detection

Speaker Identification

Speech Separation

Text Normalization

Punctuation Prediction

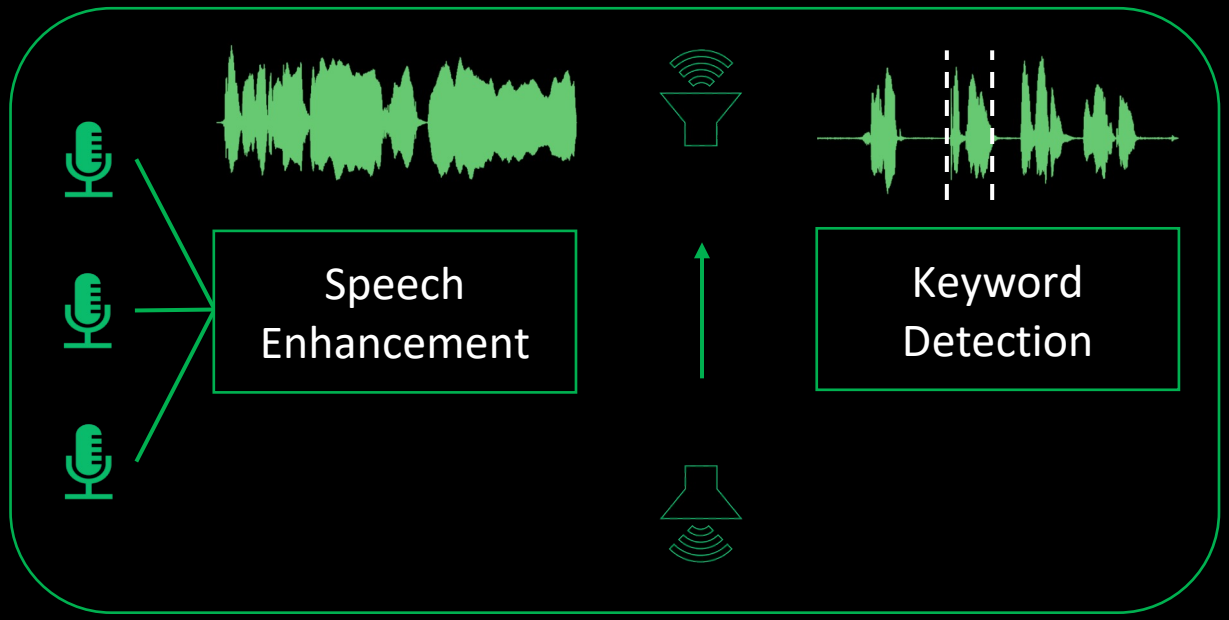
Салют, какая погода в Москве?

Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Language Generation

Сейчас в Москве +15 °C

Voice assistant pipeline



Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

Emotion Detection

Speaker Identification

Speech Separation

Text Normalization

Punctuation Prediction

Салют, какая погода в Москве?

Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

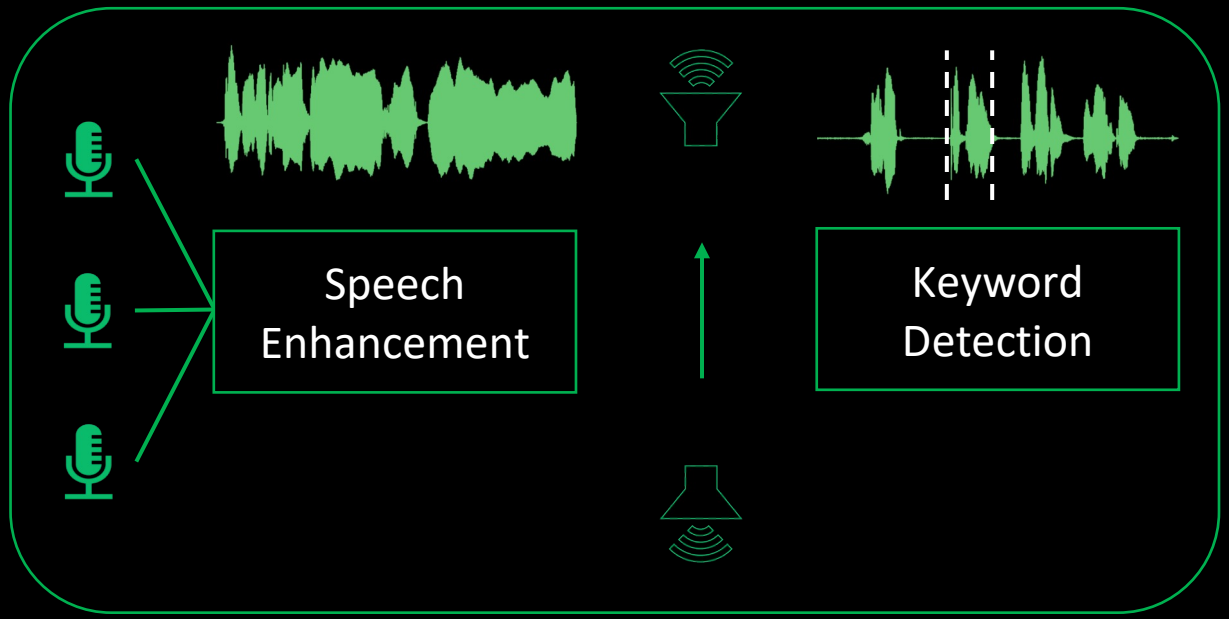
Language Generation

Сейчас в Москве +15 ºC

Text Normalization

Speech Synthesis

Voice assistant pipeline



Keyword Validation

Intended Query Detection

Speech Recognition

Endpoint Detection

Whisper Detection

Emotion Detection

Speaker Identification

Speech Separation

Text Normalization

Punctuation Prediction

Салют, какая погода в Москве?

Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Language Generation

Сейчас в Москве +15 ºC

Text Normalization

Speech Synthesis

Voice assistant pipeline

Какие основные требования у пользователей?
- Простота настройки и интеграции с другими системами
- Приватность и безопасность
- Качество звука и синтеза речи
- Поддержка широкого спектра возможностей
- Скорость отклика

User perceived latency

Скорость отклика
- Быстрое выполнение команд, без задержек и пауз

User perceived latency – метрика, отражающая воспринимаемое пользователем время задержки между действием и реакцией системы.

В нашем случае можем выделить два примера:
- Произношение активационной фразы <-> активация колонки
- Произношение голосового запроса <-> получение ответа

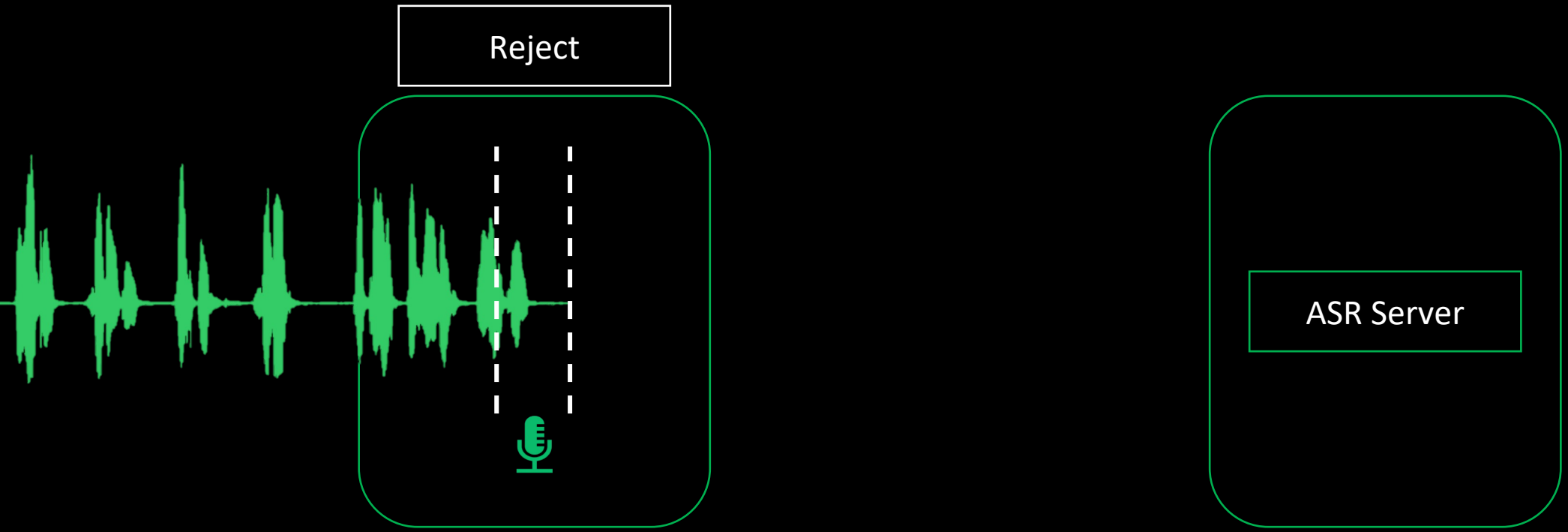
Keyword detection

Основная задача в рамках Voice Assistant Pipeline
- Детектирование ключевых фраз в аудиопотоке для активации устройства

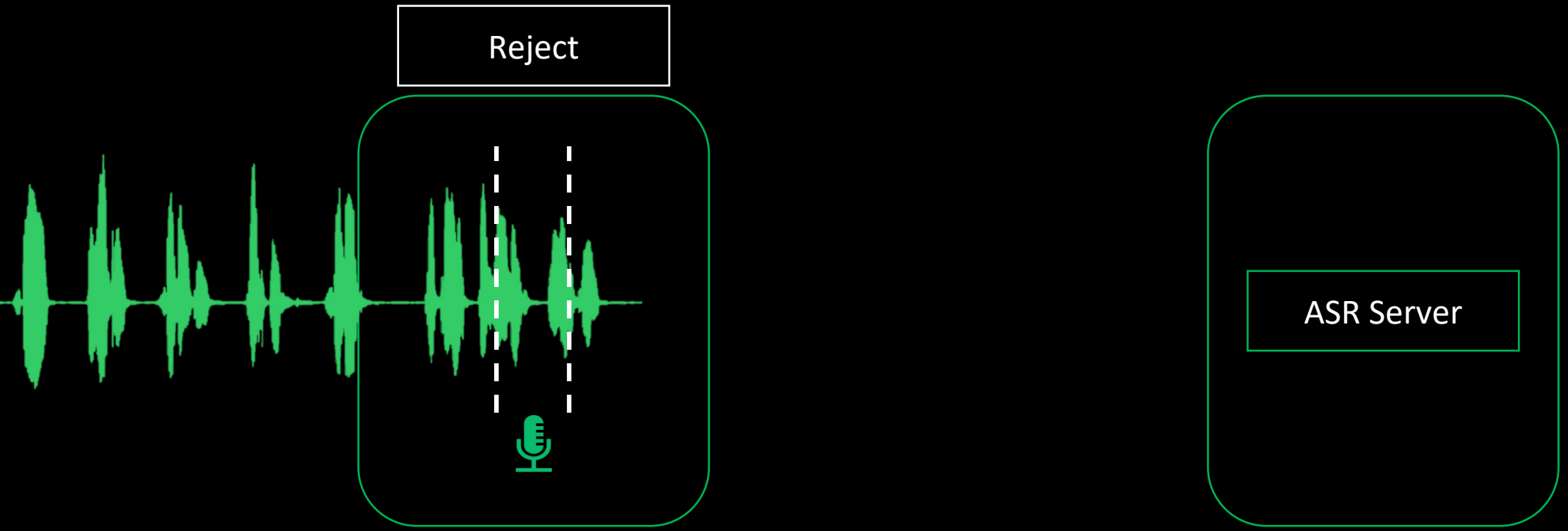
- Салют
- Сбер
- Афина
- Джой

Keyword detection: on-device

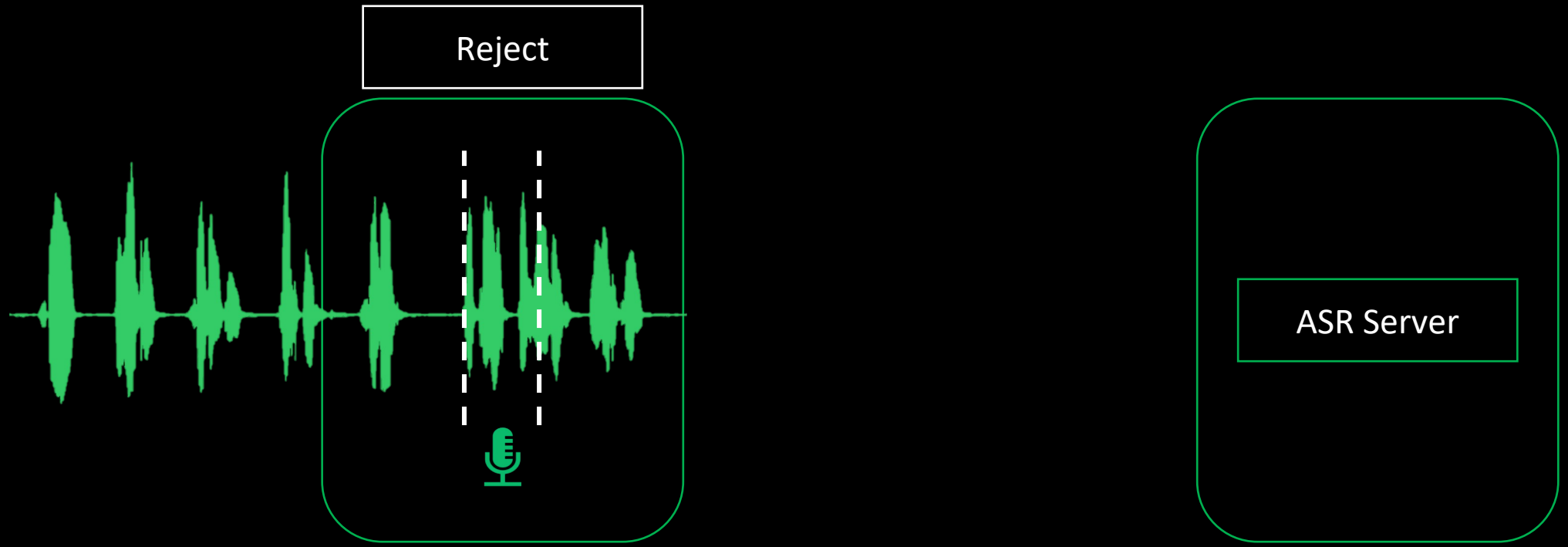
Keyword detection: on-device



Keyword detection: on-device

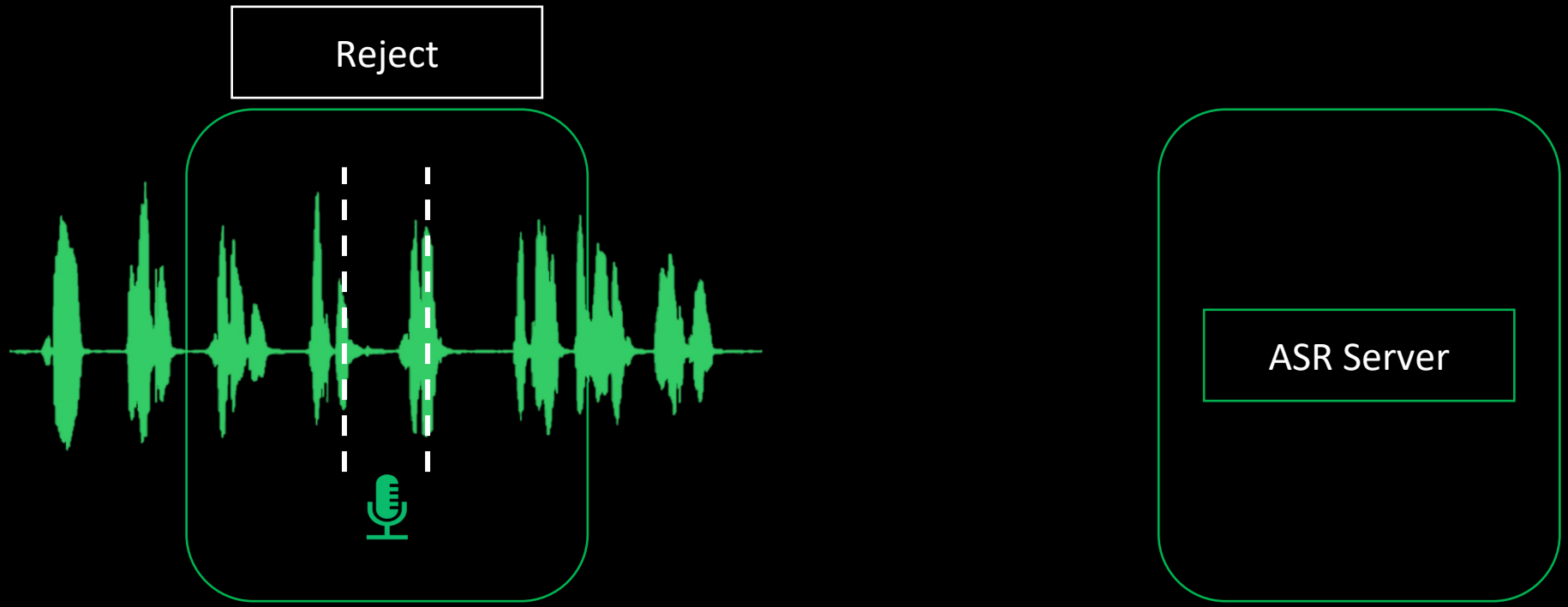


Keyword detection: on-device

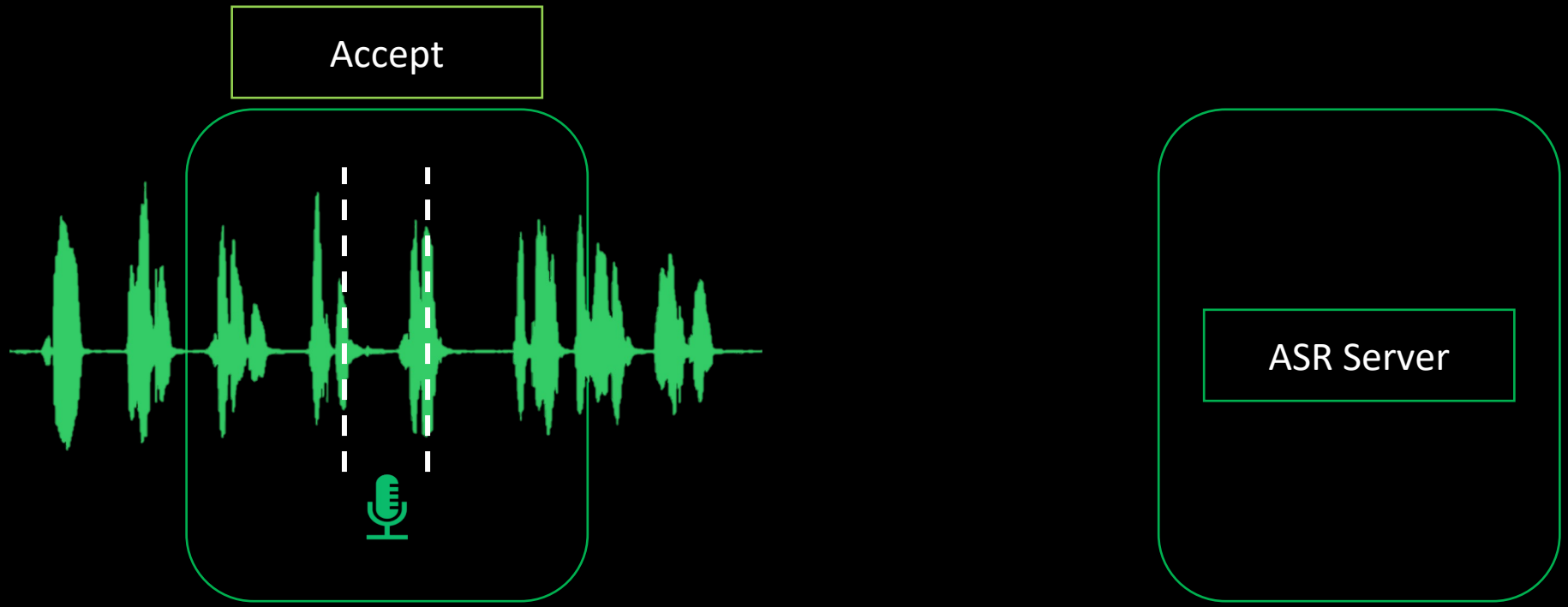


Keyword detection: on-device

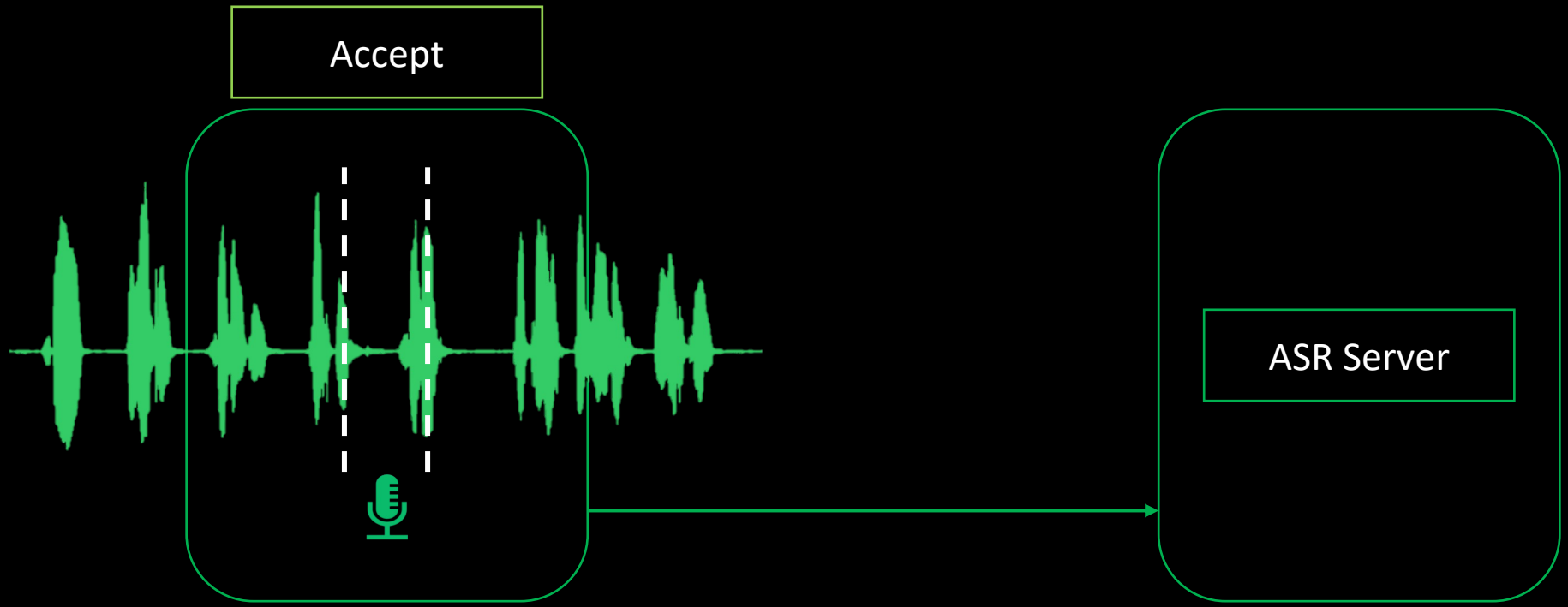
Keyword detection: on-device



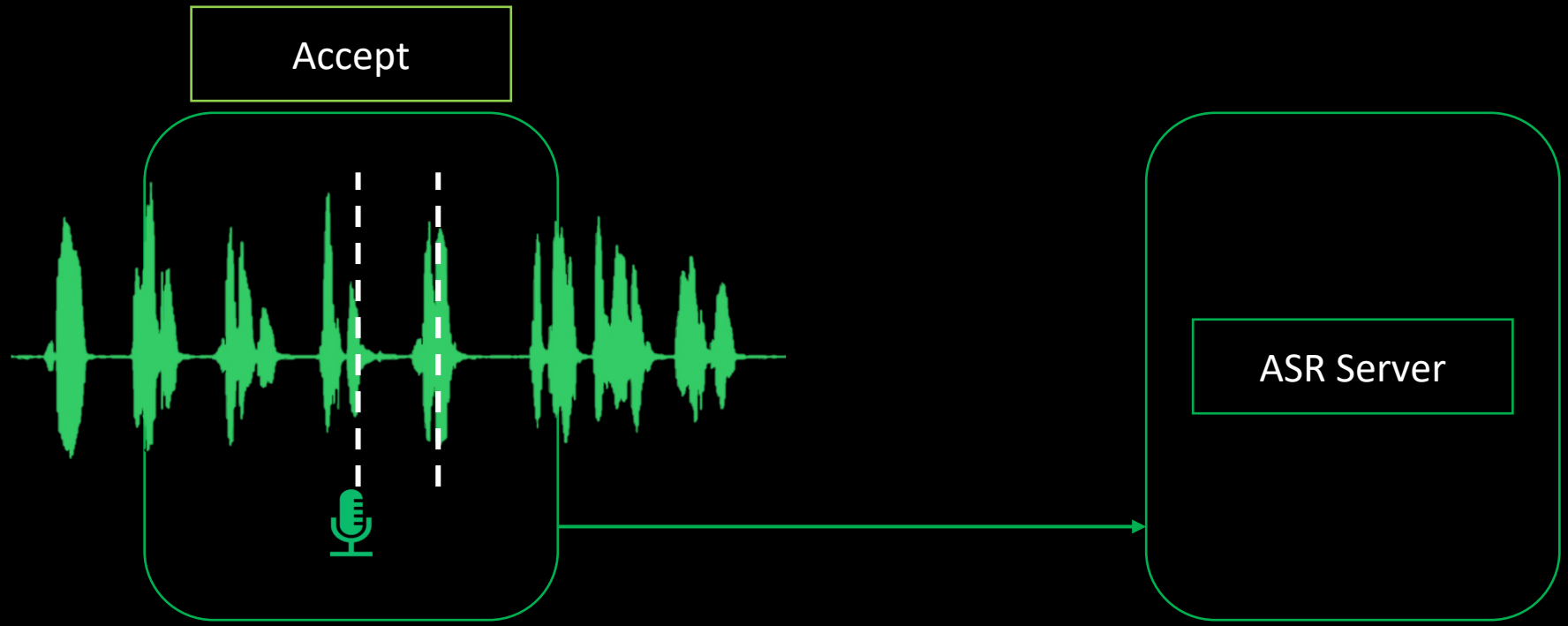
Keyword detection: on-device



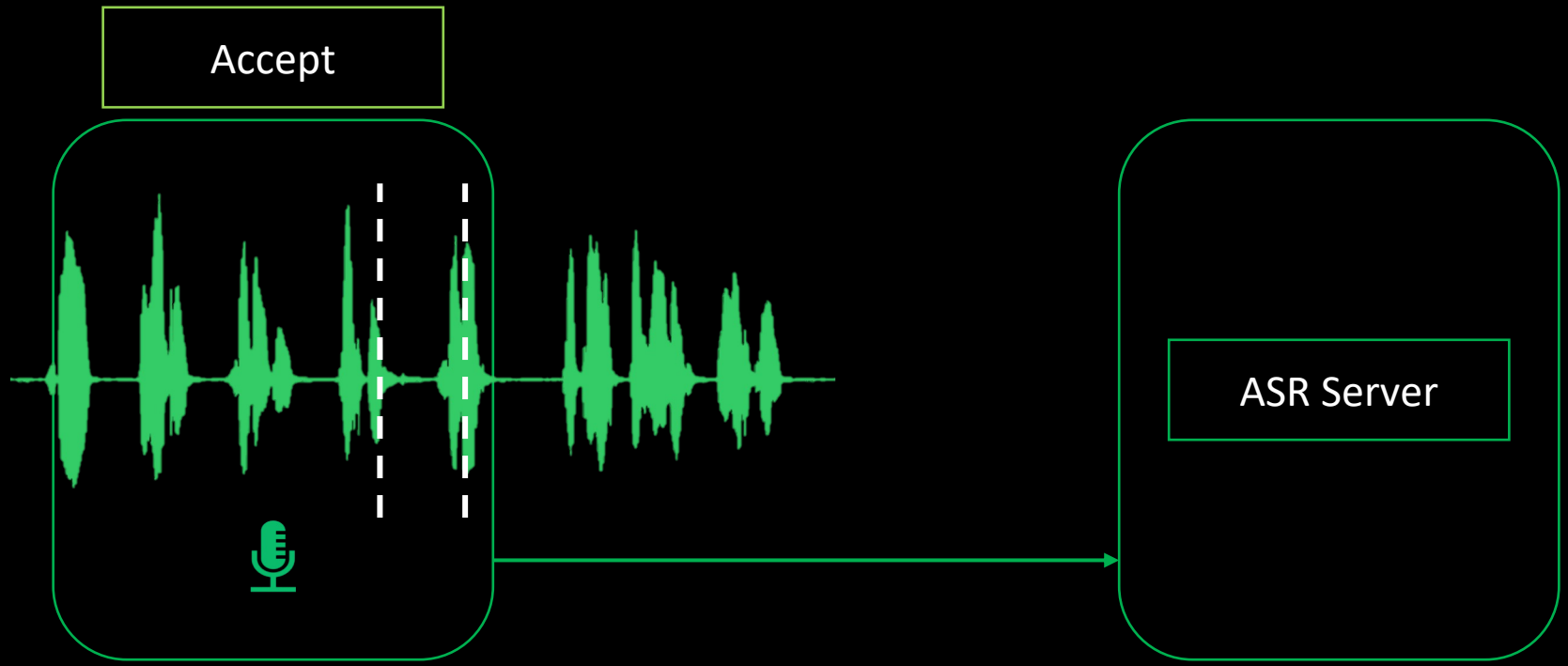
Keyword detection: on-device



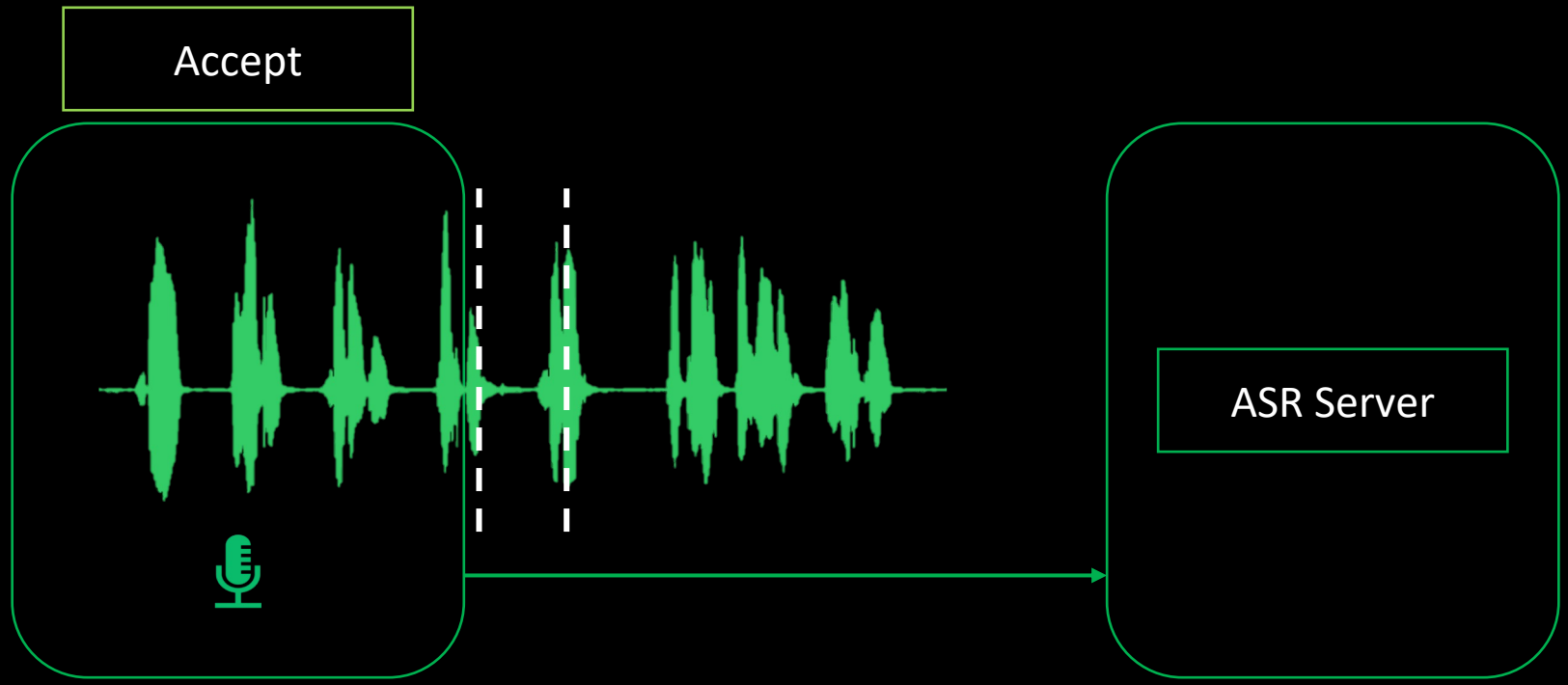
Keyword detection: on-device



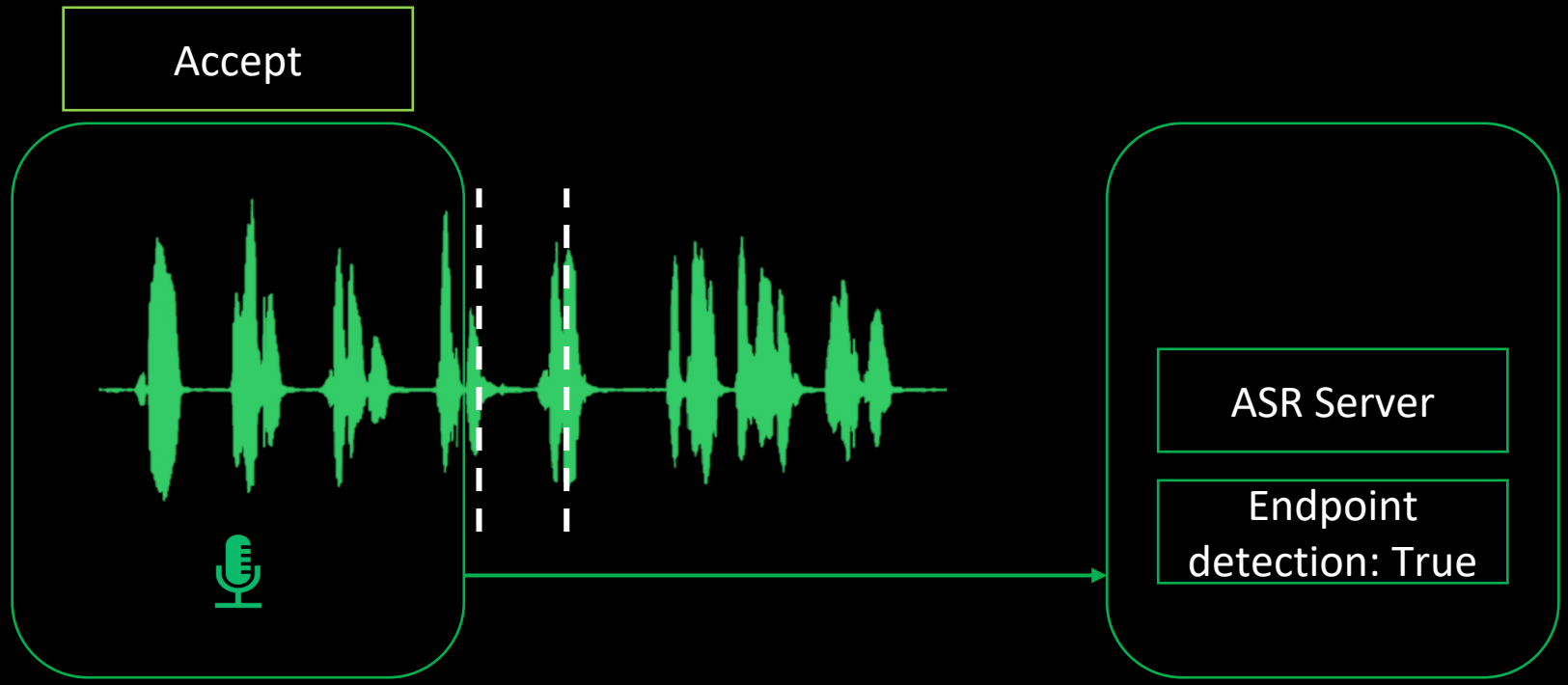
Keyword detection: on-device



Keyword detection: on-device



Keyword detection: on-device



Keyword detection: on-device

Keyword detection: on-device

ASR Server

Endpoint detection: True

Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Keyword detection: on-device

Преимущества:

- Простота реализации системы
- Скорость инференса

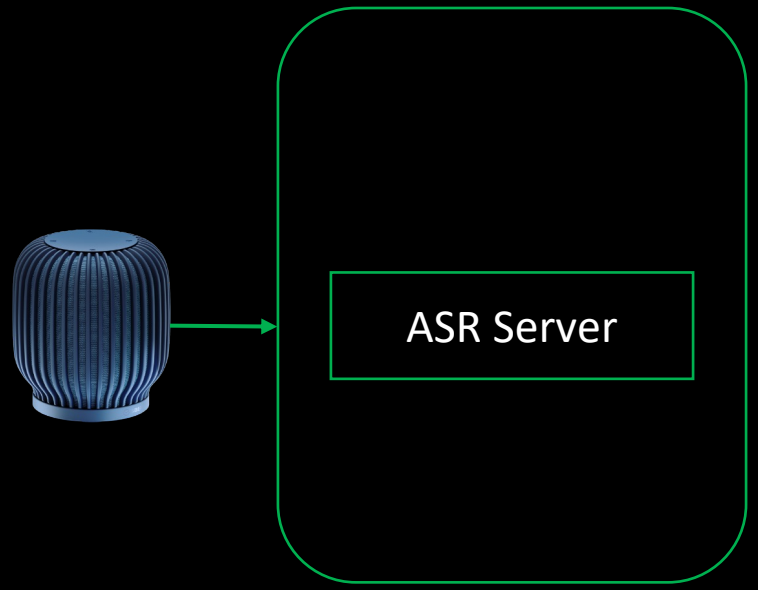
Недостатки:
- Компромисс между точностью, ресурсоемкостью и скоростью

Keyword detection: on-device

Почему on-device?

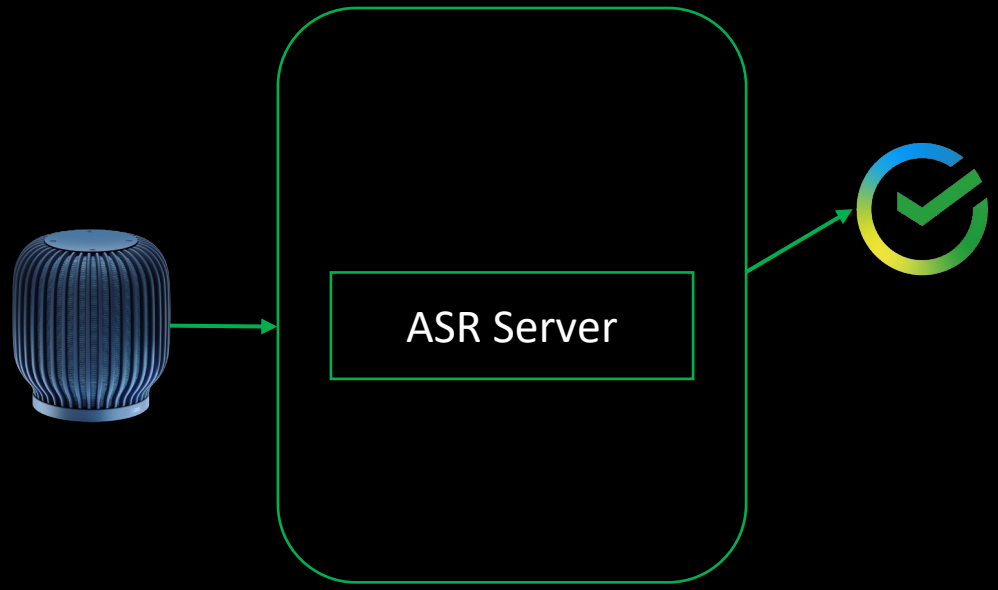
Keyword detection: on-device

Почему on-device?
- Конфиденциальность



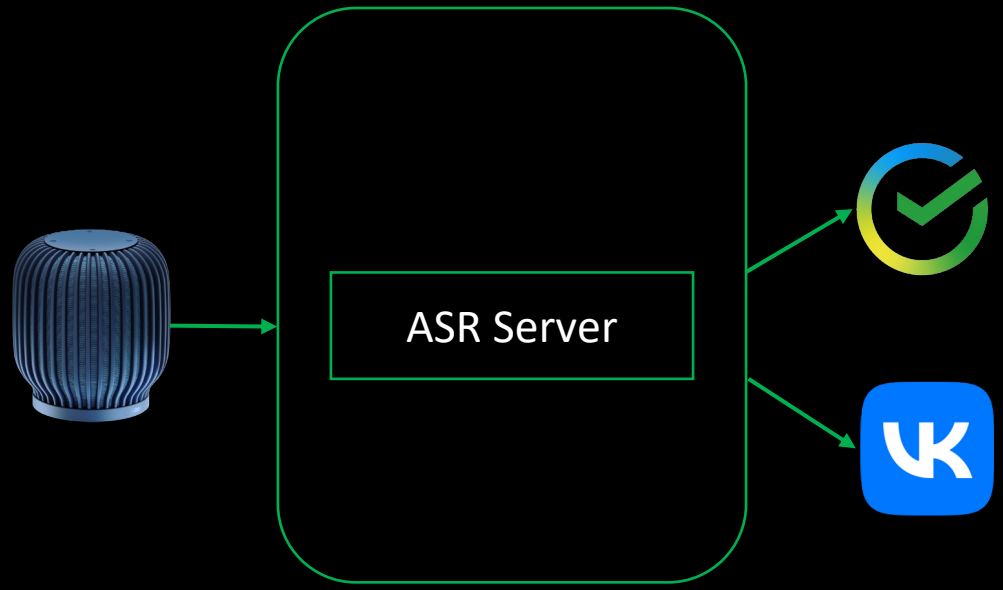
Keyword detection: on-device

Почему on-device?
- Конфиденциальность



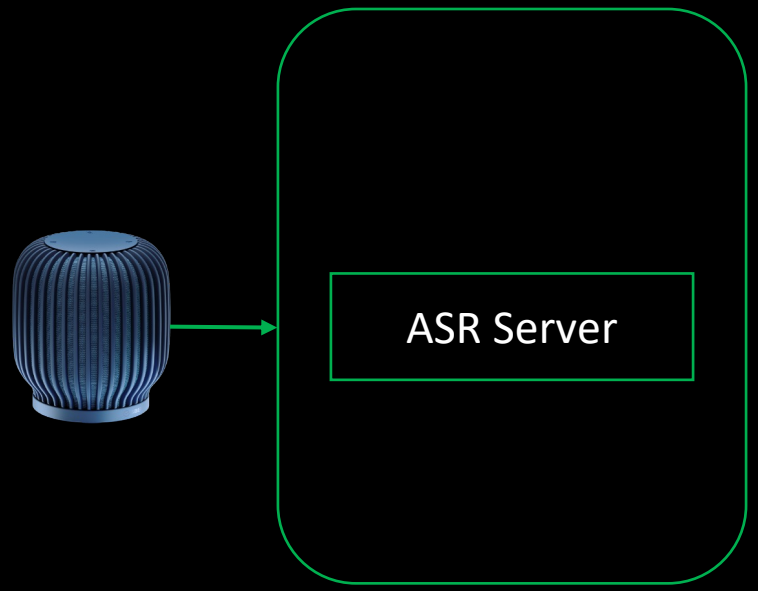
Keyword detection: on-device

Почему on-device?
- Конфиденциальность



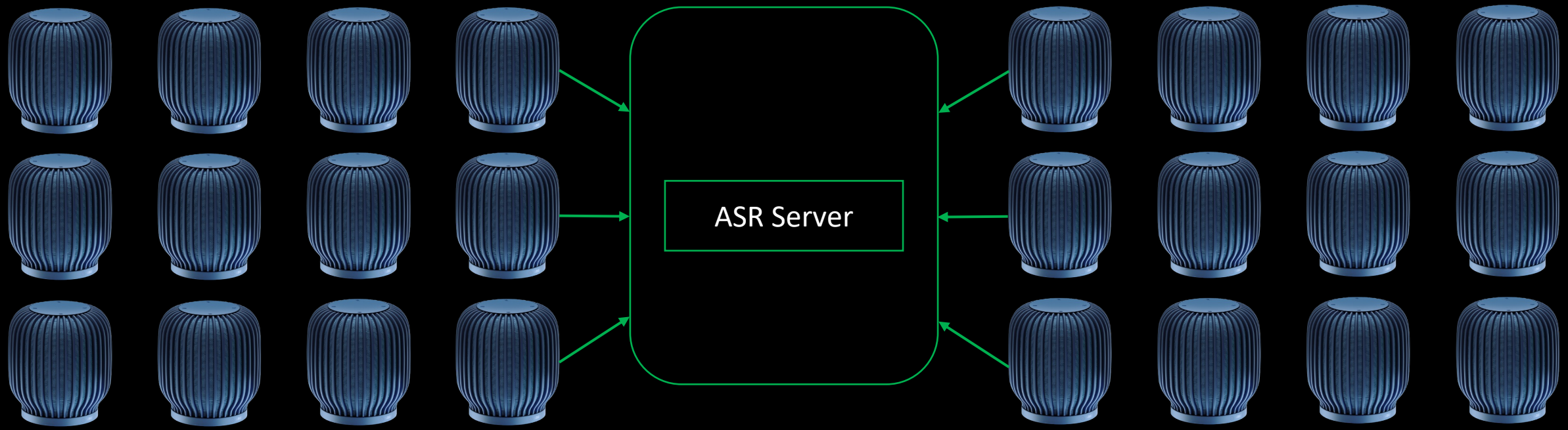
Keyword detection: on-device

Почему on-device?
- Конфиденциальность
- Вычислительные мощности

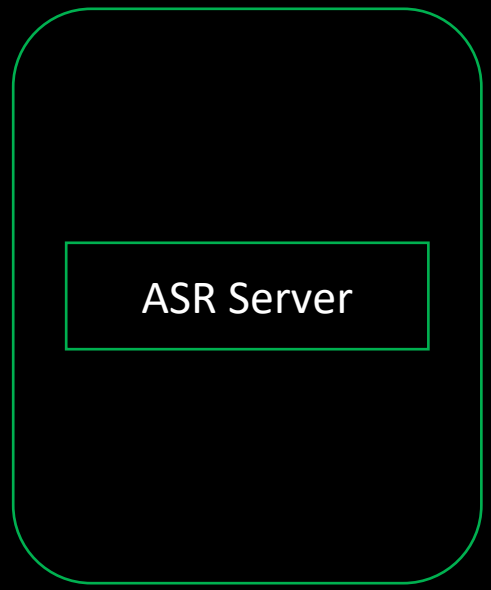


Keyword detection: on-device

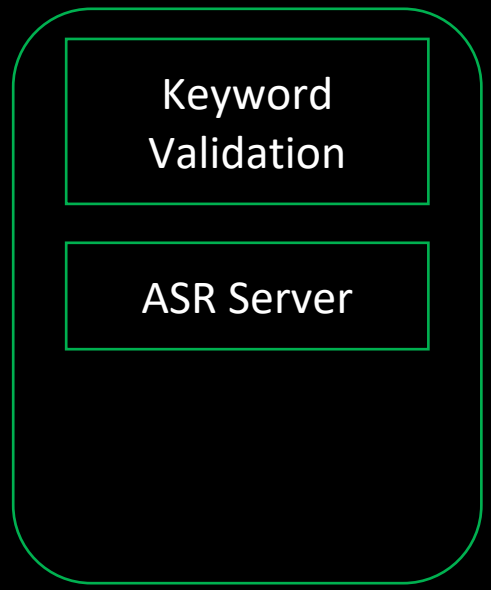
Почему on-device?
- Конфиденциальность
- Вычислительные мощности



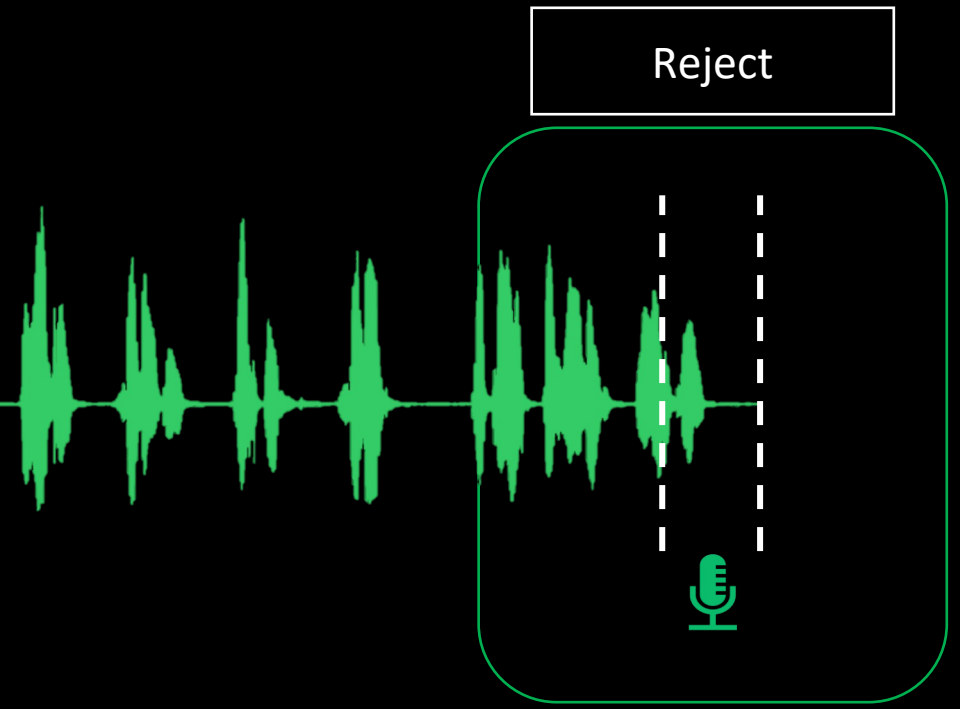
Keyword detection: cascade system

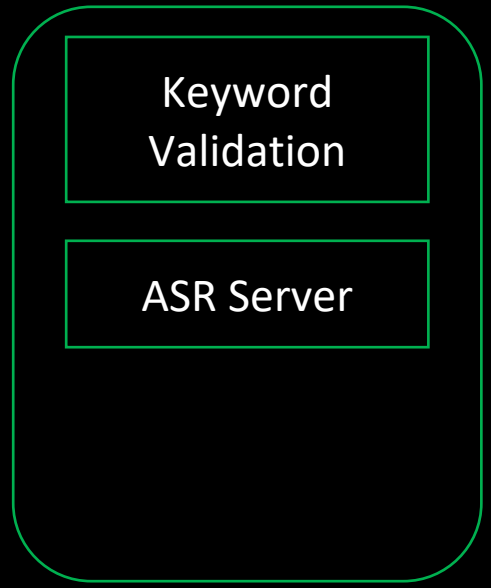


Keyword detection: cascade system

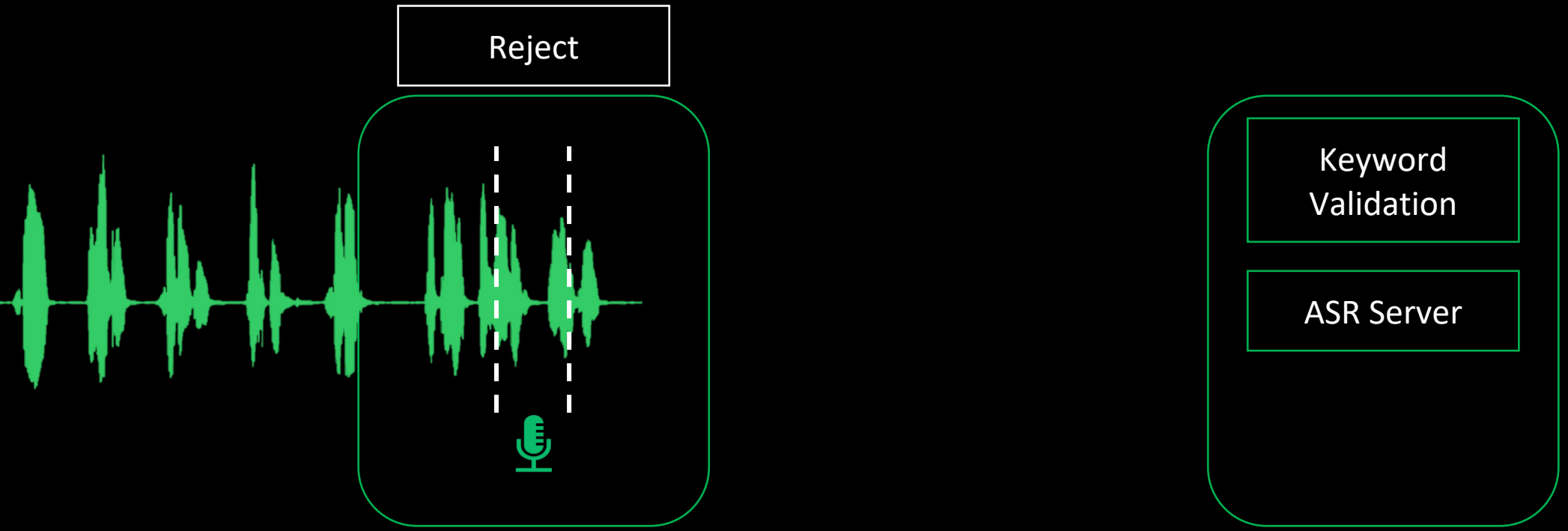


Keyword detection: cascade system

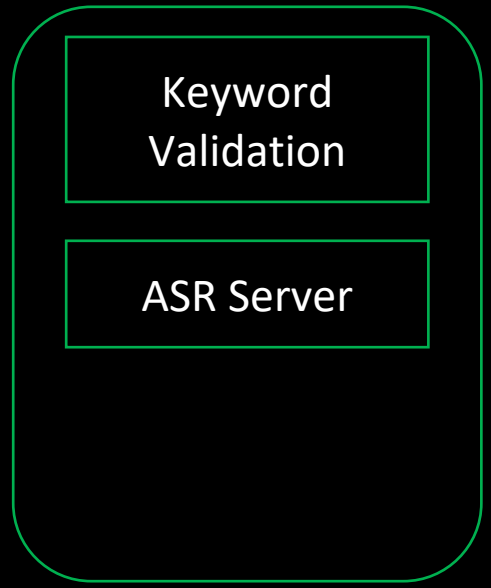




Keyword detection: cascade system

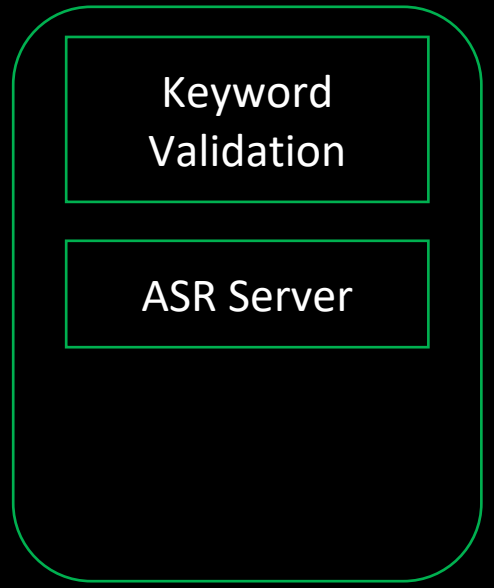


Keyword detection: cascade system

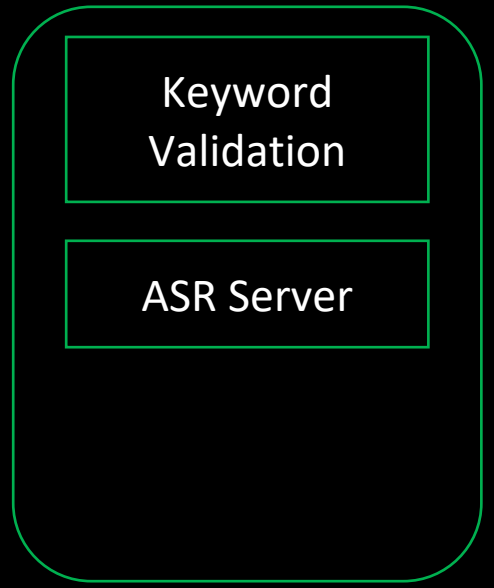


Keyword detection: cascade system

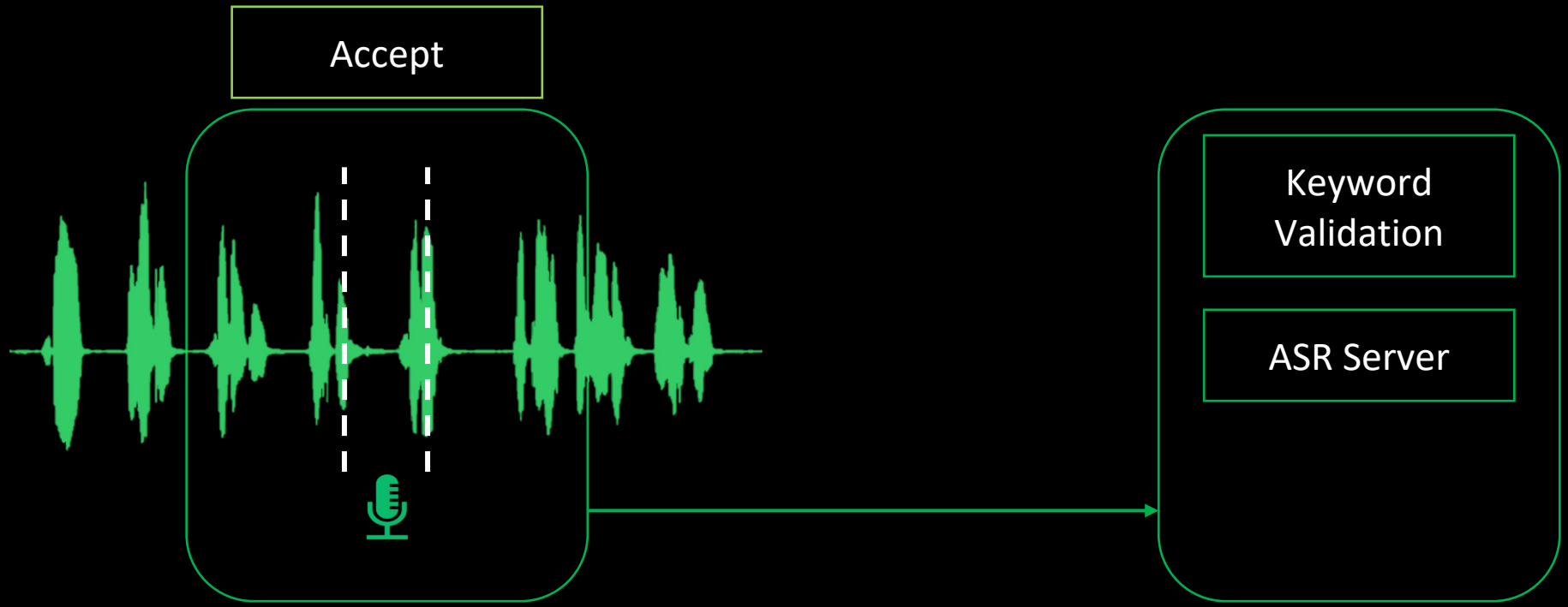
Keyword detection: cascade system



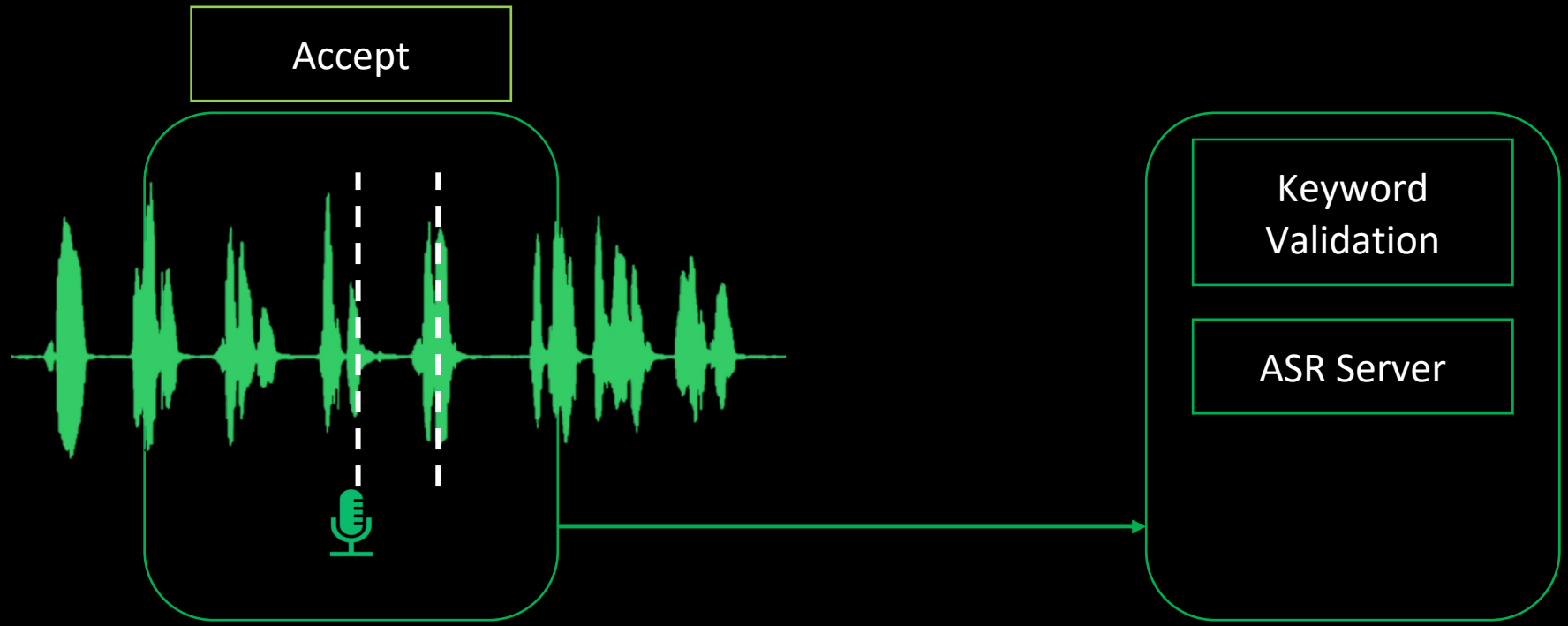
Keyword detection: cascade system



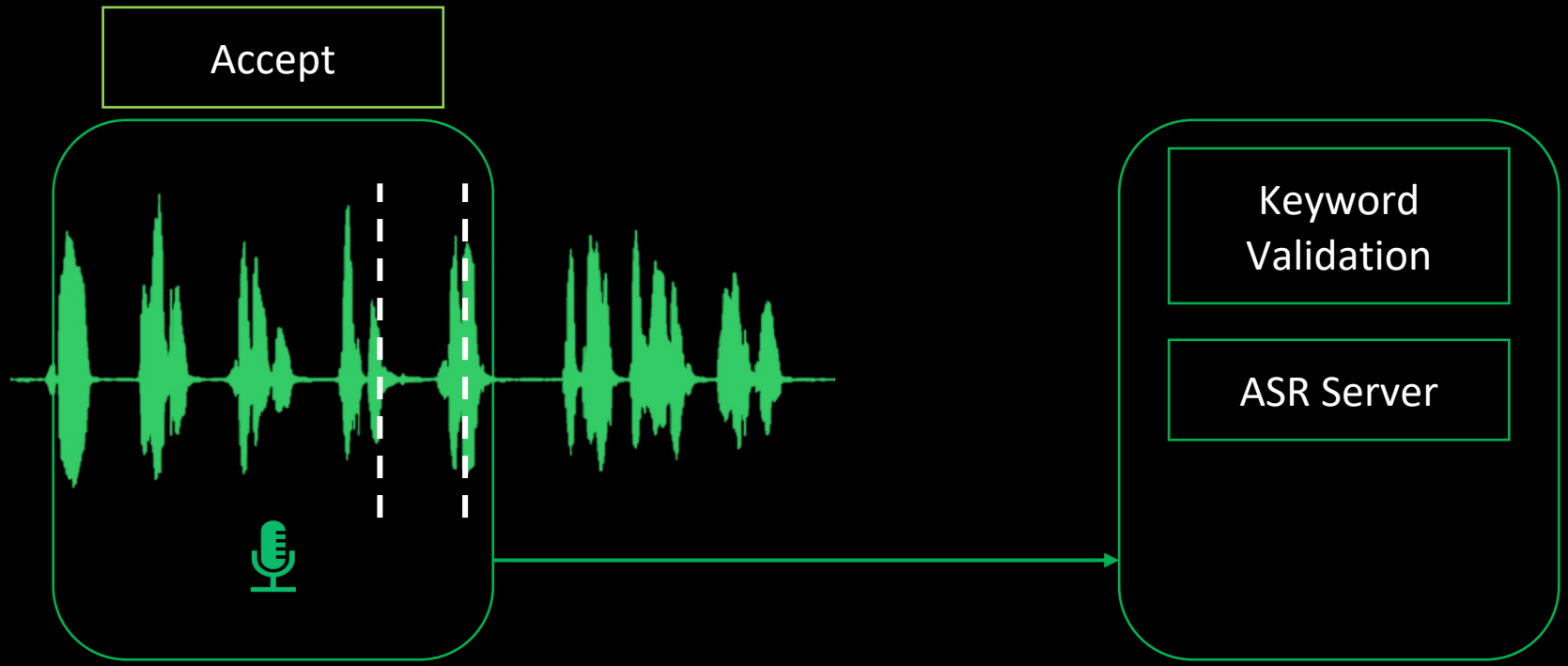
Keyword detection: cascade system



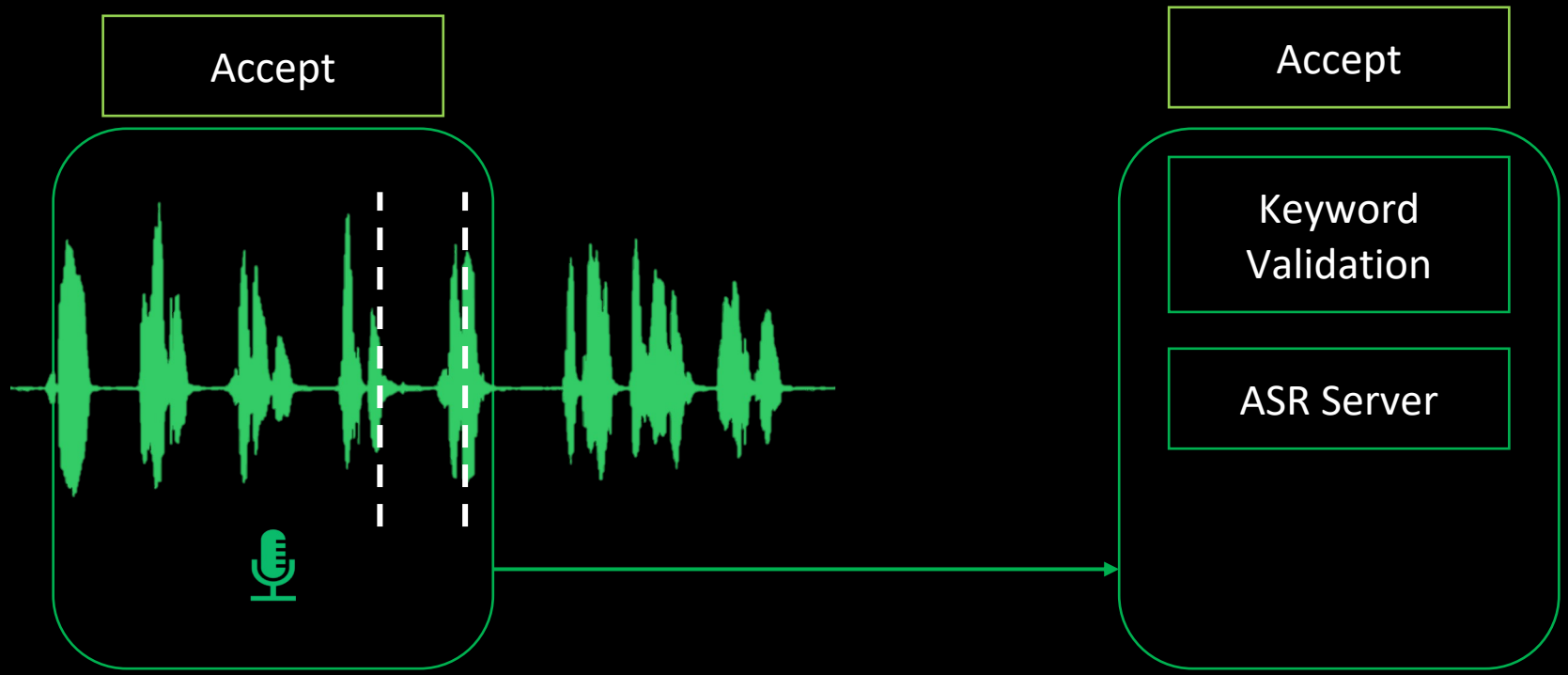
Keyword detection: cascade system



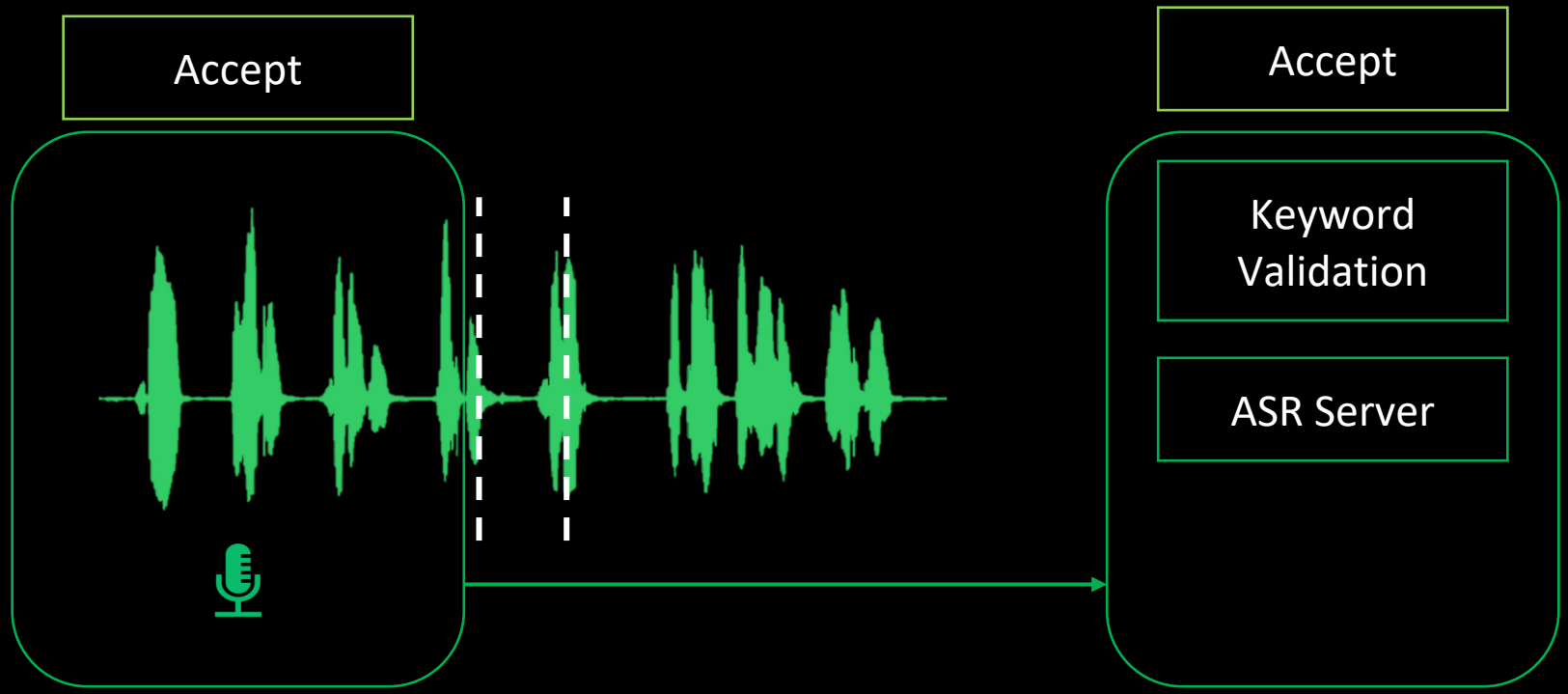
Keyword detection: cascade system



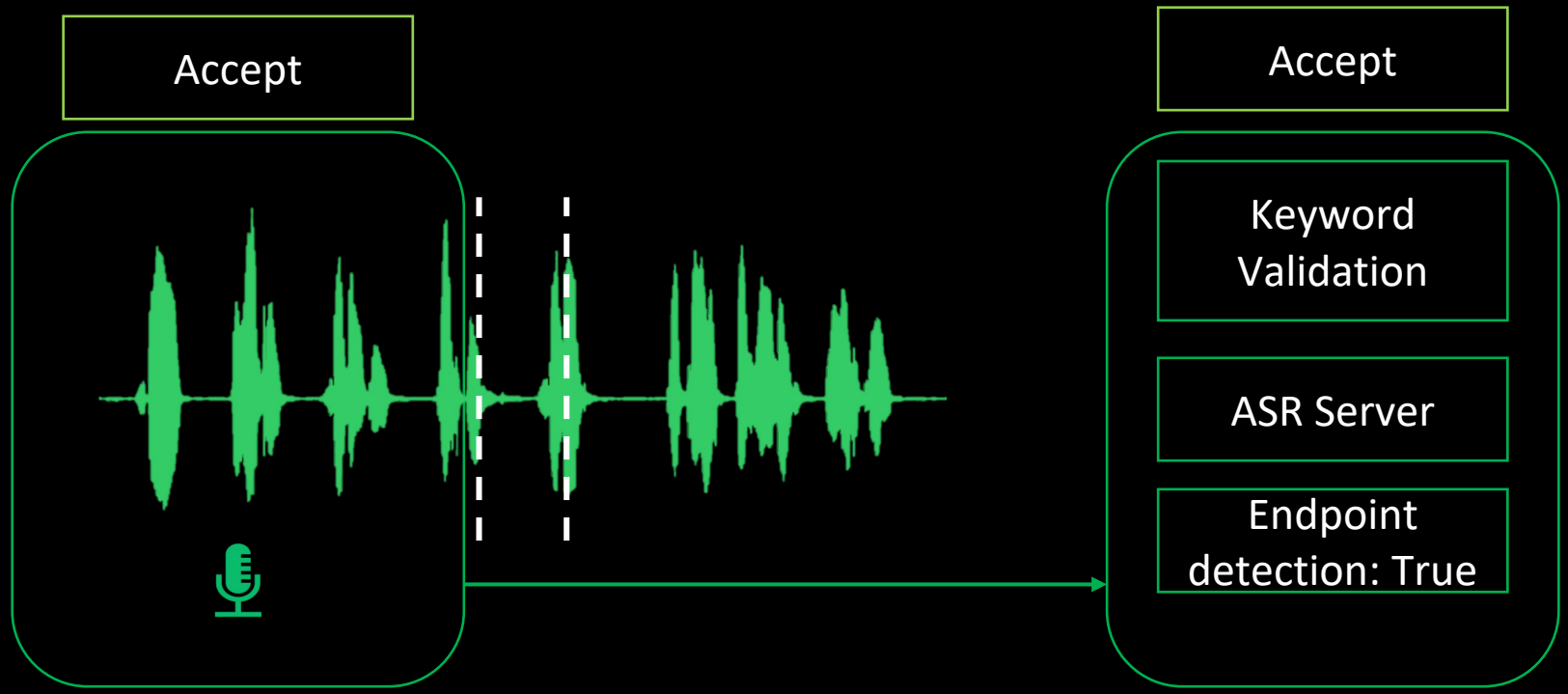
Keyword detection: cascade system



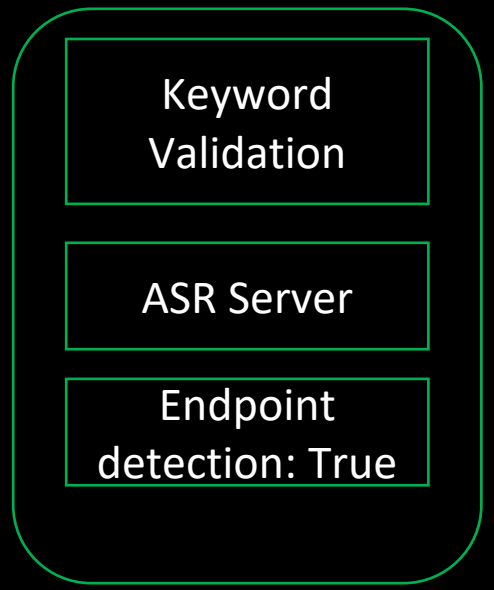
Keyword detection: cascade system



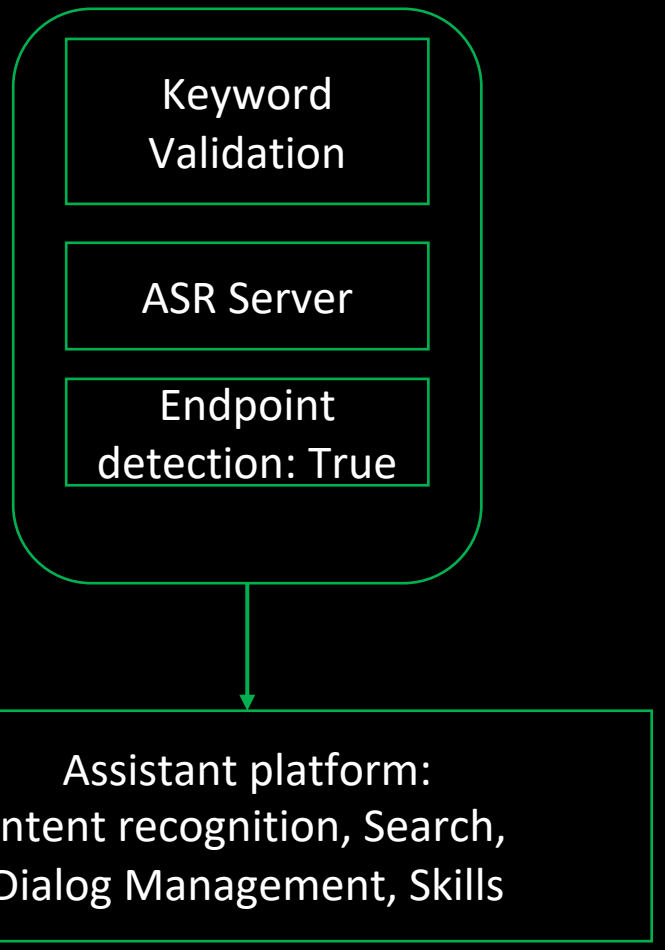
Keyword detection: cascade system



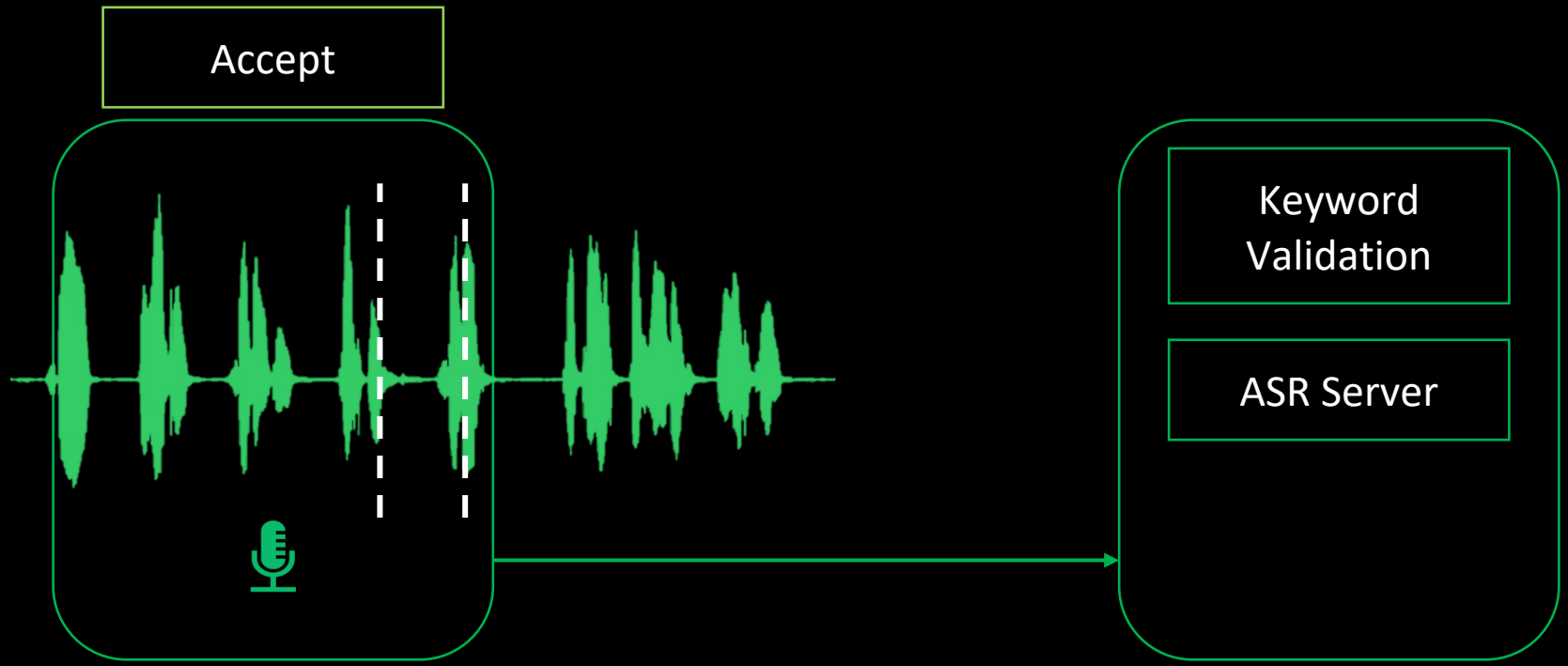
Keyword detection: cascade system



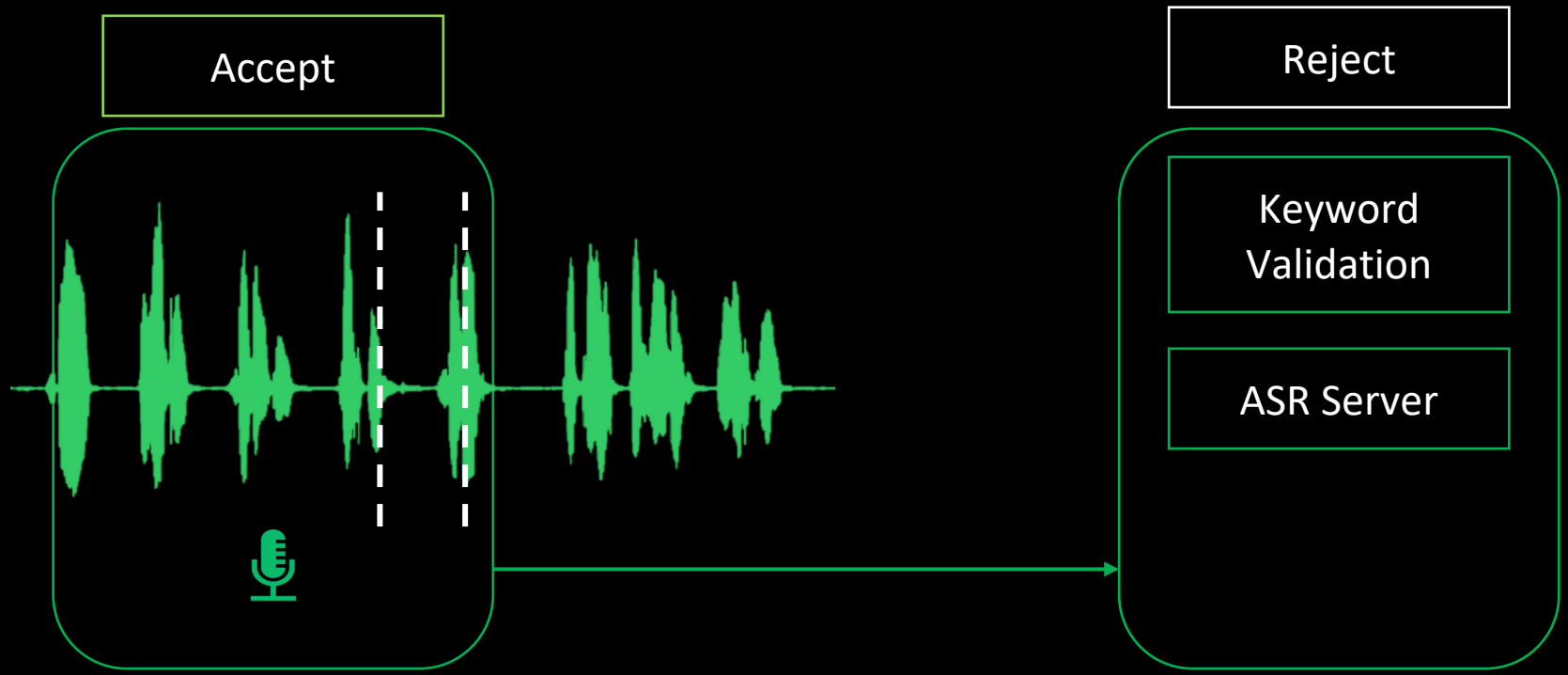
Keyword detection: cascade system



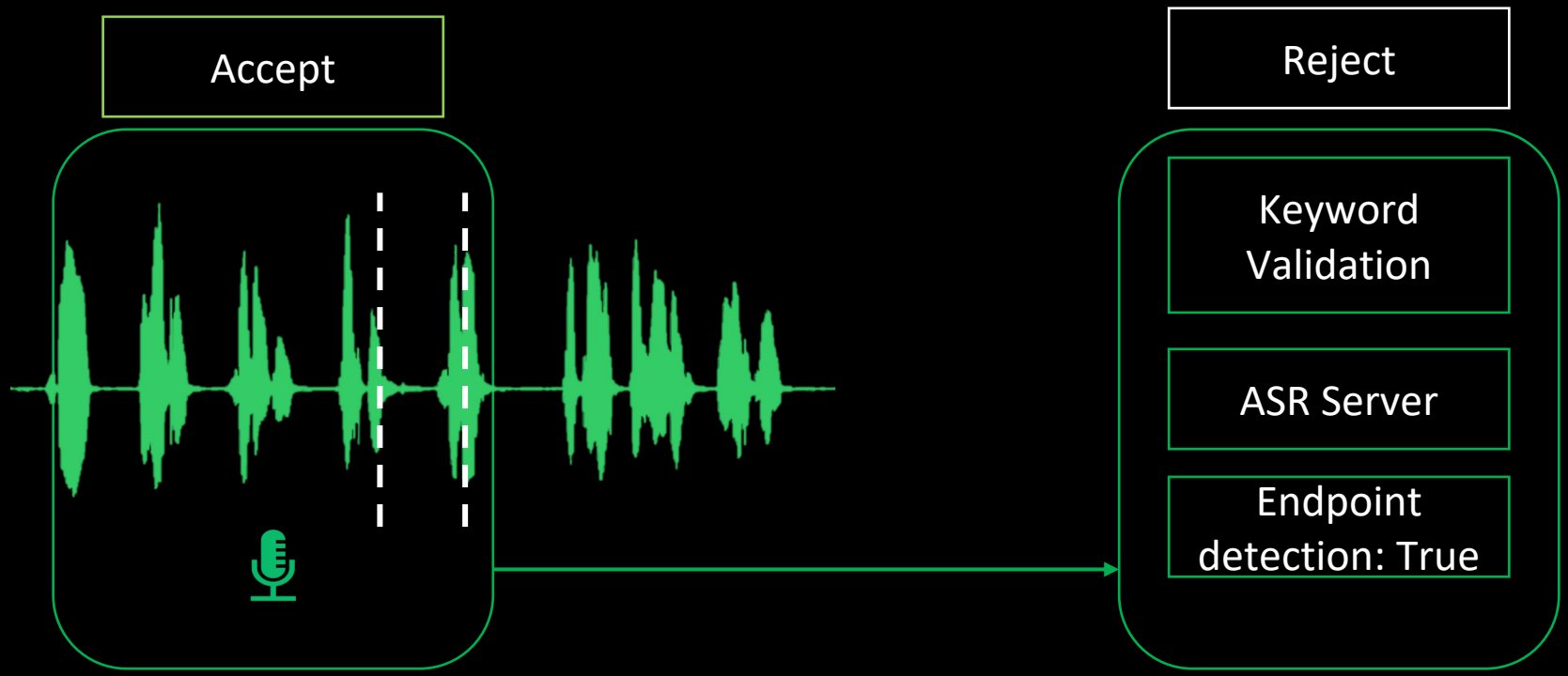
Keyword detection: cascade system



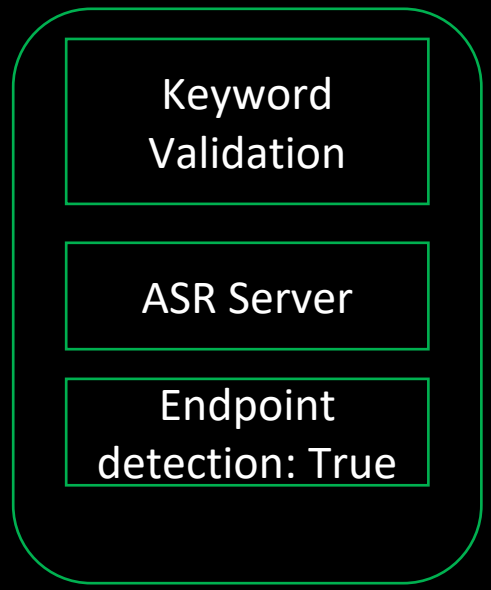
Keyword detection: cascade system



Keyword detection: cascade system



Keyword detection: cascade system



Keyword detection: cascade system

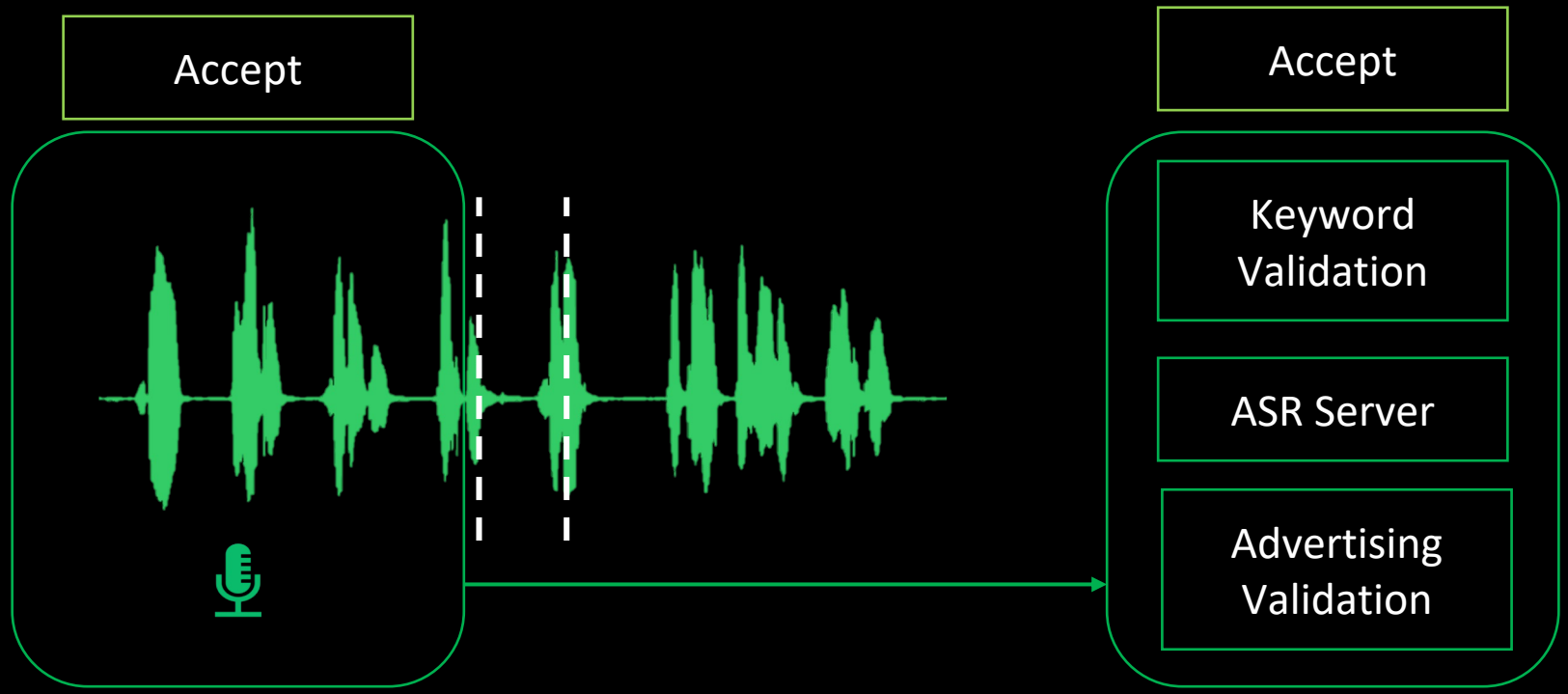
Преимущества:
- Высокая точность
- Гибкость системы

Недостатки:
- Дополнительные вычислительные расходы

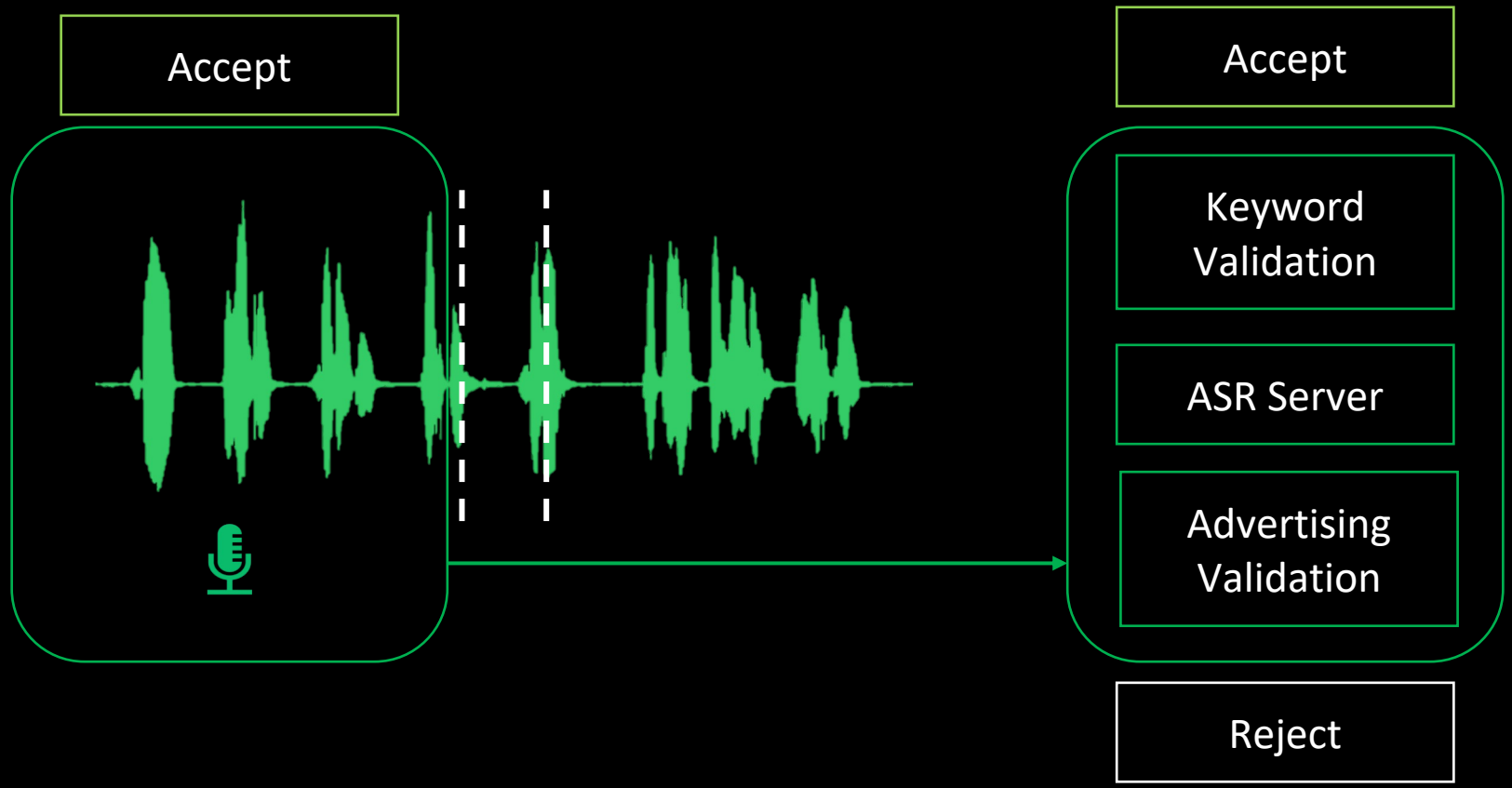
Keyword detection: complex cascade system

Что может быть еще?
- Валидация рекламы

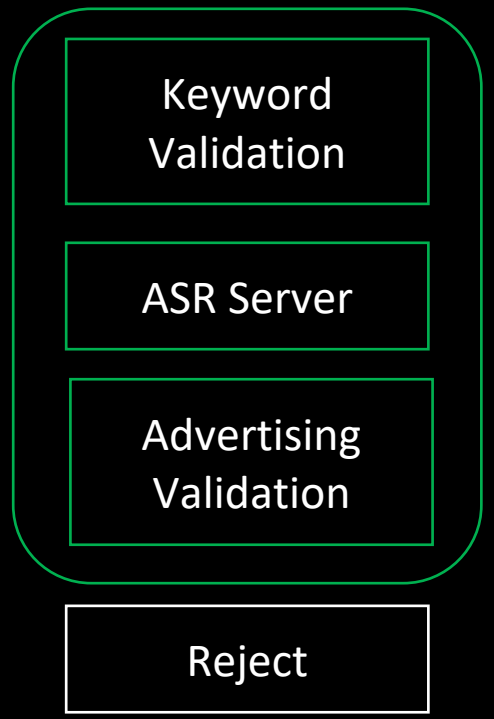
Keyword detection: complex cascade system



Keyword detection: complex cascade system

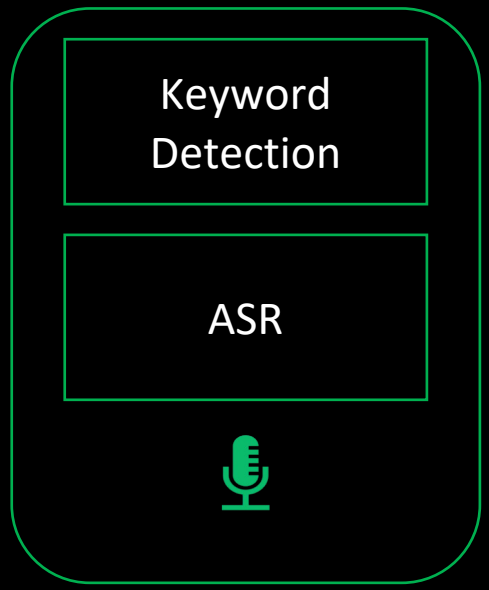


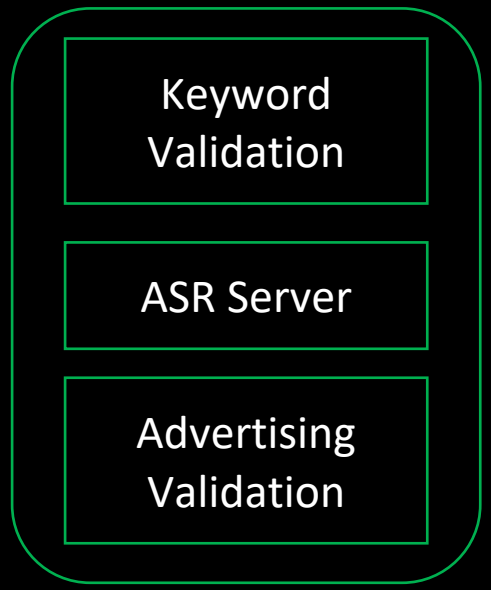
Keyword detection: complex cascade system



Keyword detection: complex cascade system

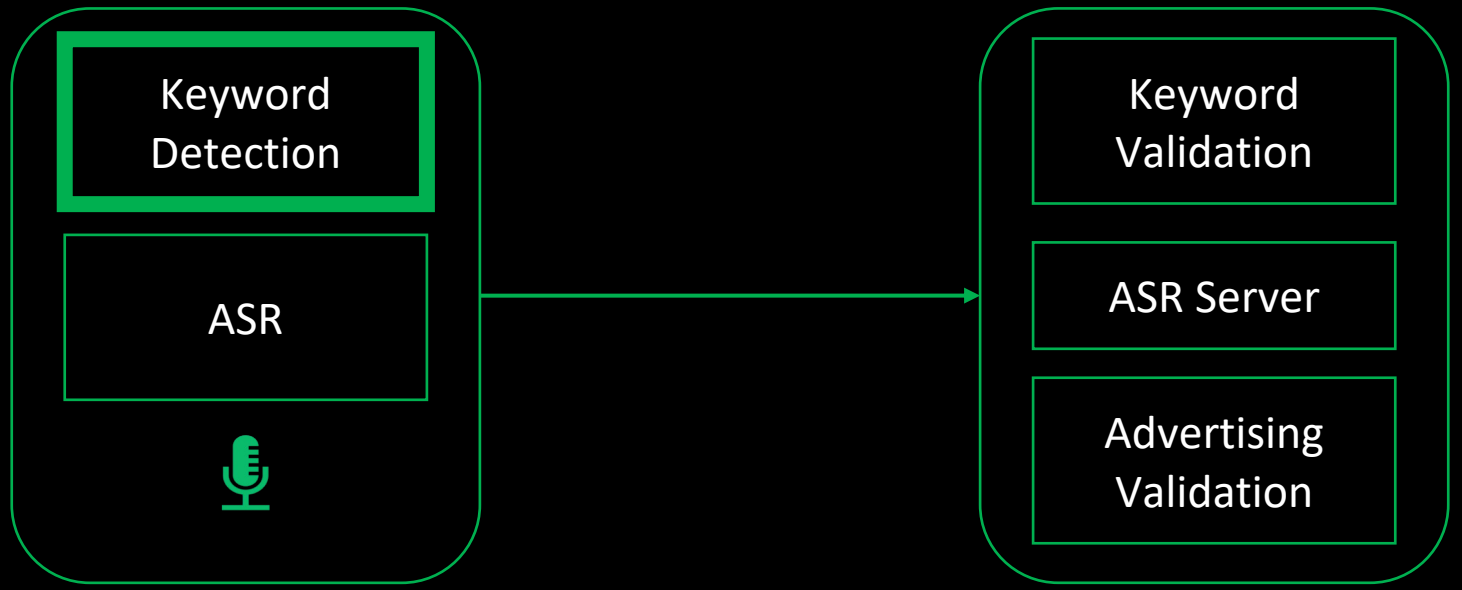
Что может быть еще?
- Быстрые команды, habr





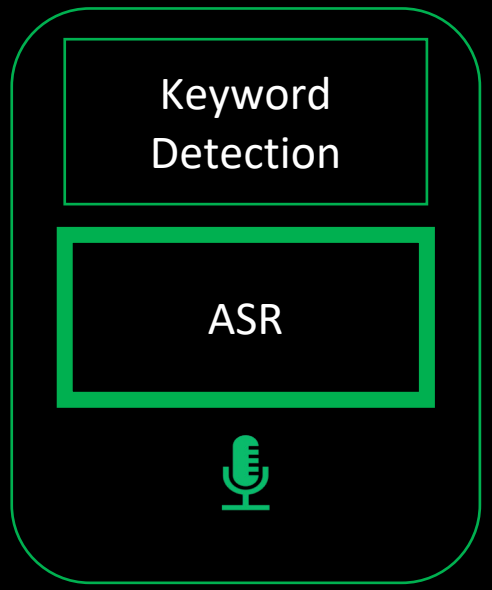
Keyword detection: complex cascade system

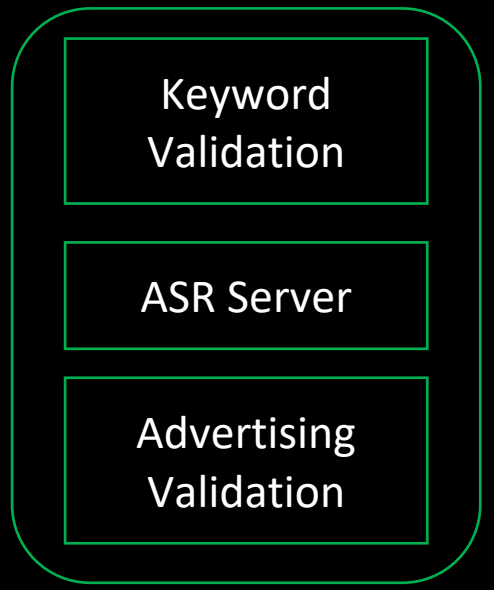
Что может быть еще?
- Быстрые команды, habr



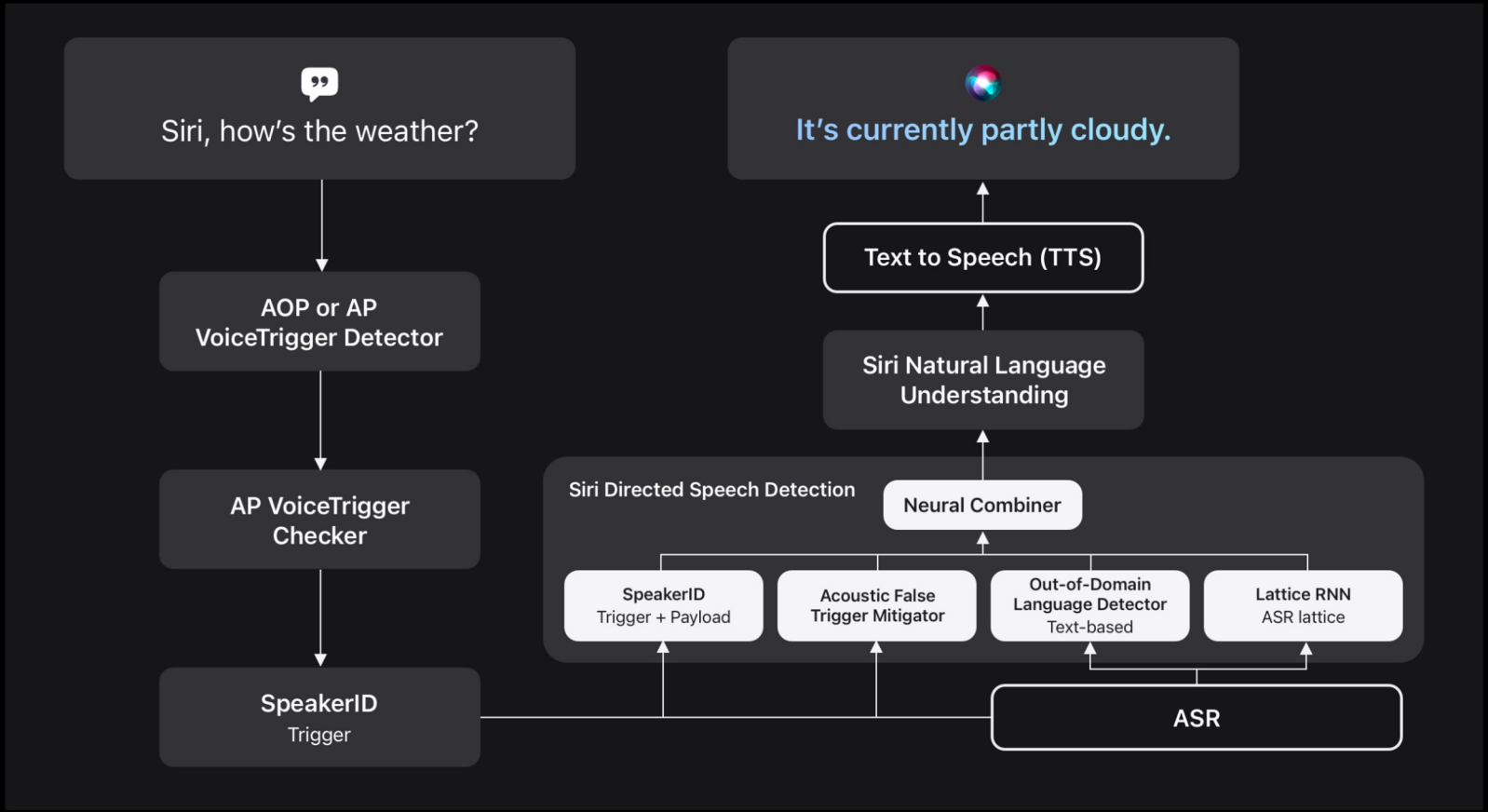
Keyword detection: complex cascade system

Что может быть еще?
- Быстрые команды, habr





Keyword detection: complex cascade system



Voice Trigger System for Siri: Apple, 2023

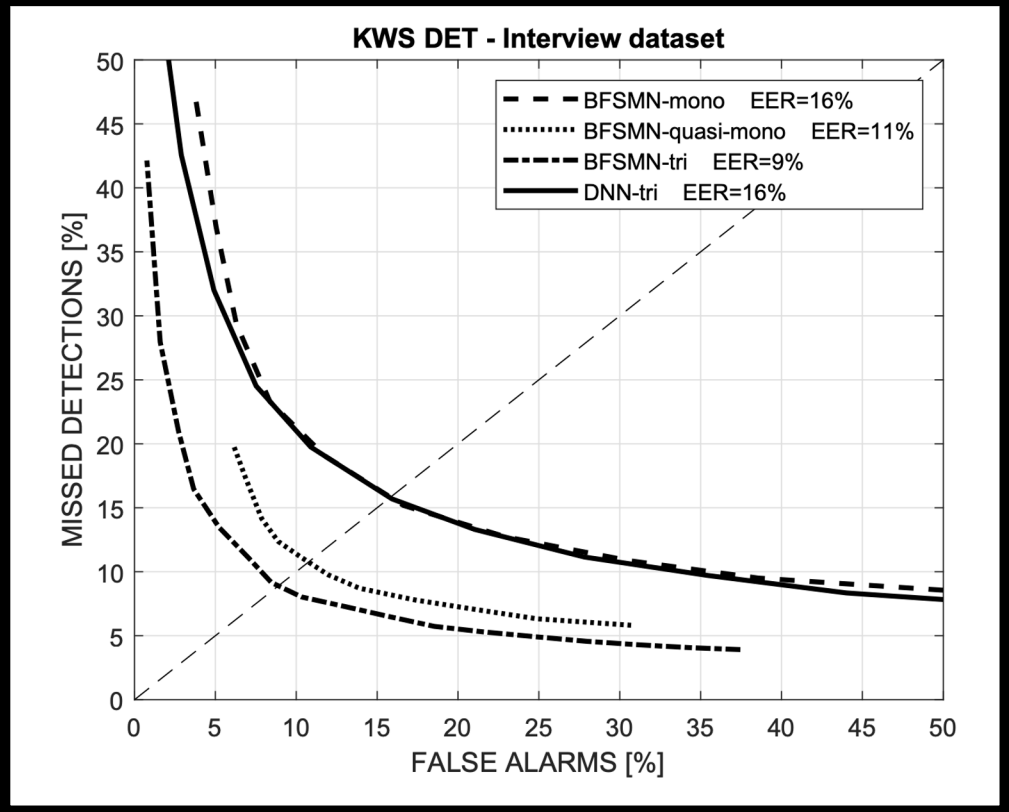
Keyword detection: метрики

FAR: False (Activation / Accepts / Alarm) Rate
- FAR = FP / (FP + TN)

FRR: False Rejects Rate
- FRR = FN / (FN + TP)

FAh: False Accepts per hour
- Количество ложных срабатываний в час

Detection Error Tradeoff



User perceived latency

Скорость отклика
- Быстрое выполнение команд, без задержек и пауз

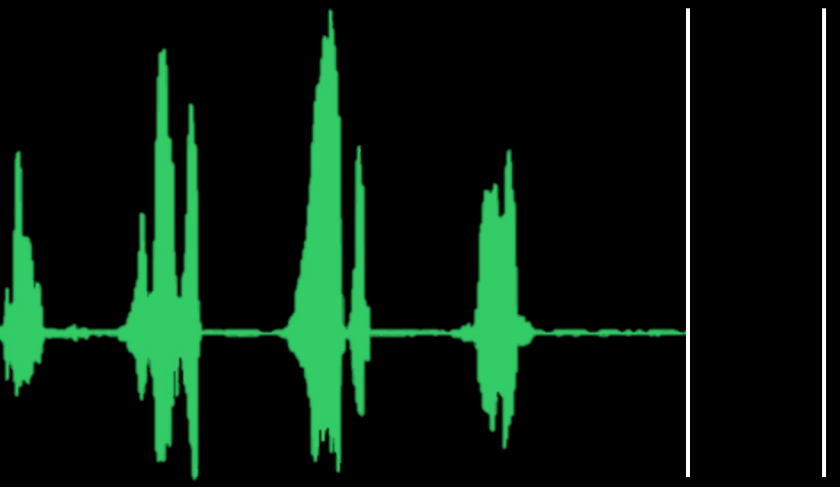
User perceived latency – метрика, отражающая воспринимаемое пользователем время задержки между действием и реакцией системы.

В нашем случае можем выделить два примера:
- Произношение активационной фразы <-> активация колонки
- Произношение голосового запроса <-> получение ответа

Endpoint detection

Основная задача в рамках Voice Assistant Pipeline
- Определение конца голосового запроса пользователя для завершения приема аудиосигнала и начала подготовки ответа

Endpoint detection



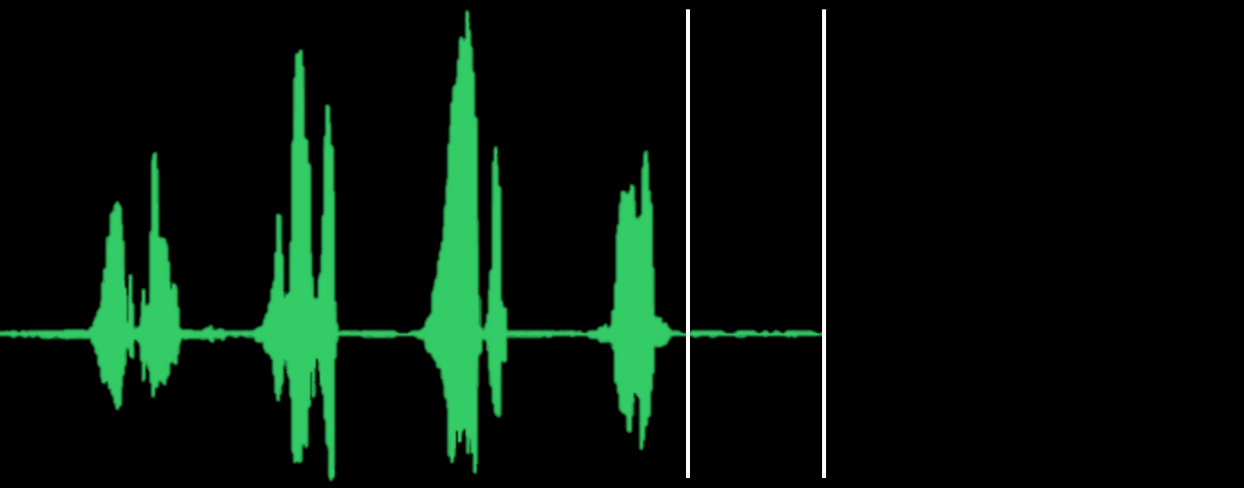
0 < x < 1

100ms

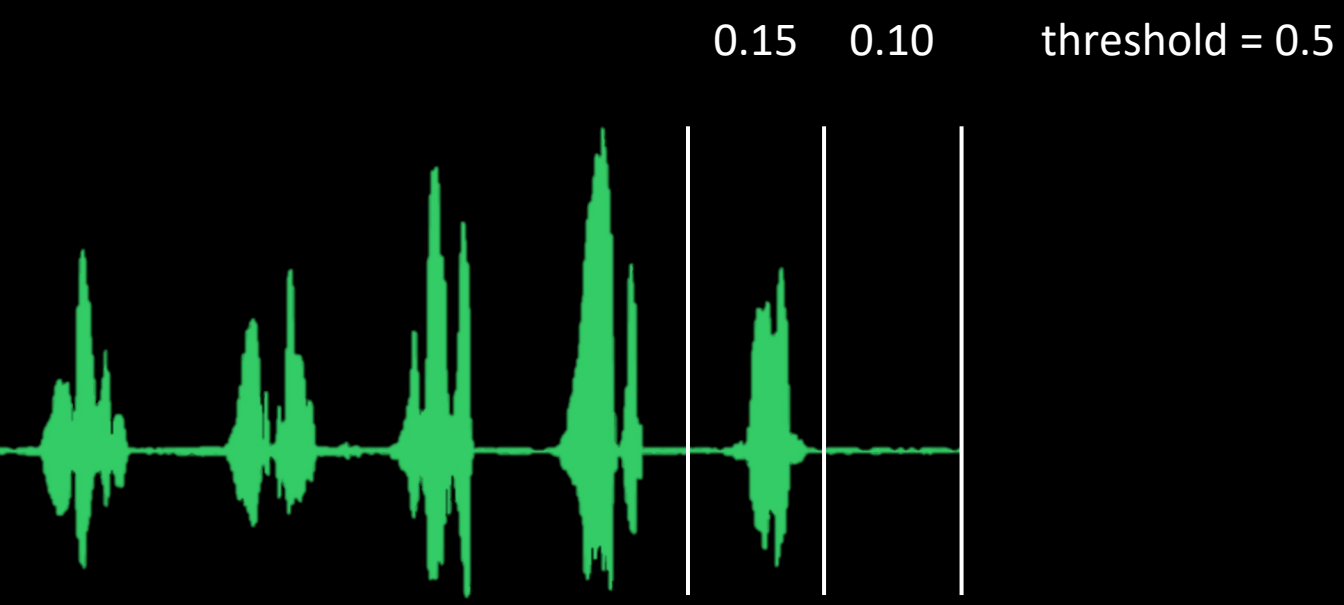
threshold

1 – endpoint: true
0 – endpoint: false

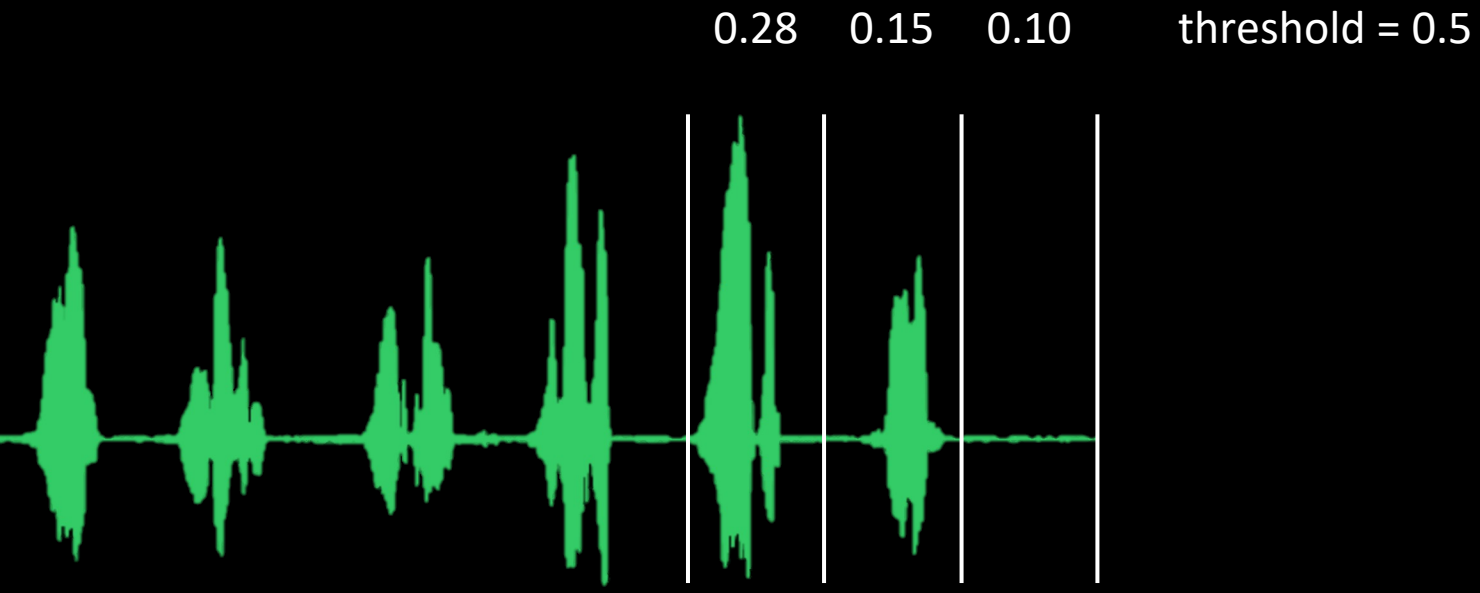
Endpoint detection



Endpoint detection

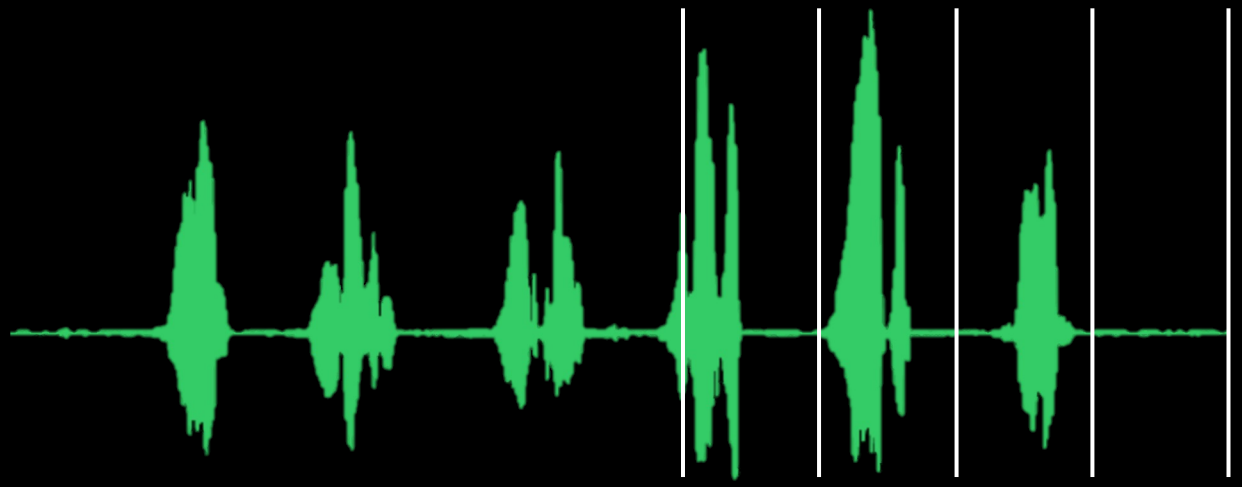


Endpoint detection

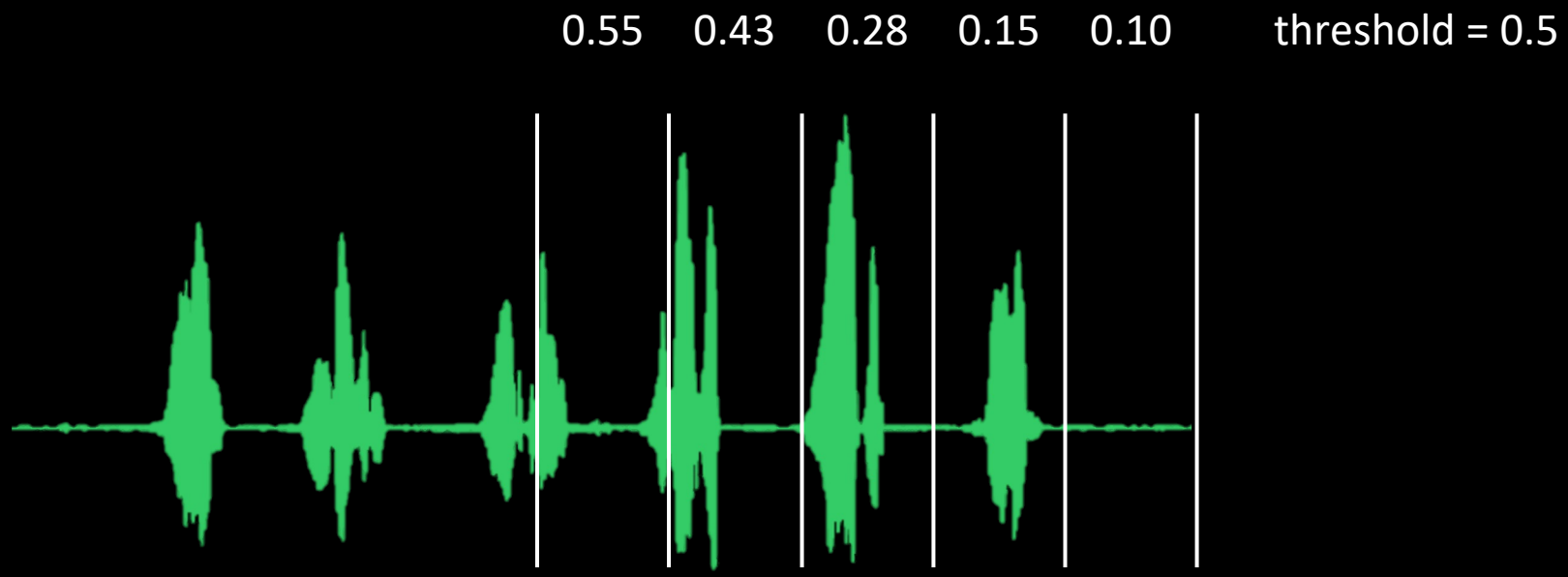


Endpoint detection

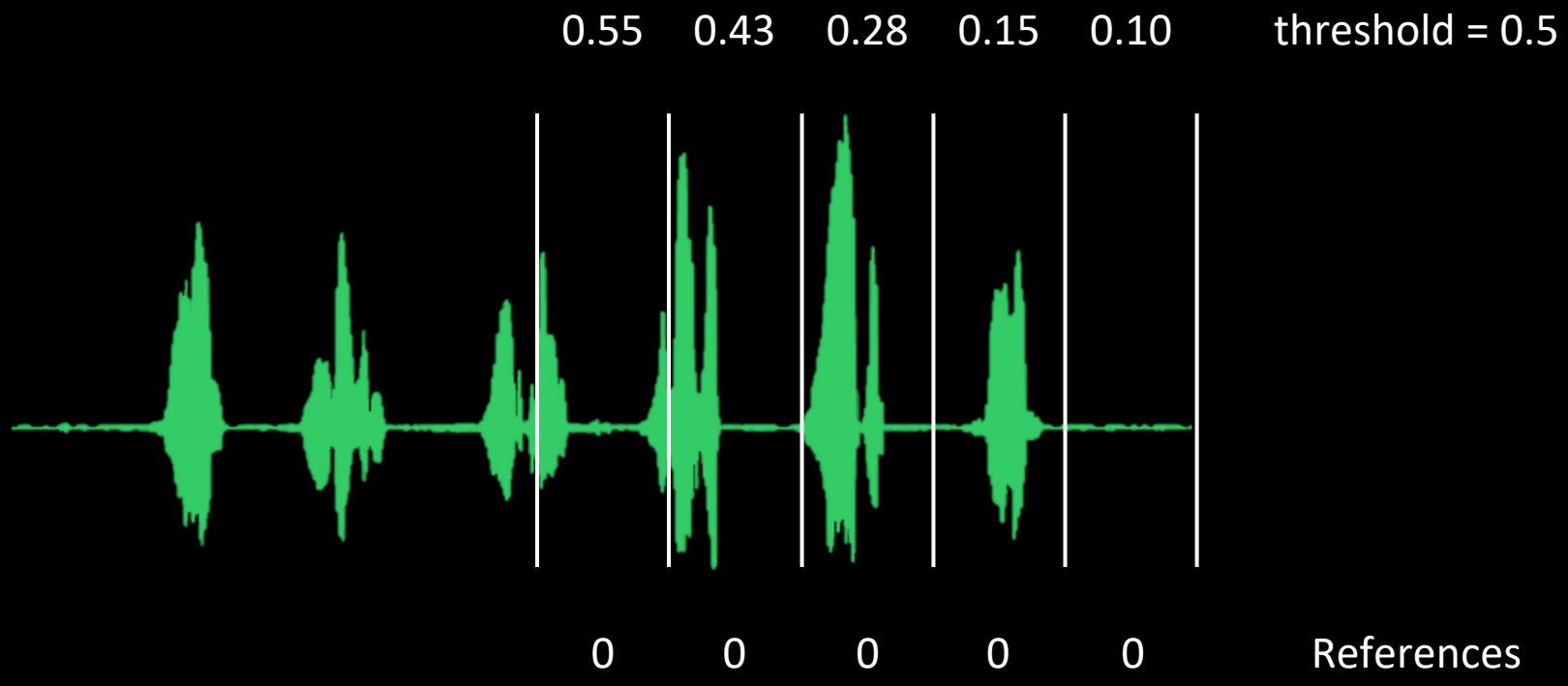
0.43 0.28 0.15 0.10 threshold = 0.5



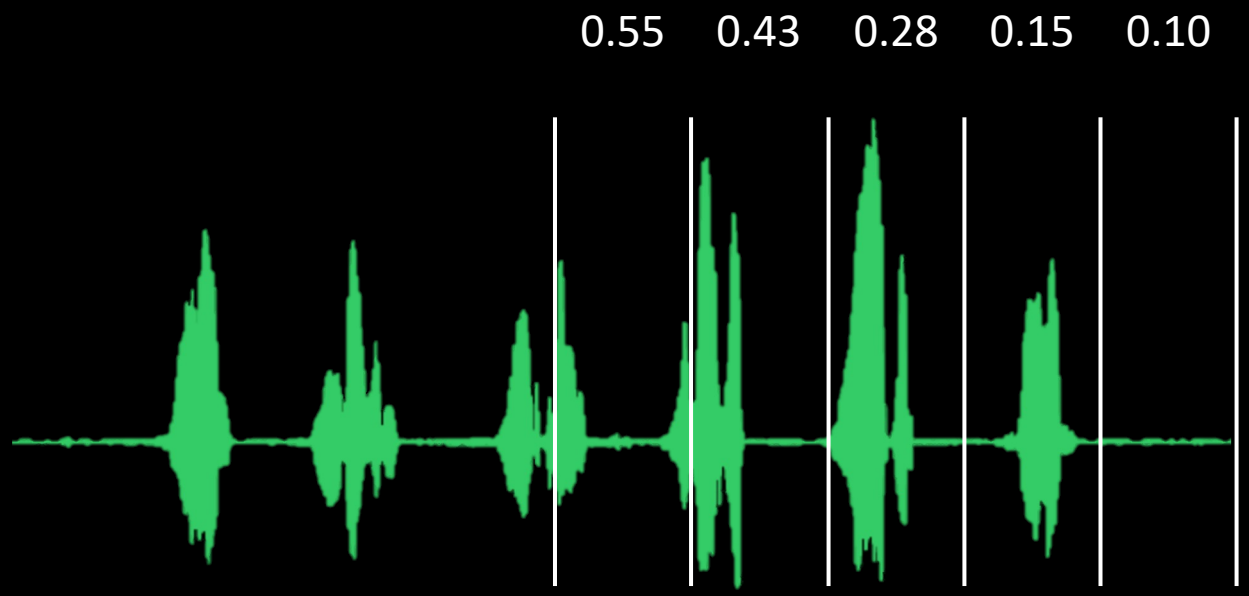
Endpoint detection



Endpoint detection

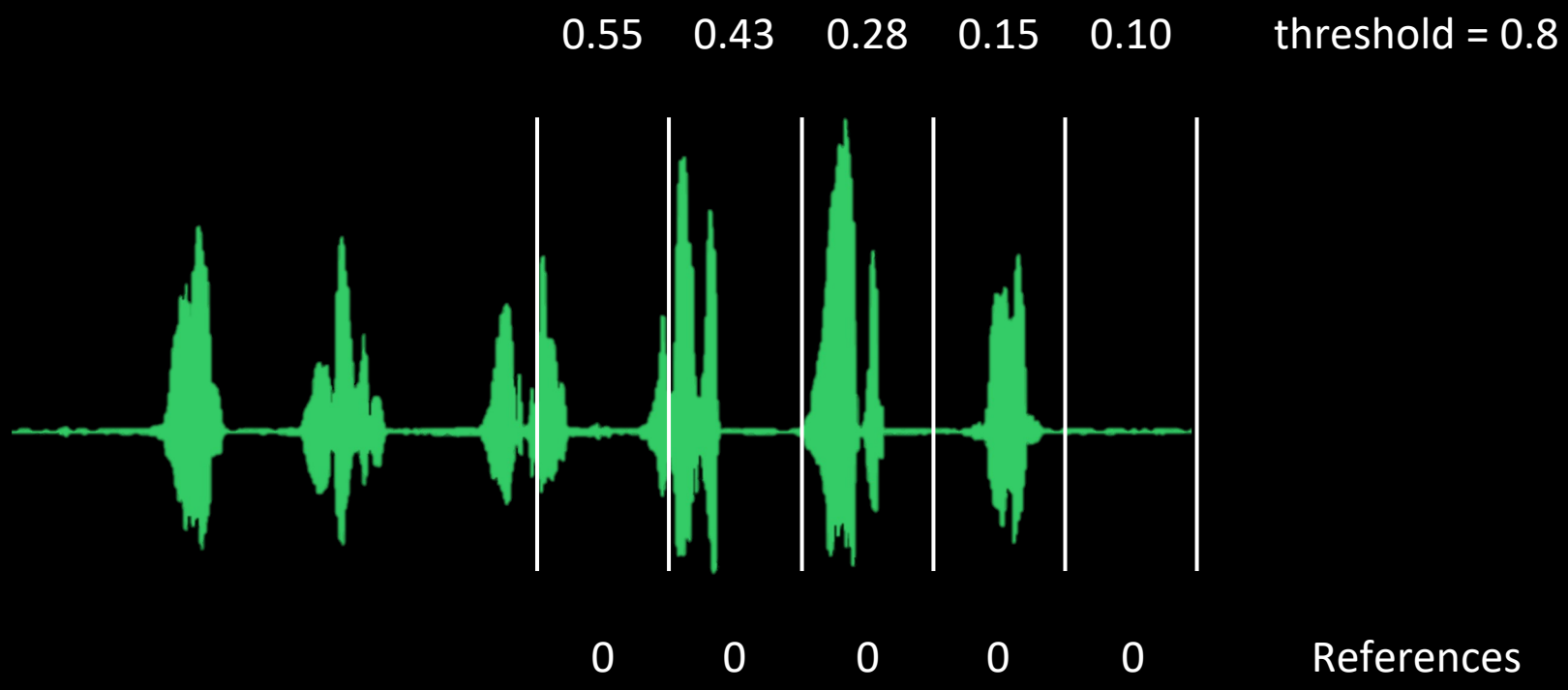


Endpoint detection

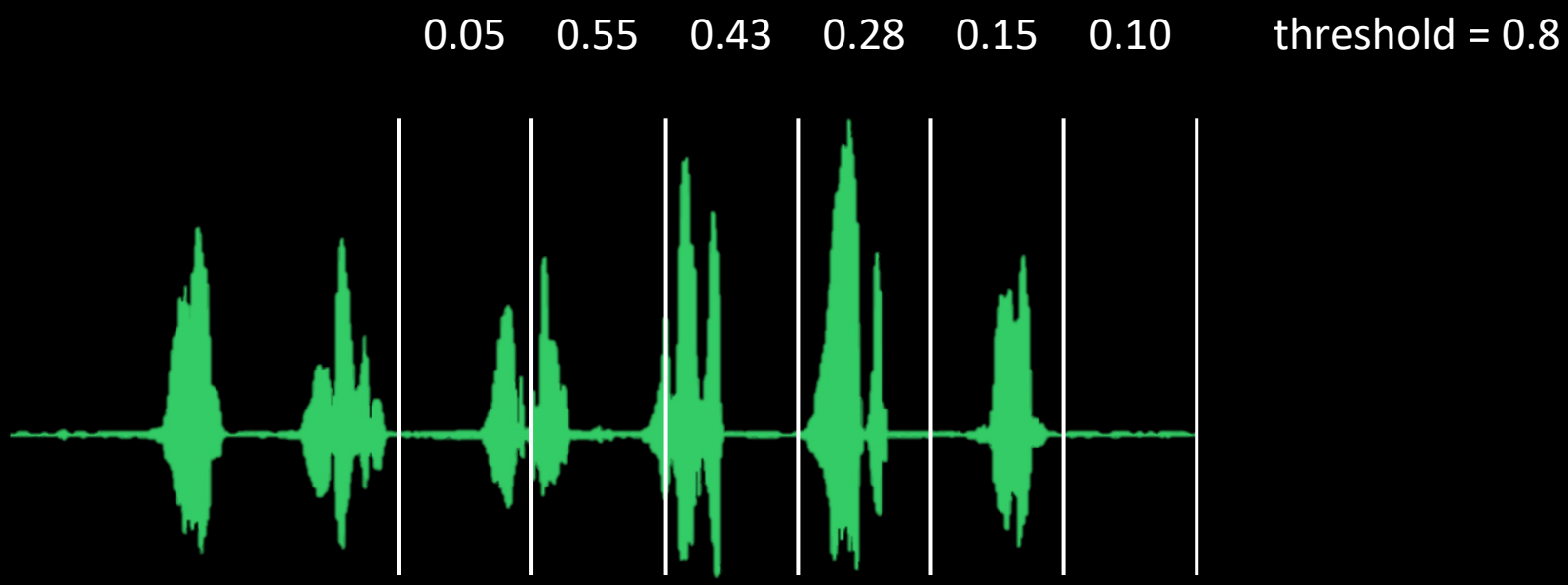


Early activation

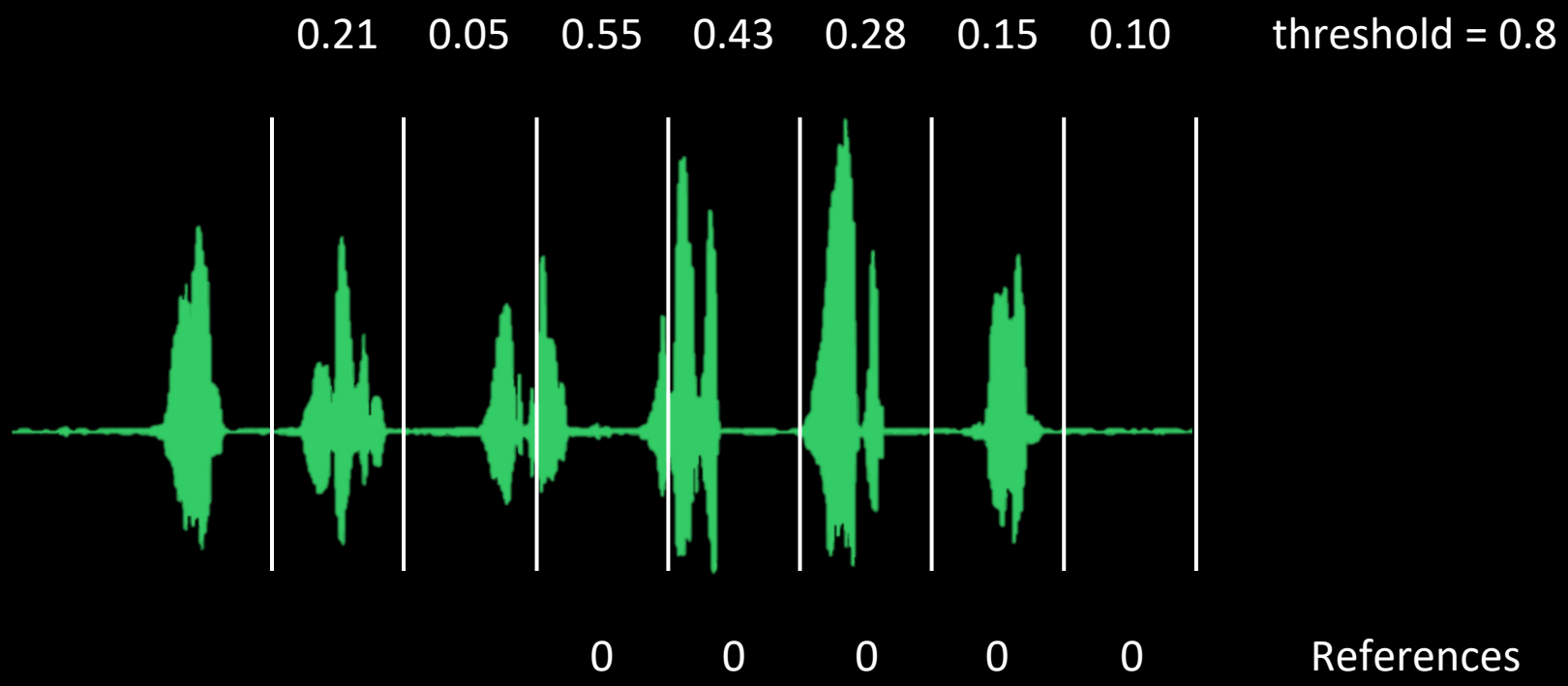
Endpoint detection



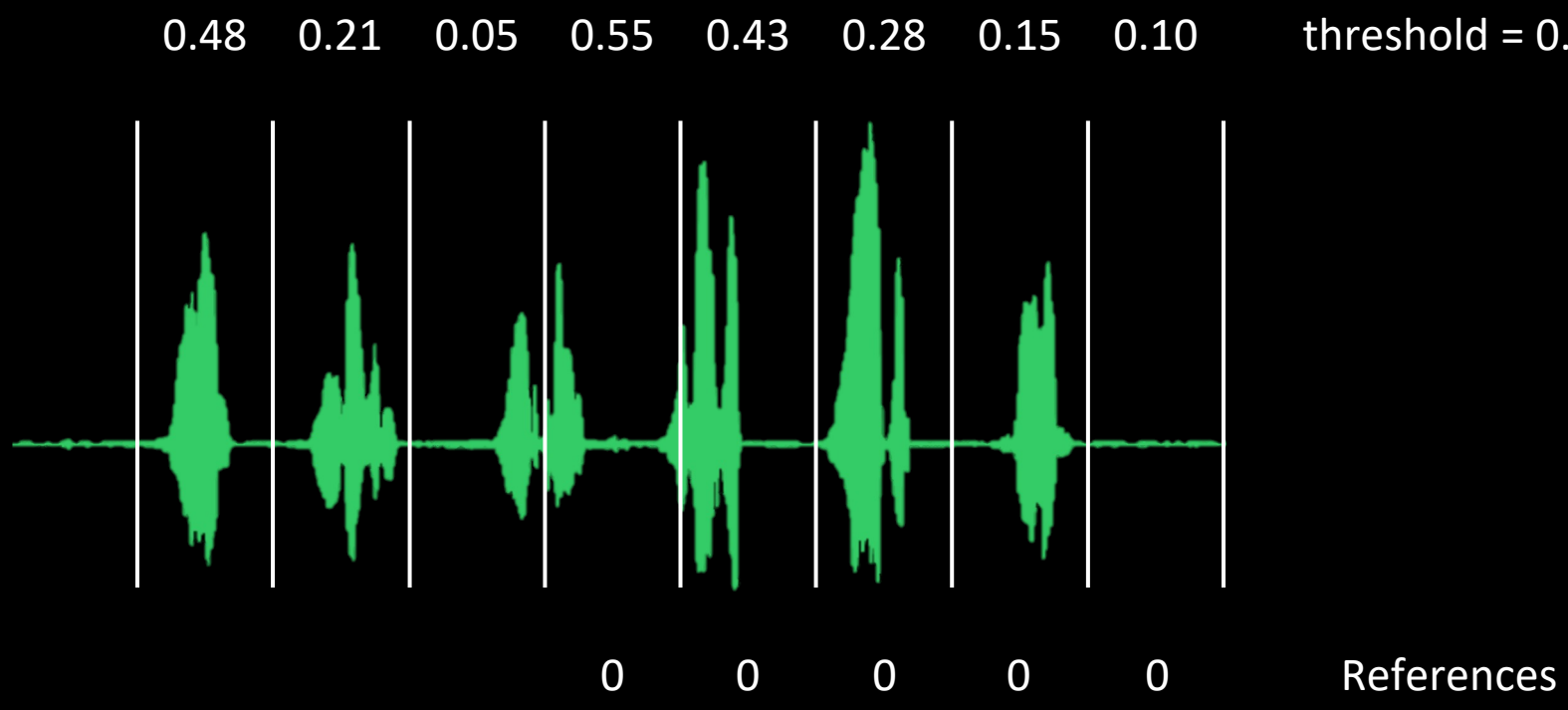
Endpoint detection



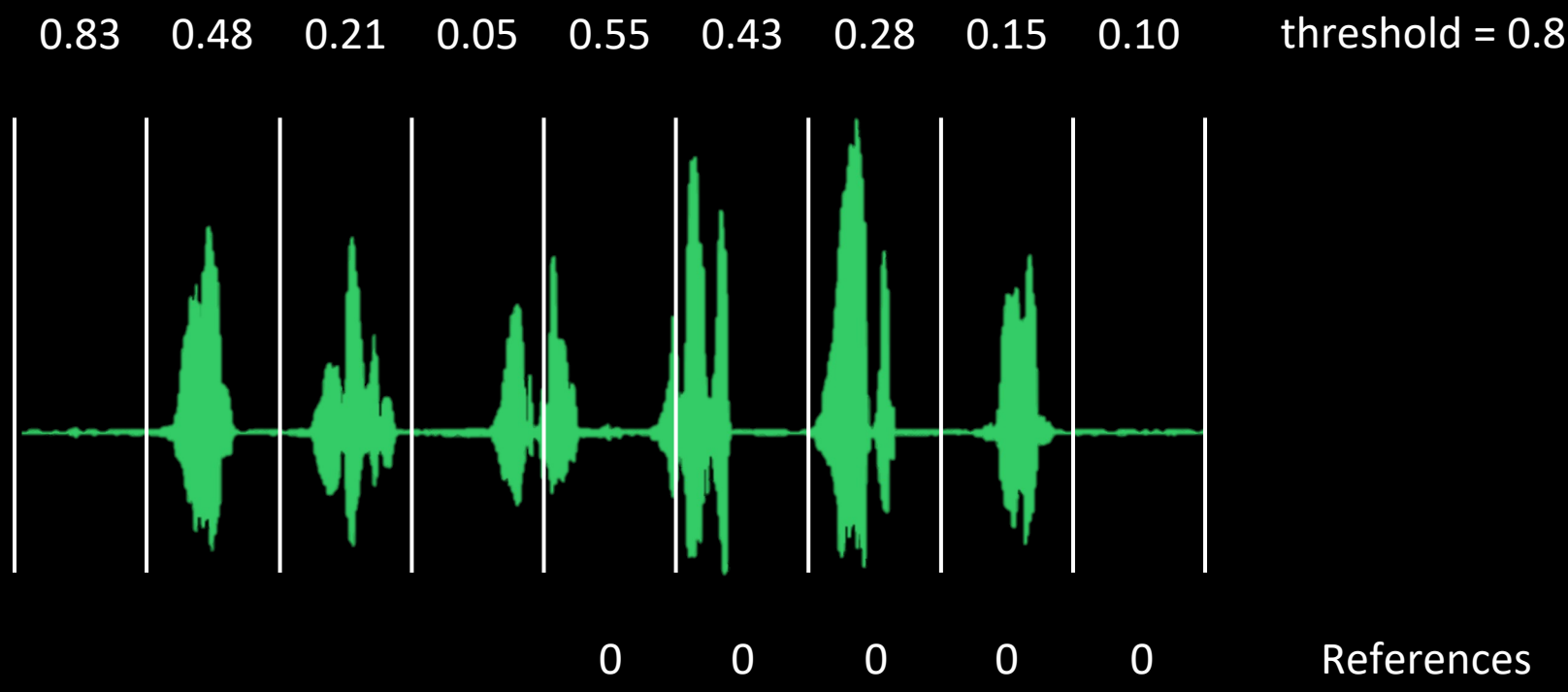
Endpoint detection



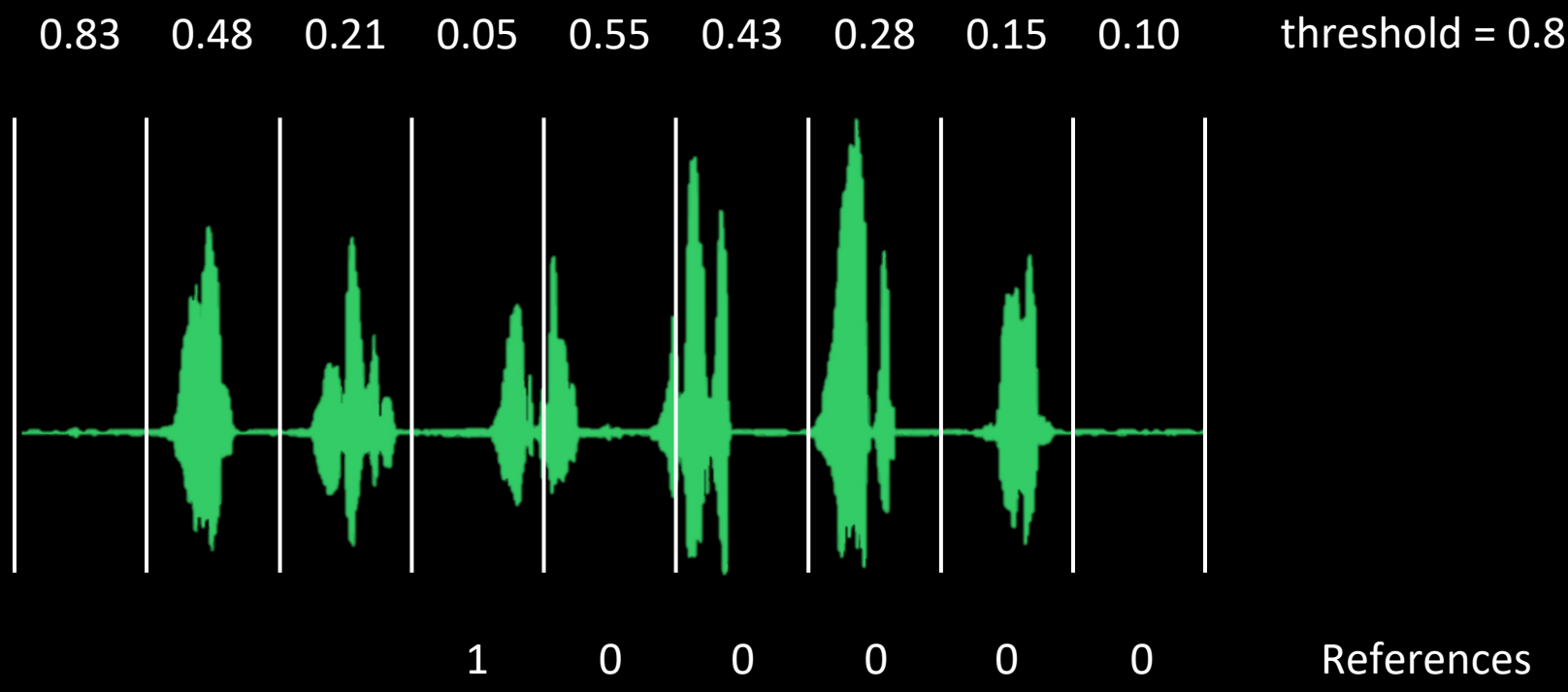
Endpoint detection



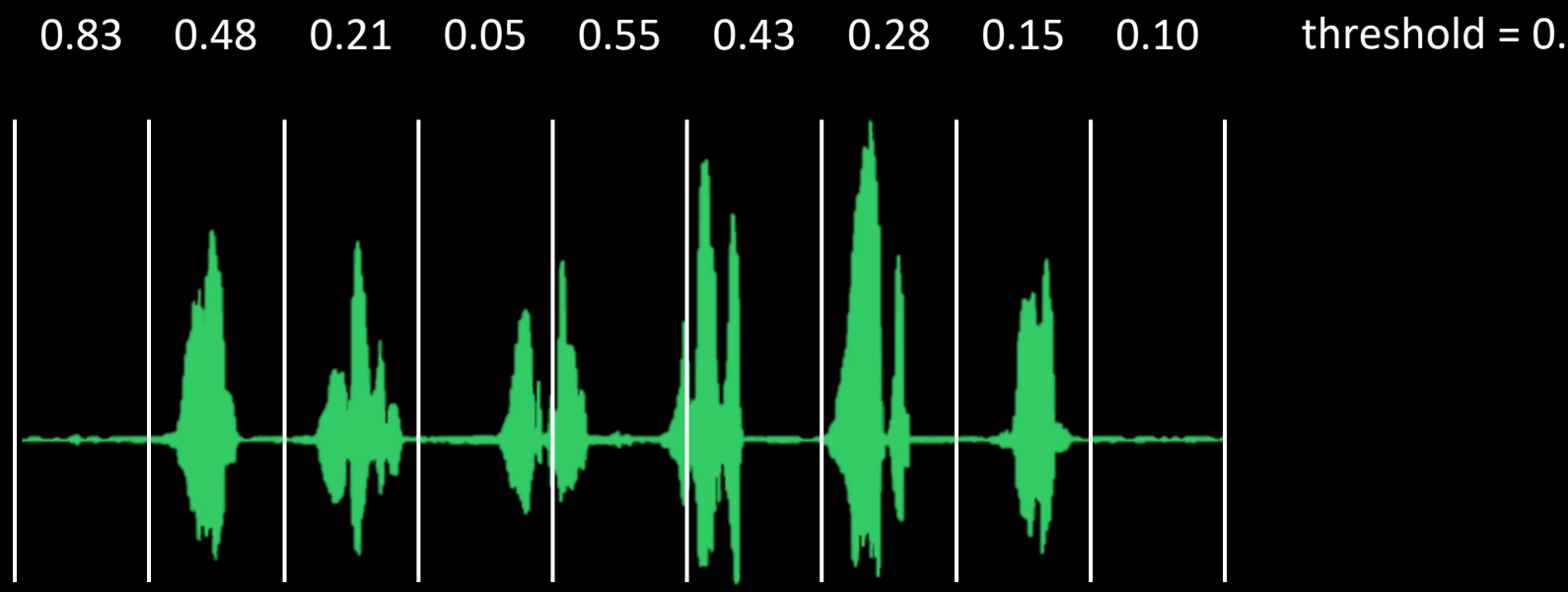
Endpoint detection



Endpoint detection



Endpoint detection



Late activation == latency 300ms

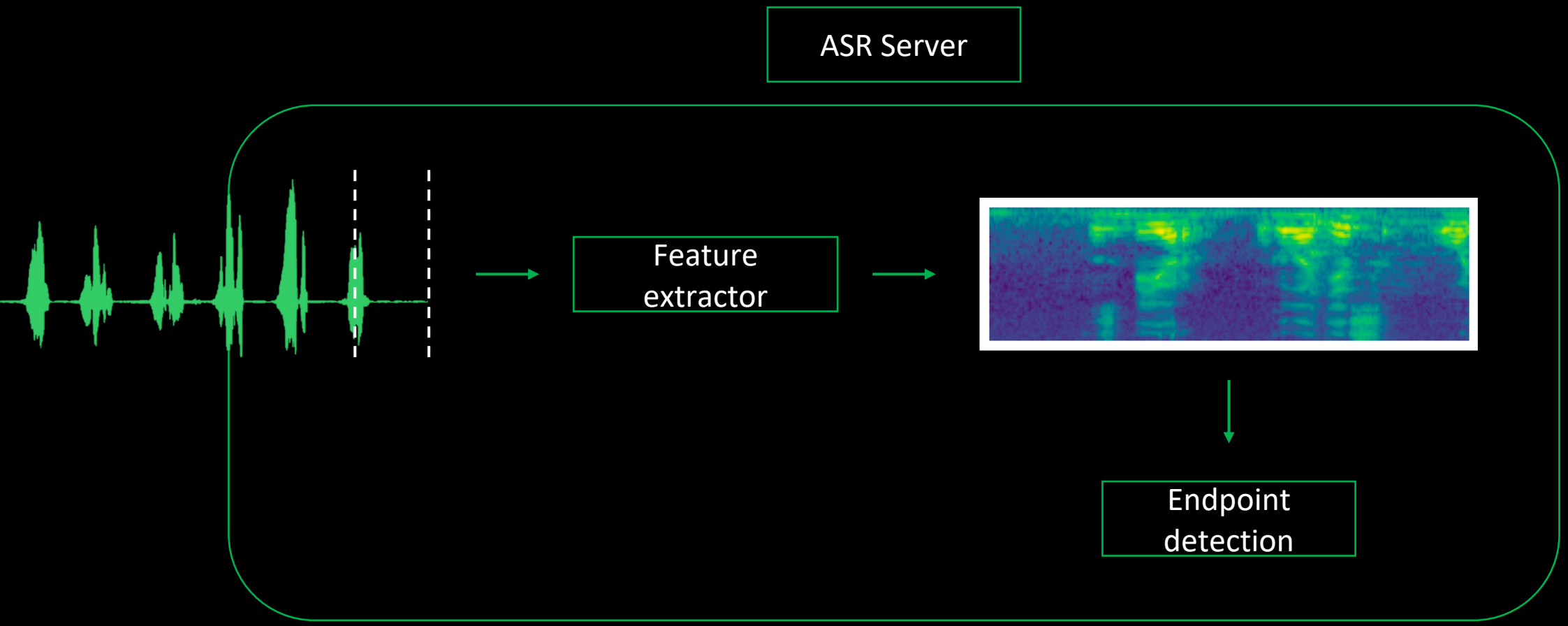
Endpoint detection: метрики

WERR: Word Error Rate Reduction
- WERR = $(WER_{old} - WER_{new}) / WER_{old} * 100\%$

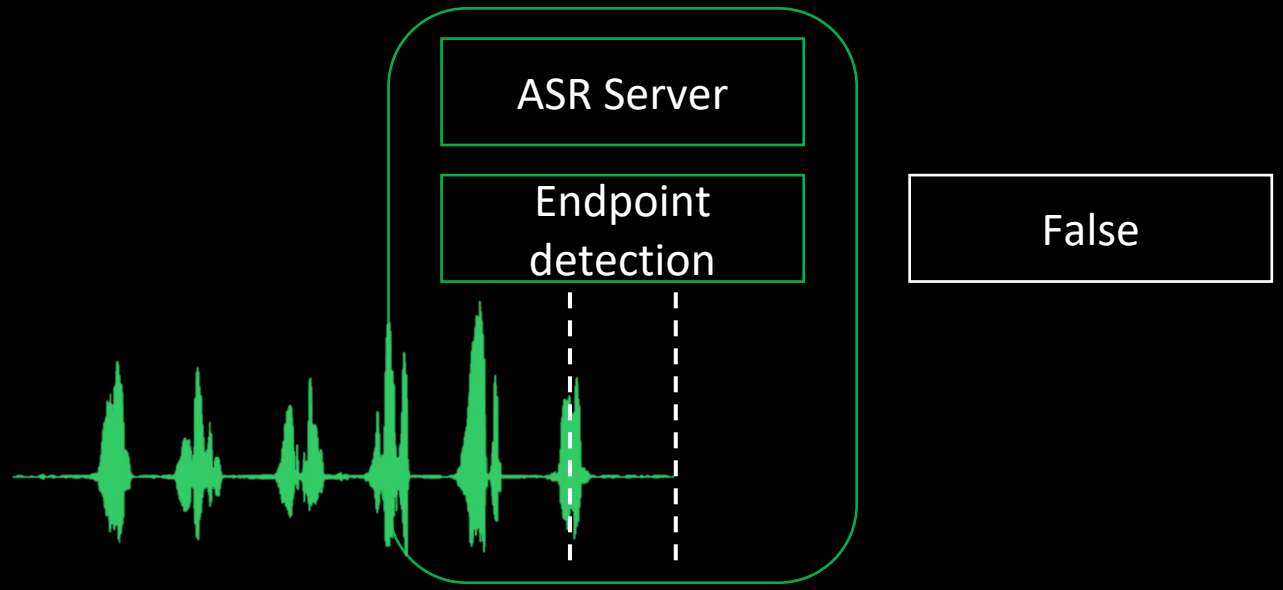
EEPR: Early Endpointing Rate
- Как часто предсказываем конец фразы до последнего слова в запросе

Latency
- Задержка между последним словом в запросе и предсказанием конца фразы
- Часто рассматривают различные квантили

Endpoint detection: audio features

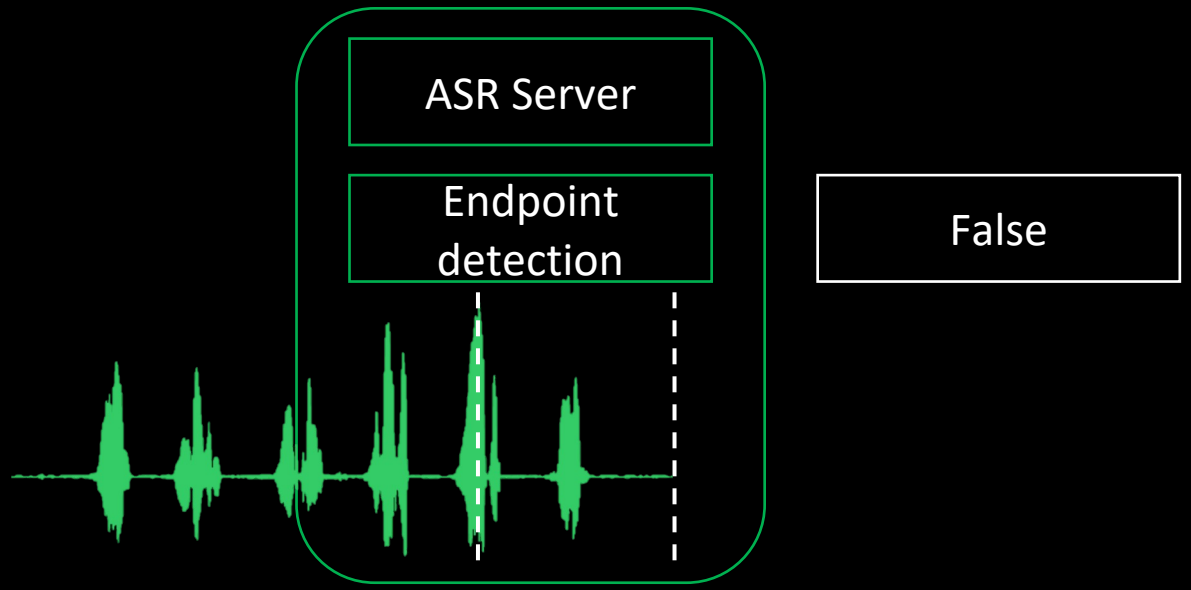


Endpoint detection: audio features



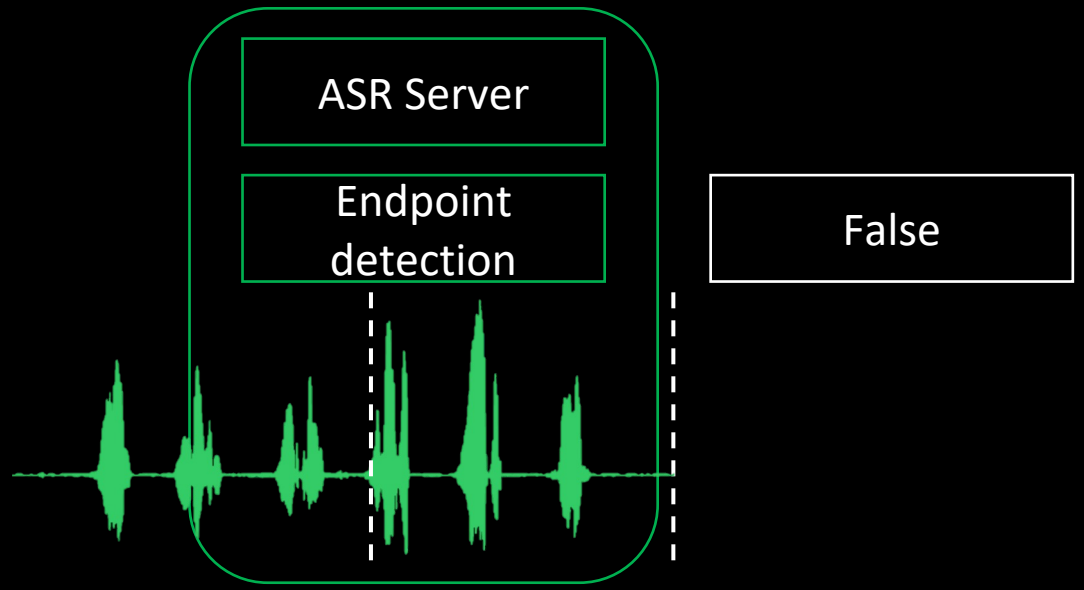
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: audio features



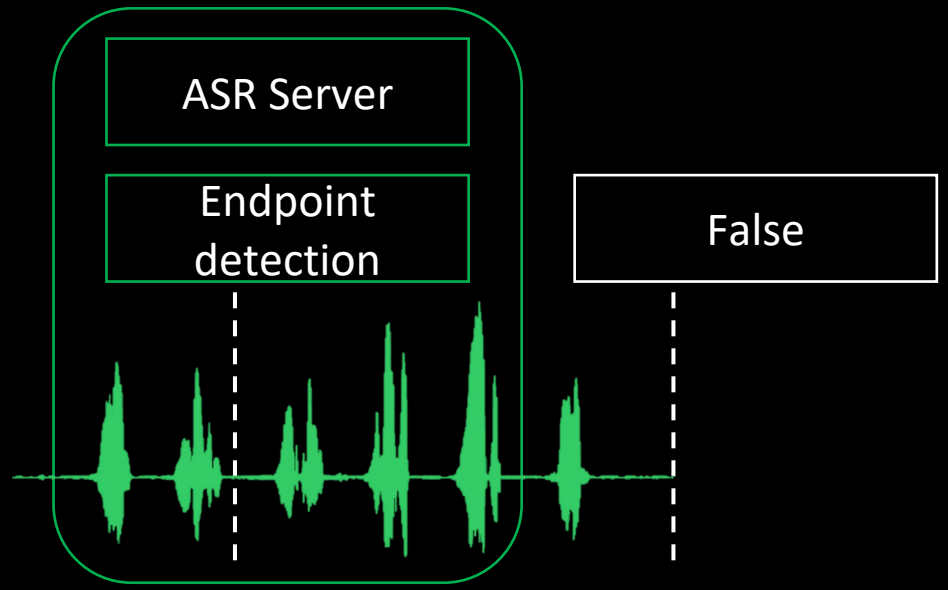
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: audio features



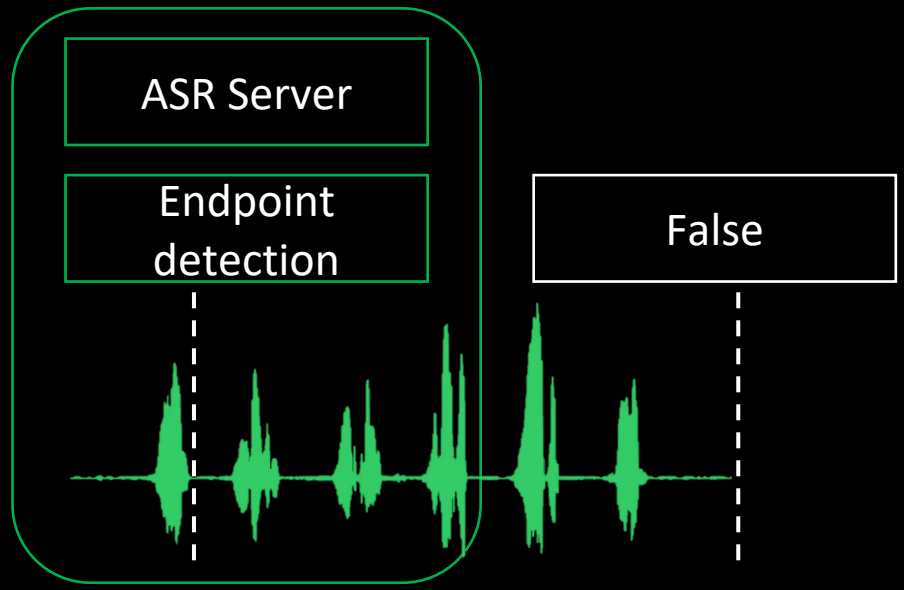
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: audio features



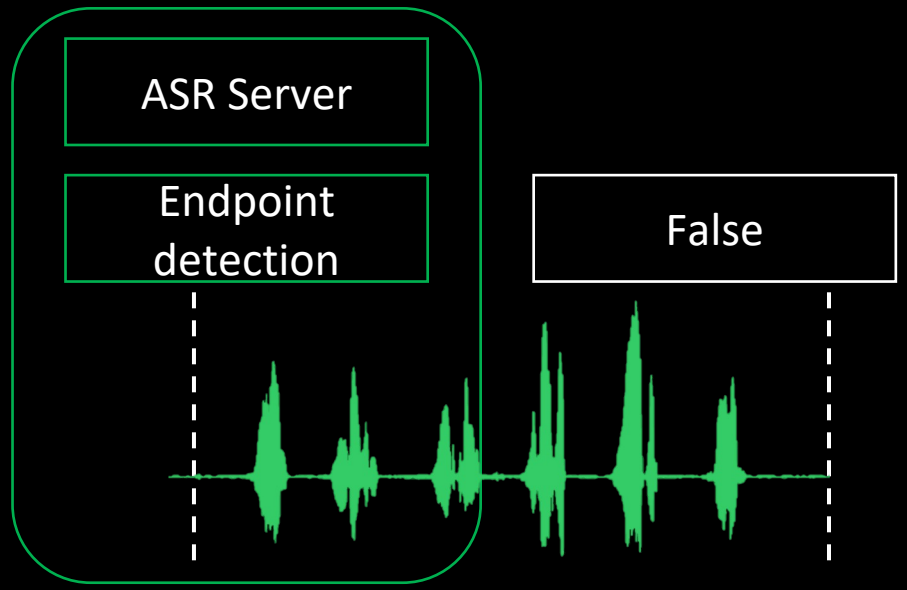
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: audio features



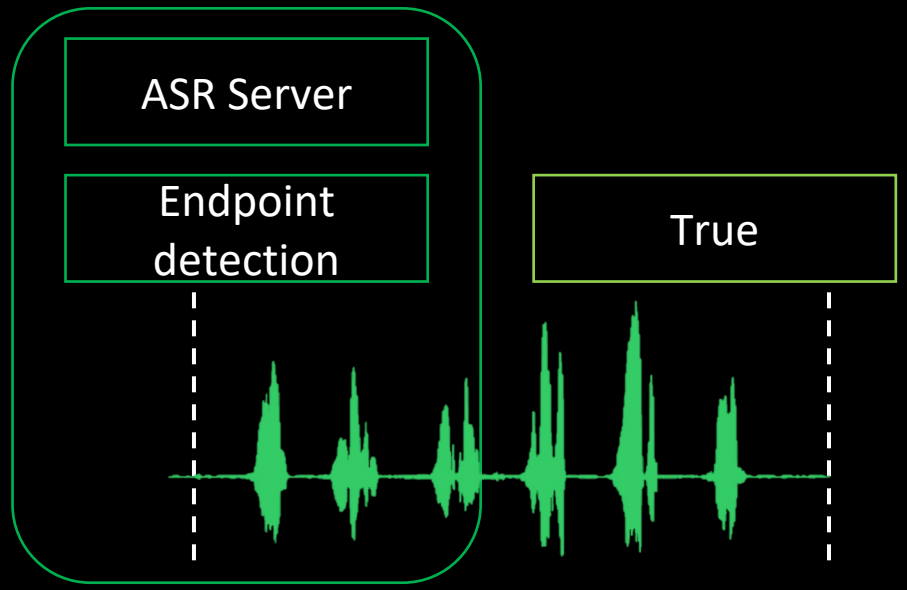
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: audio features



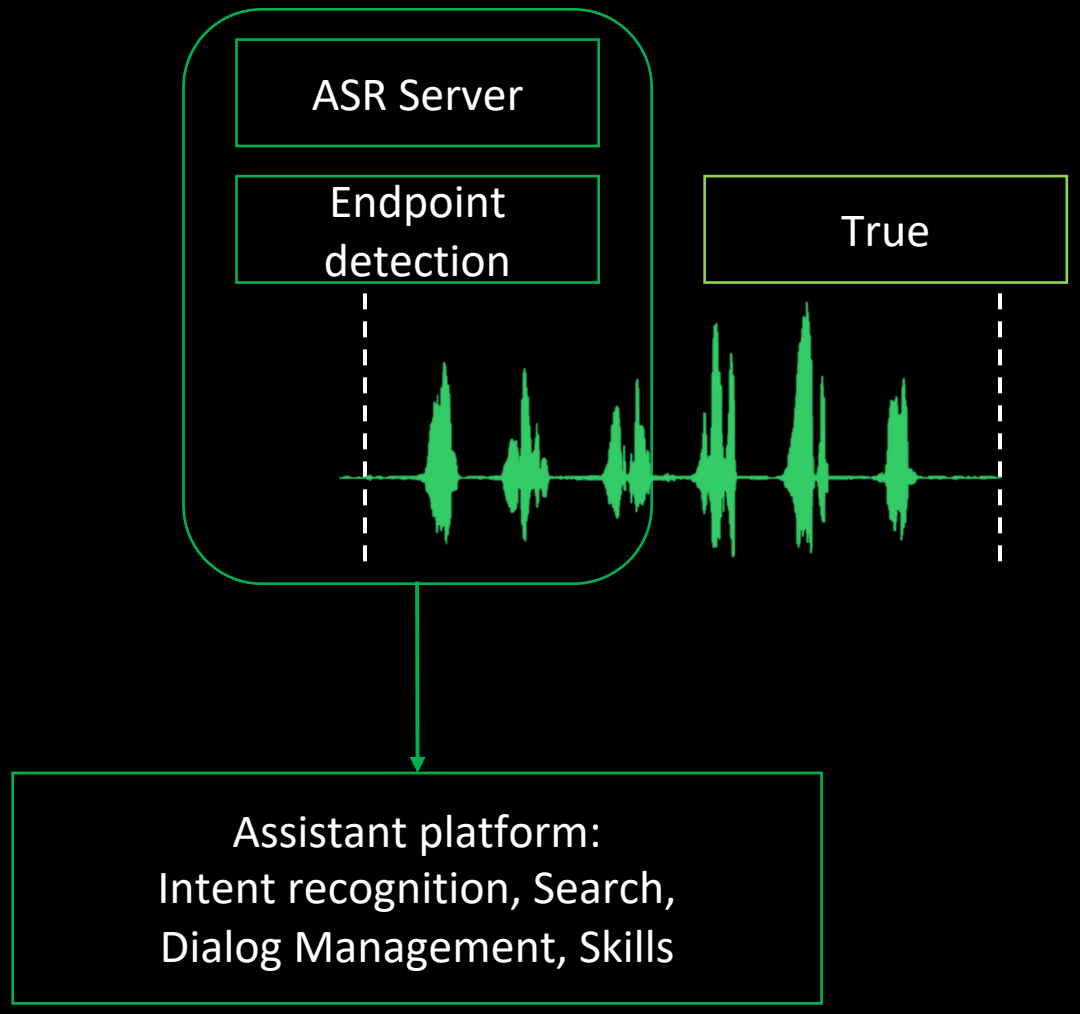
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: audio features



Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

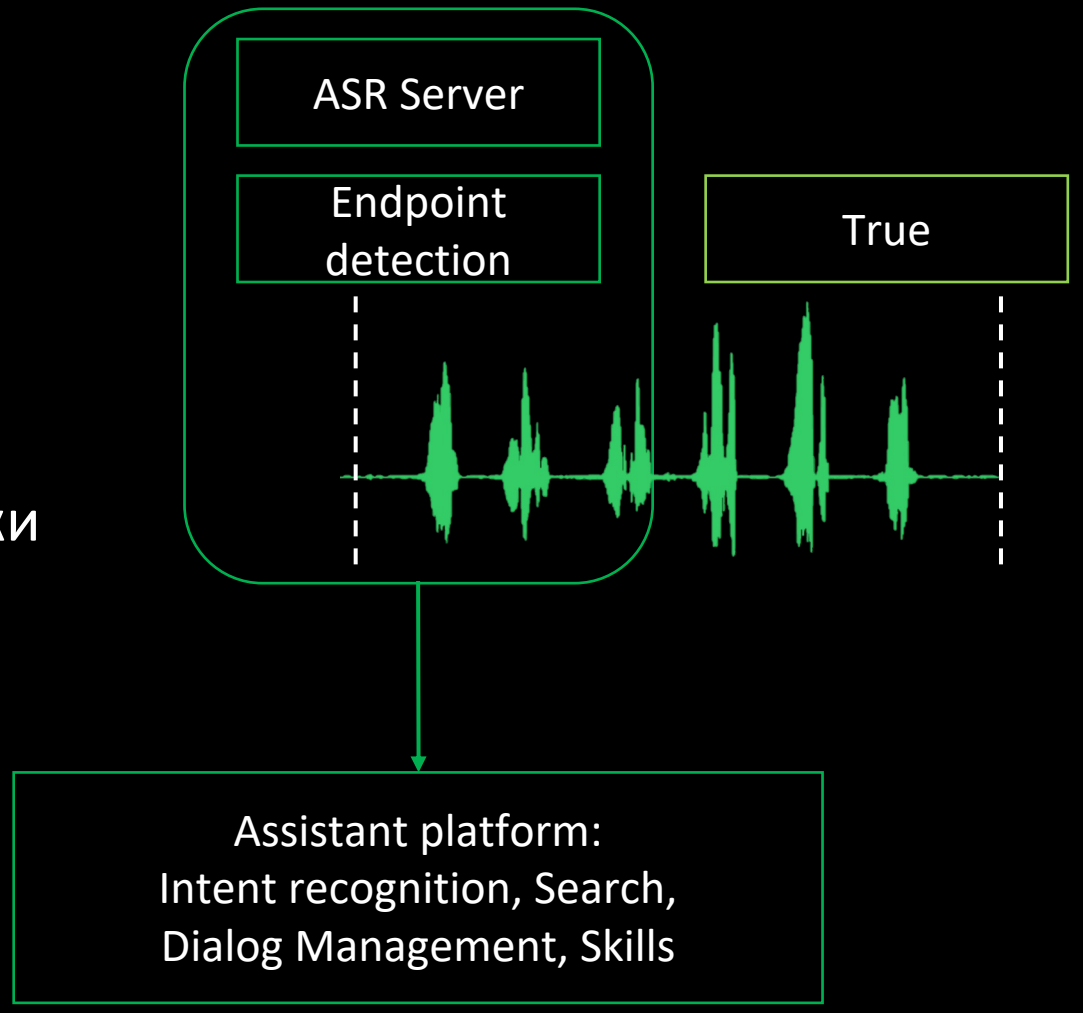
Endpoint detection: audio features



Endpoint detection: audio features

Преимущества:
- Скорость инференса
- Простота системы

Недостатки:
- Точность
- Не учитываем текстовые признаки

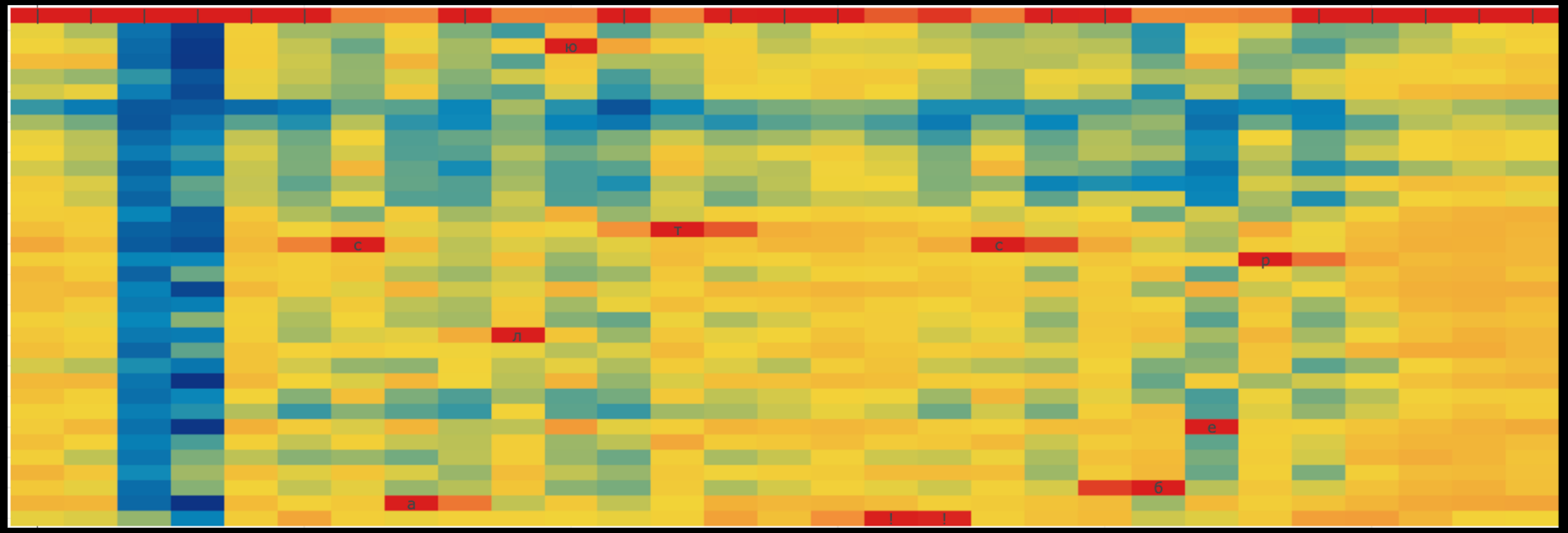


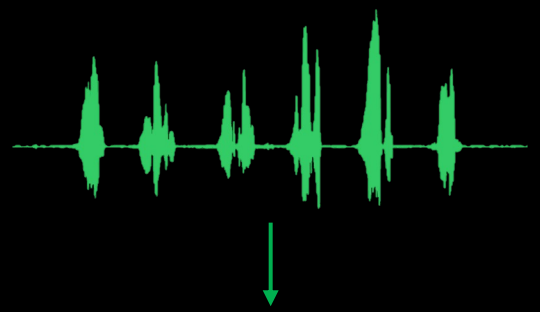
Endpoint detection: audio features

Что еще?

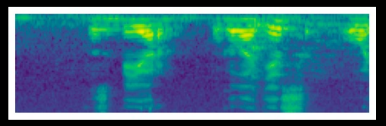
Endpoint detection: audio features

Предсказание конца запроса по логитам CTC





Feature Extractor

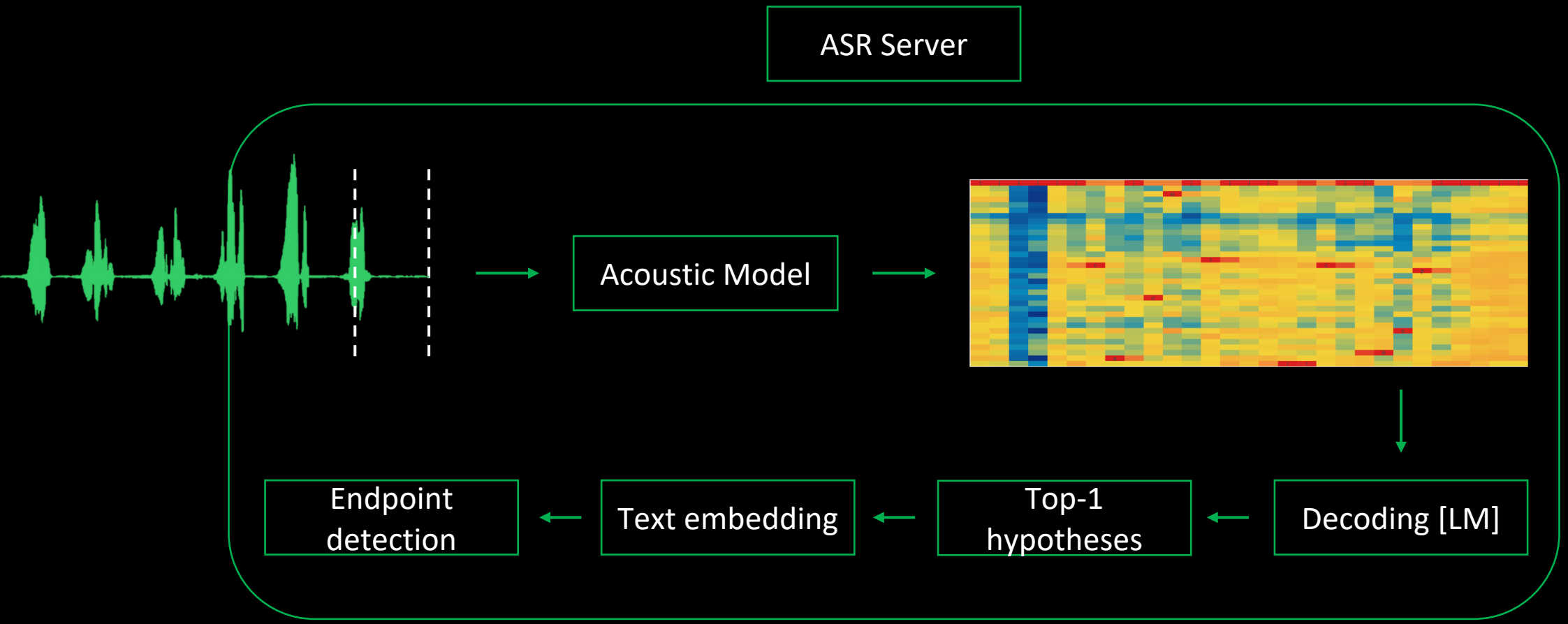


Char-CTC Acoustic Model

Endpoint detection: text features

Предсказание конца запроса на основе text features
- Оценка завершенности запроса по гипотезам ASR

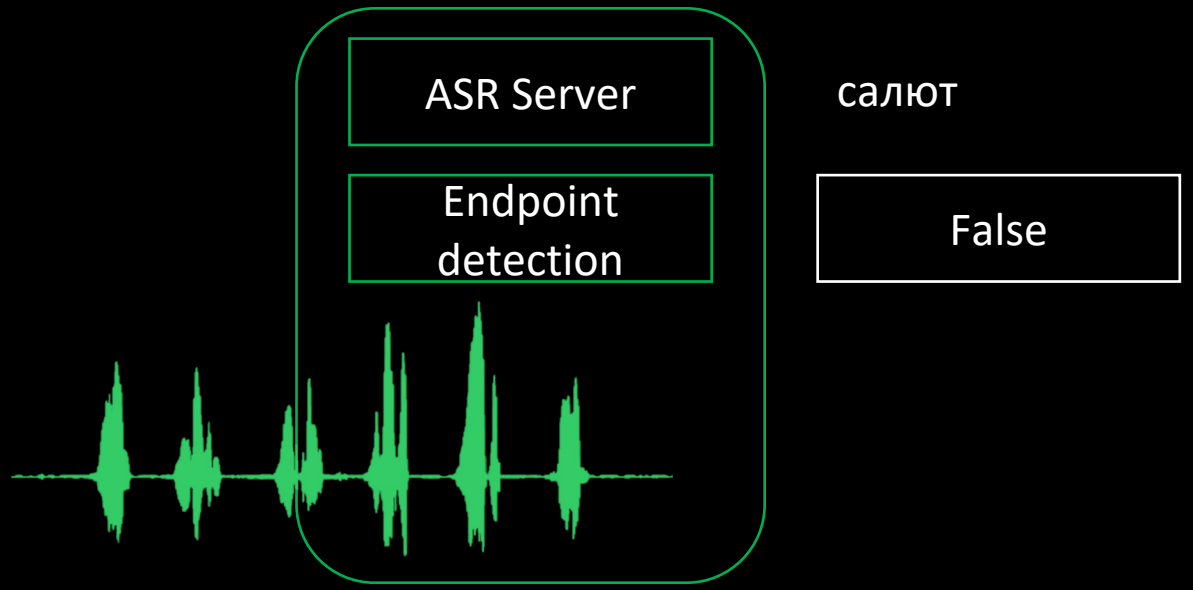
Endpoint detection: text features



Endpoint detection: text features

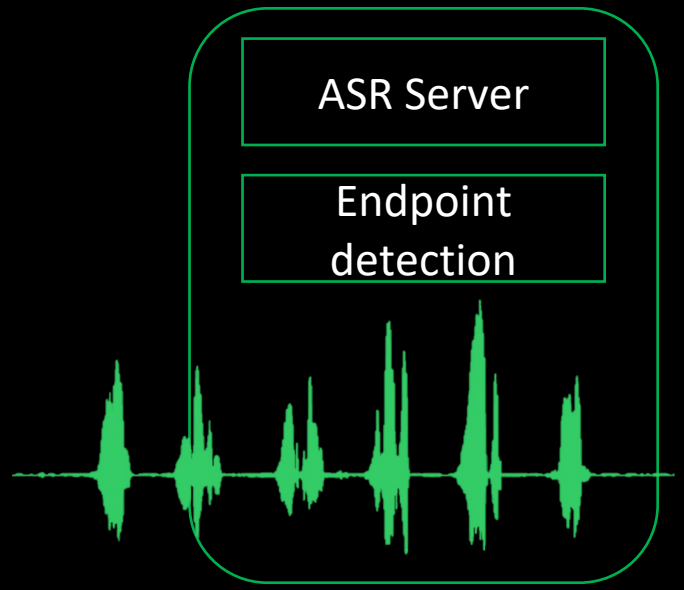
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: text features



Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: text features

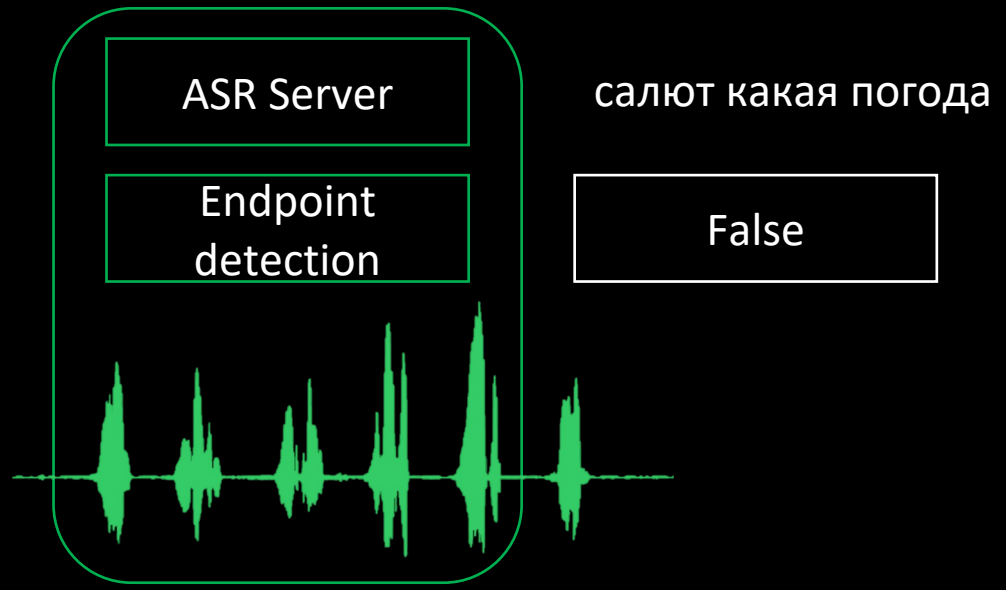


салют какая

False

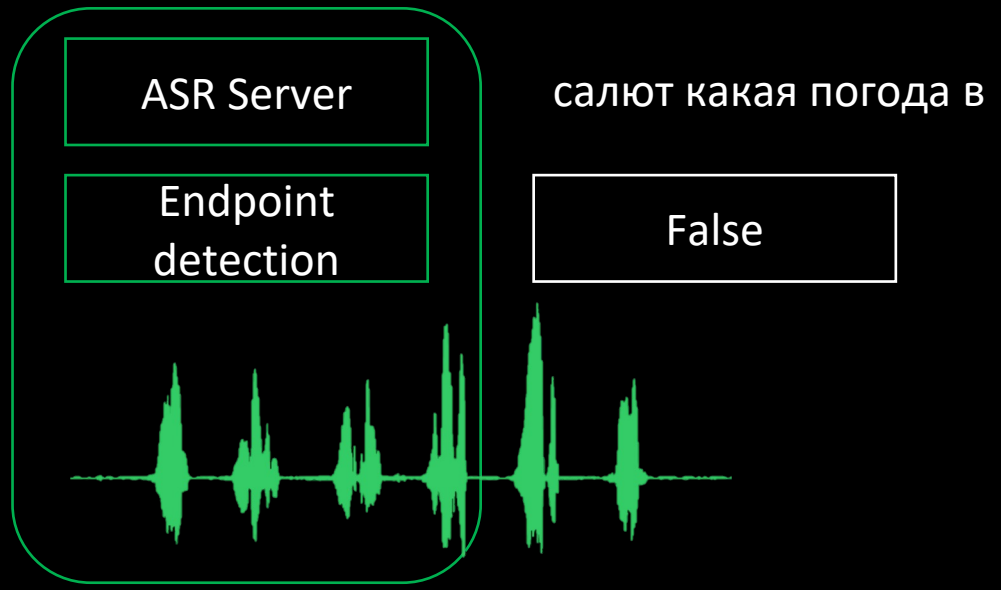
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: text features



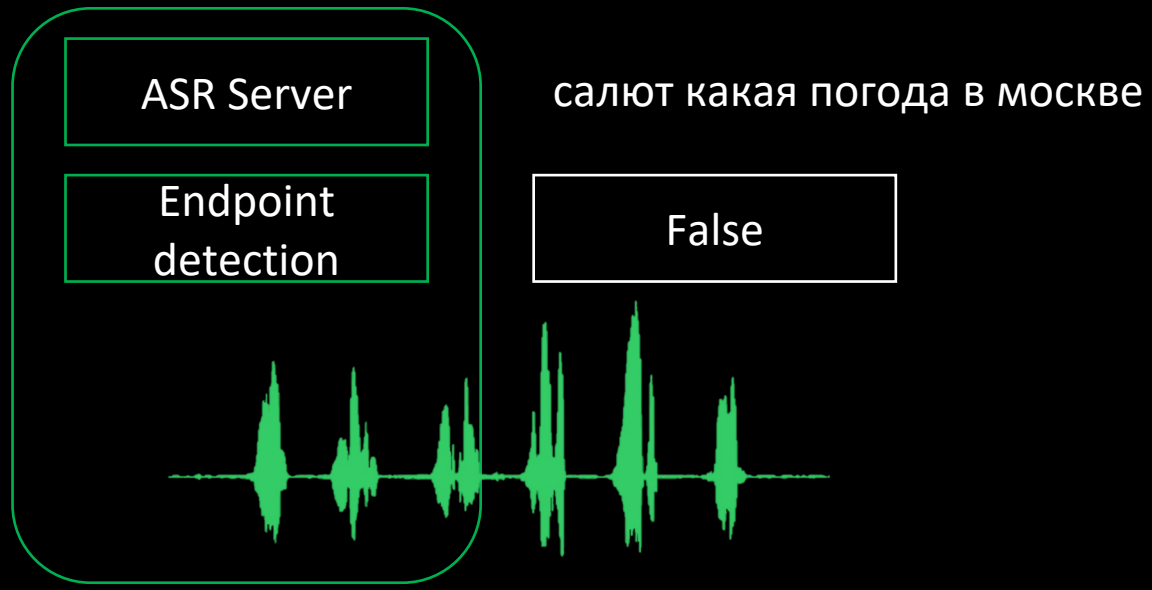
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: text features



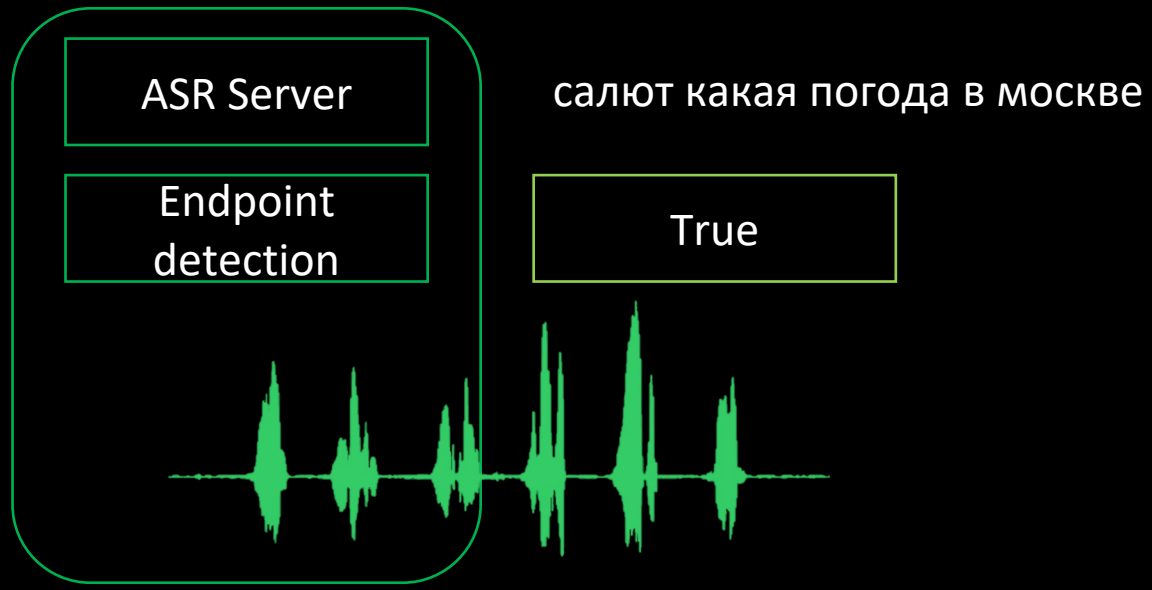
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: text features



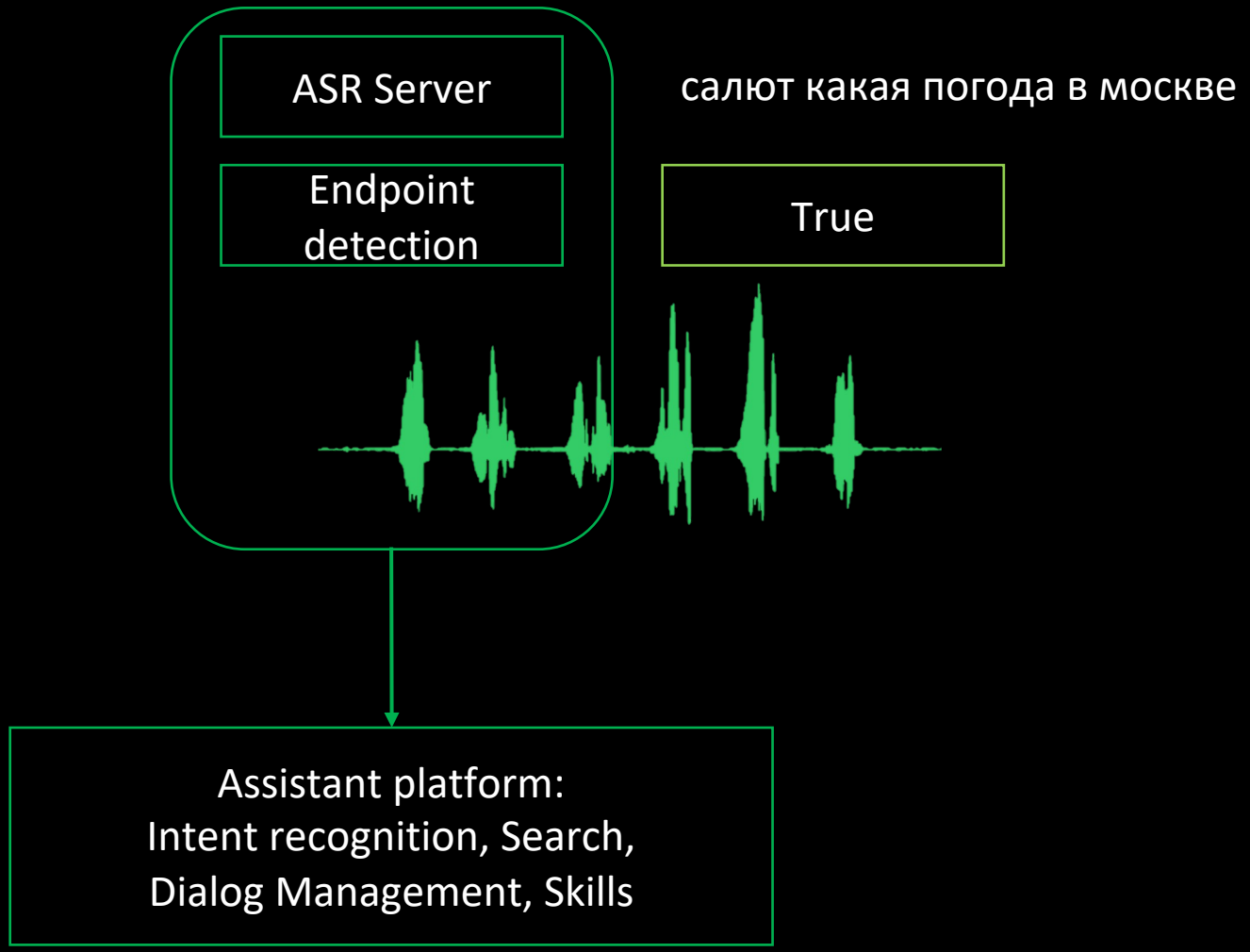
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

Endpoint detection: text features



Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

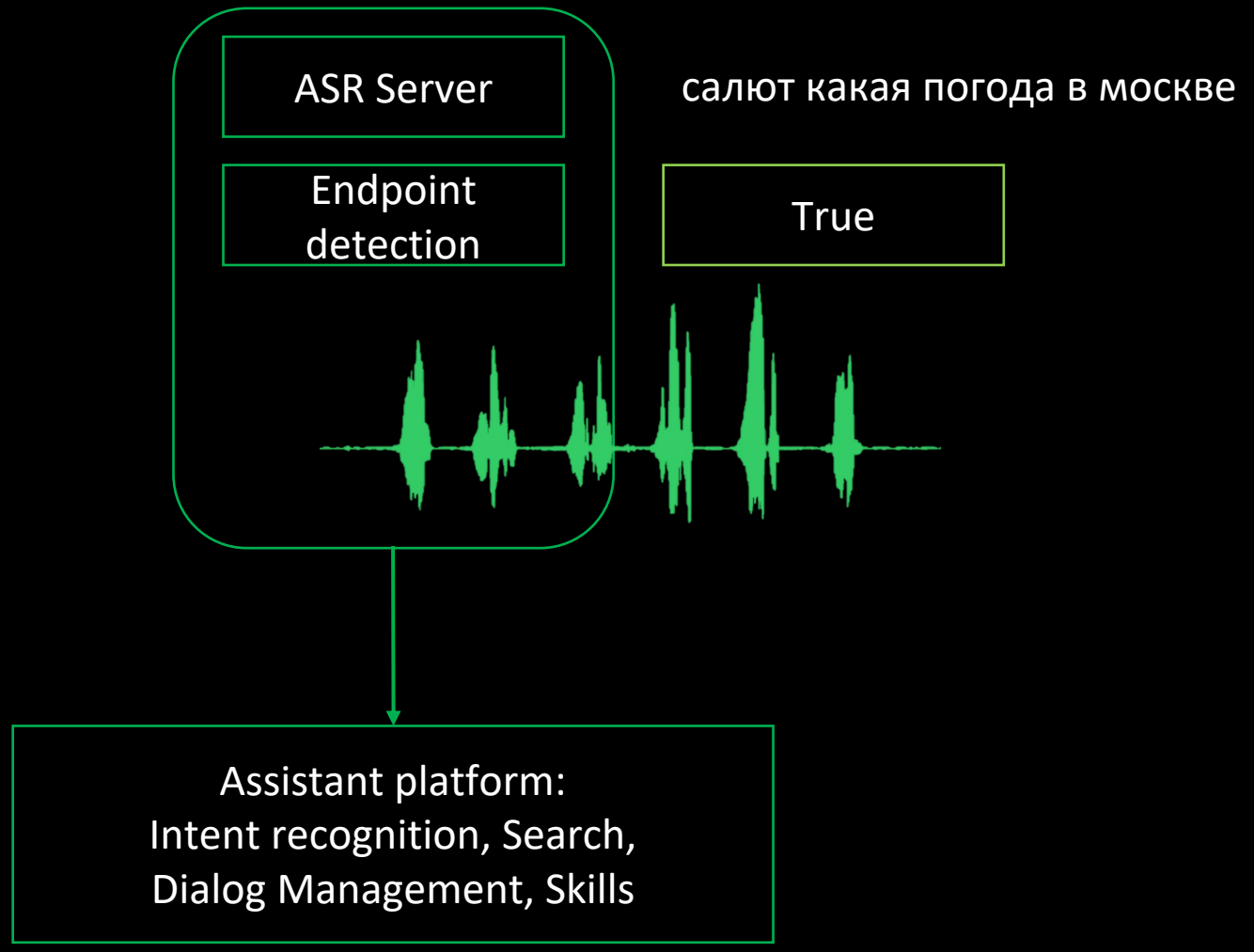
Endpoint detection: text features



Endpoint detection: text features

Преимущества:
- Скорость инференса
- Простота системы

Недостатки:
- Точность
- Не учитываем аудио признаки
- Ошибки ASR модели



Endpoint detection: multimodal system

Предсказание конца запроса на основе audio & text features

Endpoint detection: multimodal system

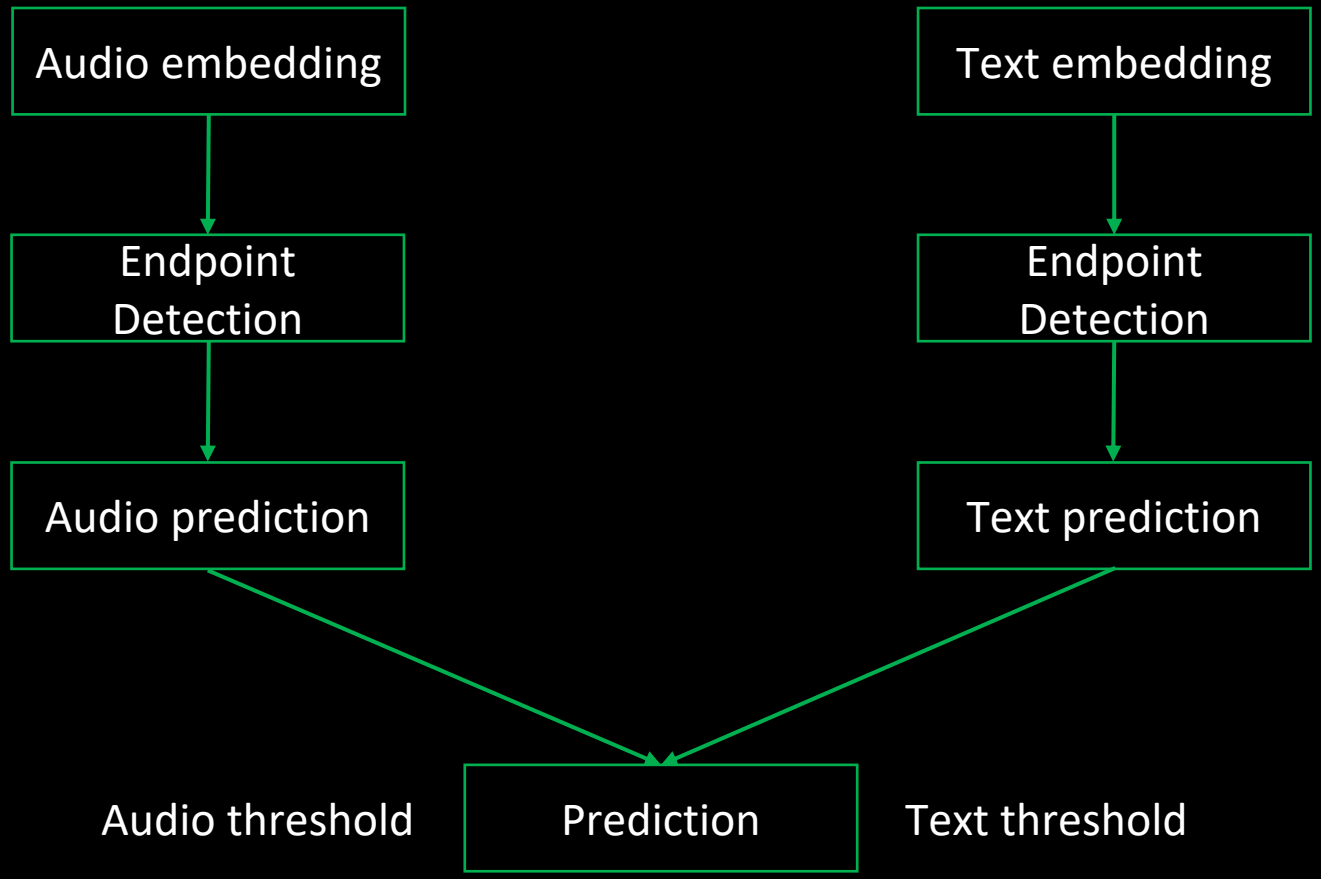
Предсказание конца запроса на основе audio & text features
Как можно объединить audio & text features?

Endpoint detection: multimodal system

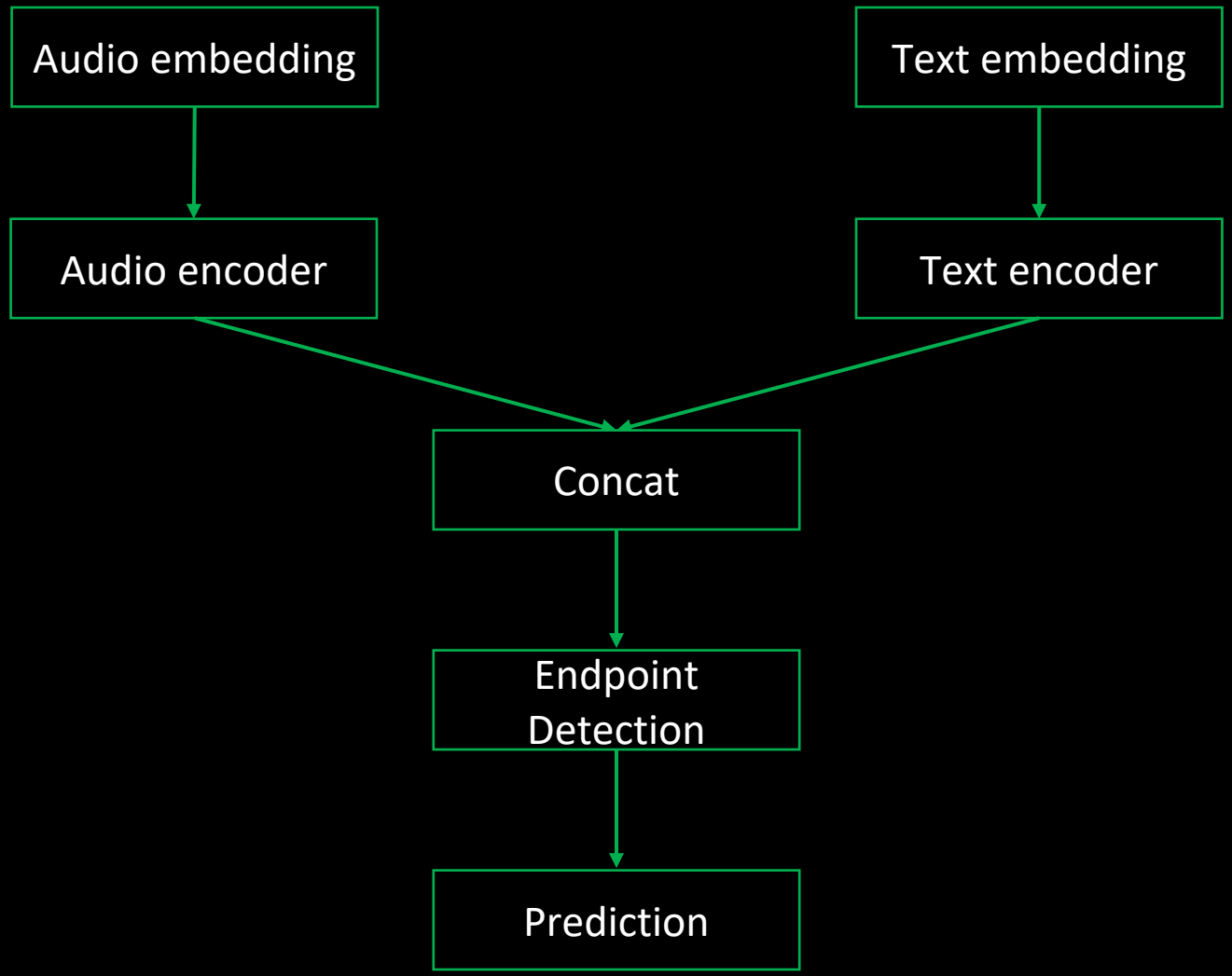
Audio embedding

Text embedding

Endpoint detection: multimodal system



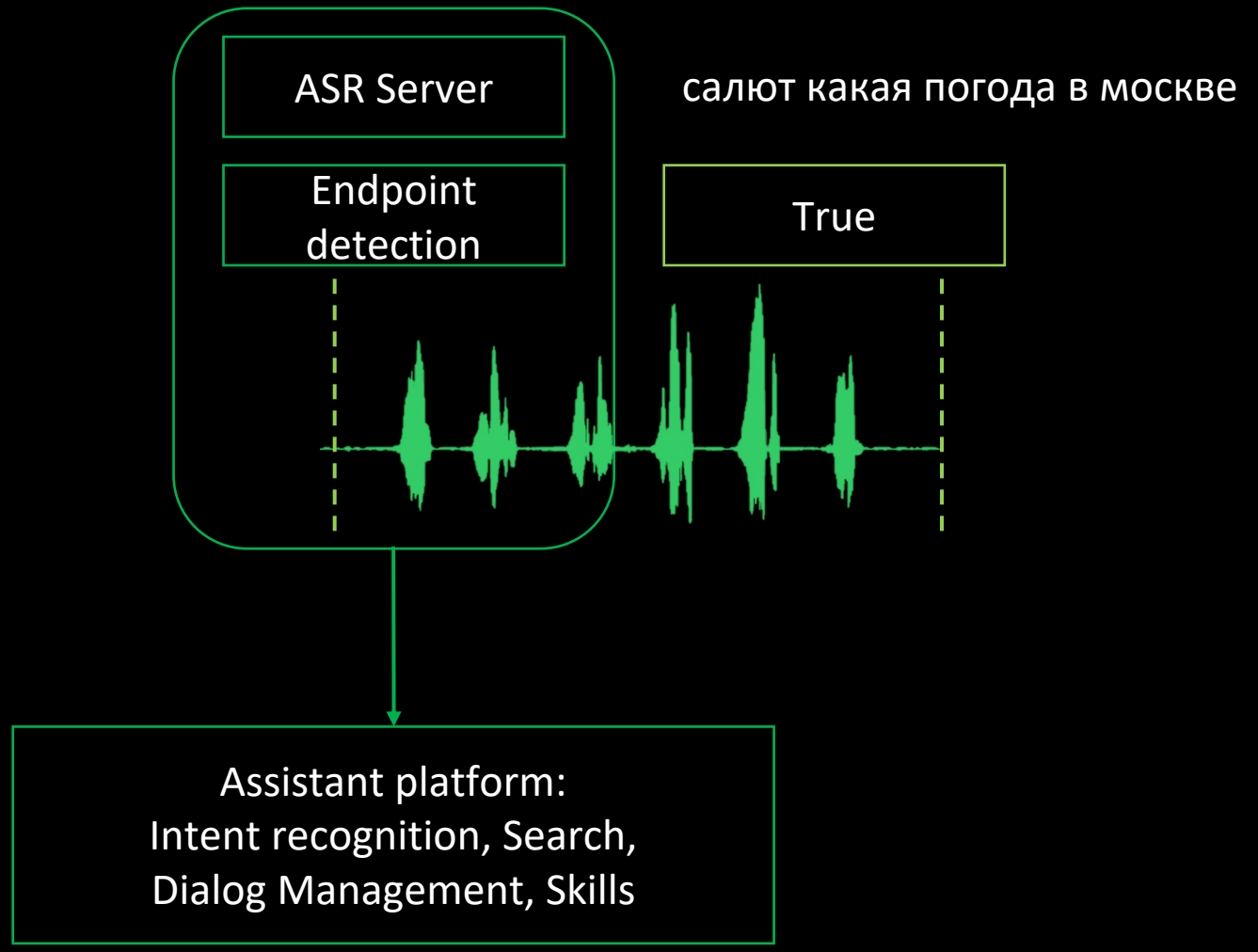
Endpoint detection: multimodal system



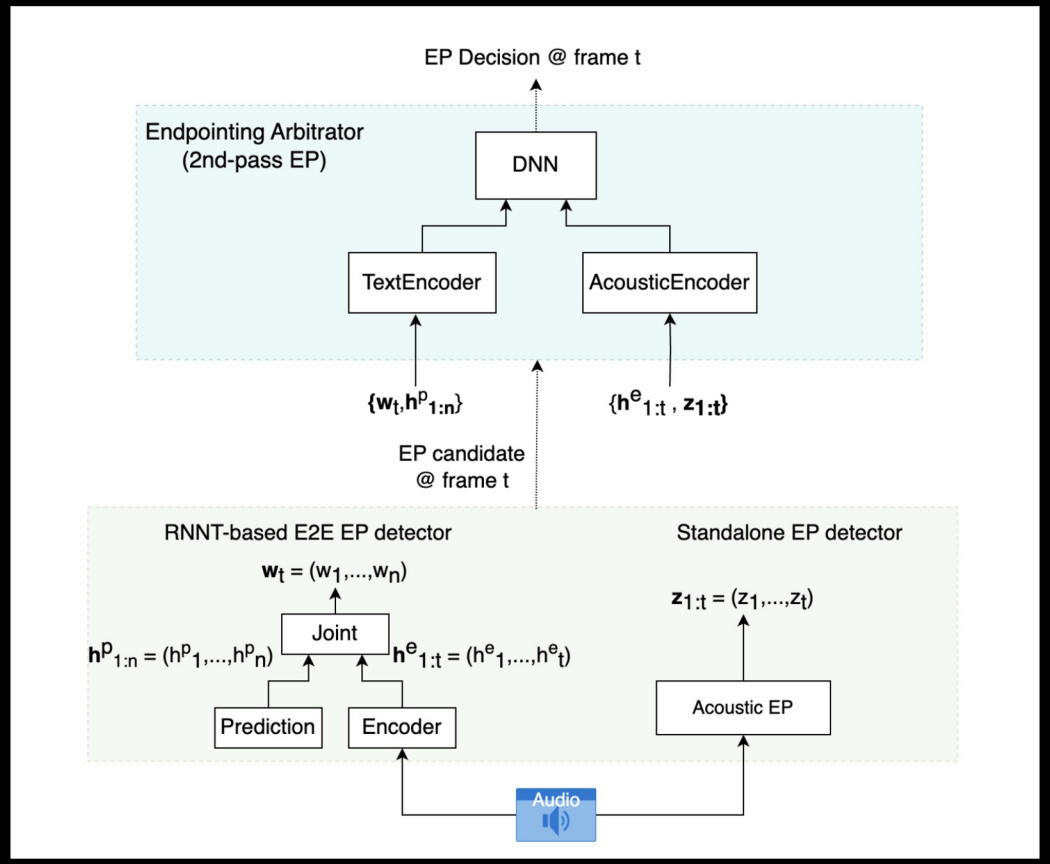
Endpoint detection: multimodal system

Преимущества:
- Точность
- Учет audio & text признаков

Недостатки:
- Сложность системы



Endpoint detection: multimodal system



Two-pass endpoint detection for speech recognition: Amazon Alexa AI, 2024

User perceived latency

Скорость отклика
- Быстрое выполнение команд, без задержек и пауз

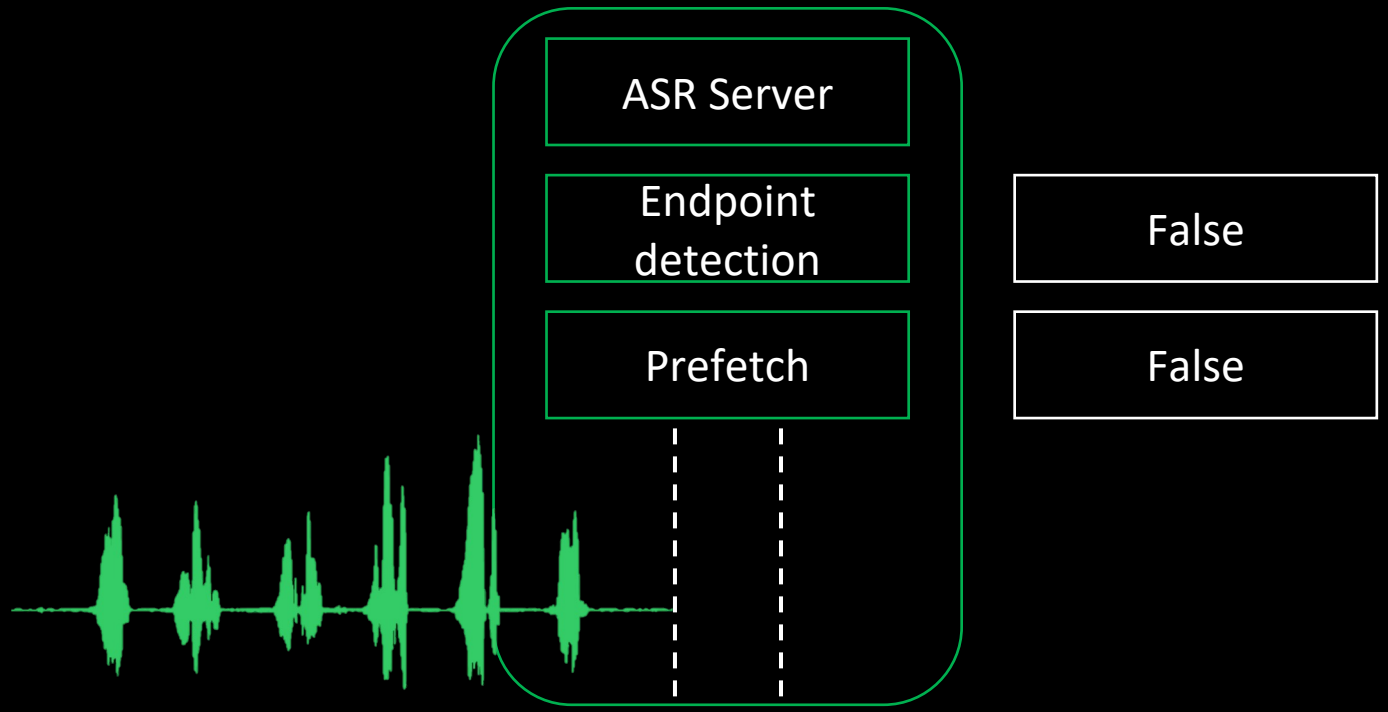
User perceived latency – метрика, отражающая воспринимаемое пользователем время задержки между действием и реакцией системы.

В нашем случае можем выделить два примера:
- Произношение активационной фразы <-> активация колонки
- Произношение голосового запроса <-> получение ответа

User perceived latency: prefetch recognition

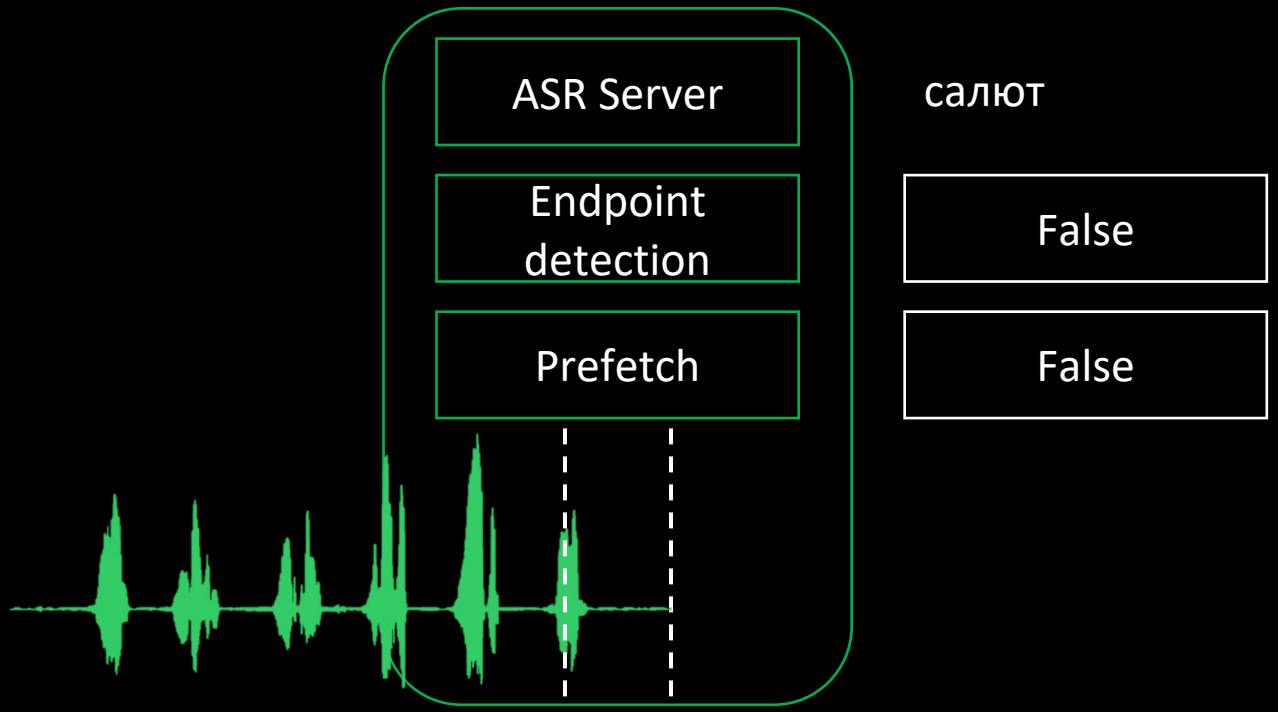
Основная задача в рамках Voice Assistant Pipeline
- Оценка завершенности частично распознанного запроса

User perceived latency: prefetch recognition



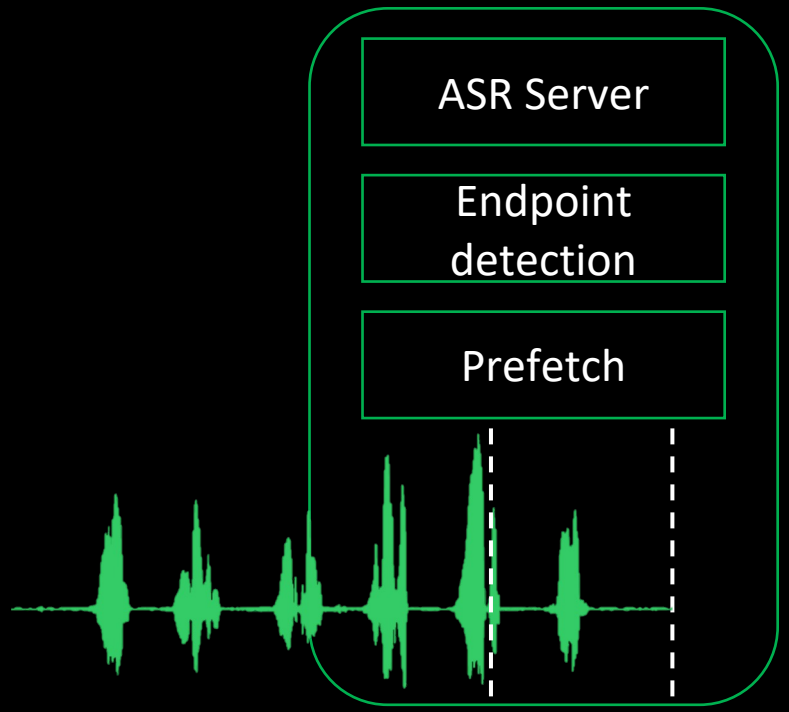
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

User perceived latency: prefetch recognition



Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

User perceived latency: prefetch recognition

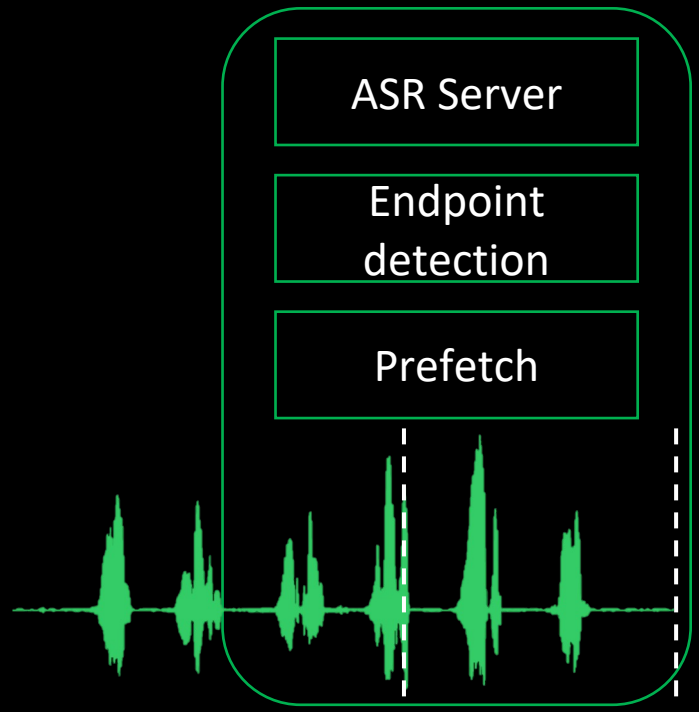


салют какая

False
False

Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

User perceived latency: prefetch recognition



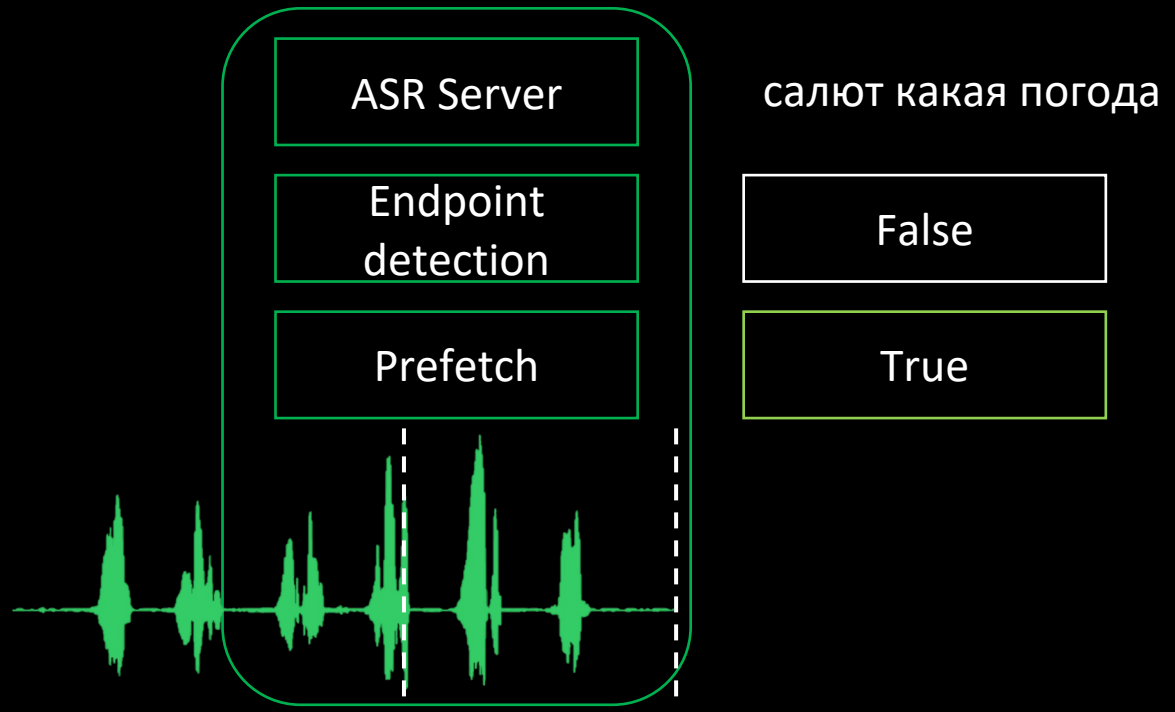
салют какая погода

False

False

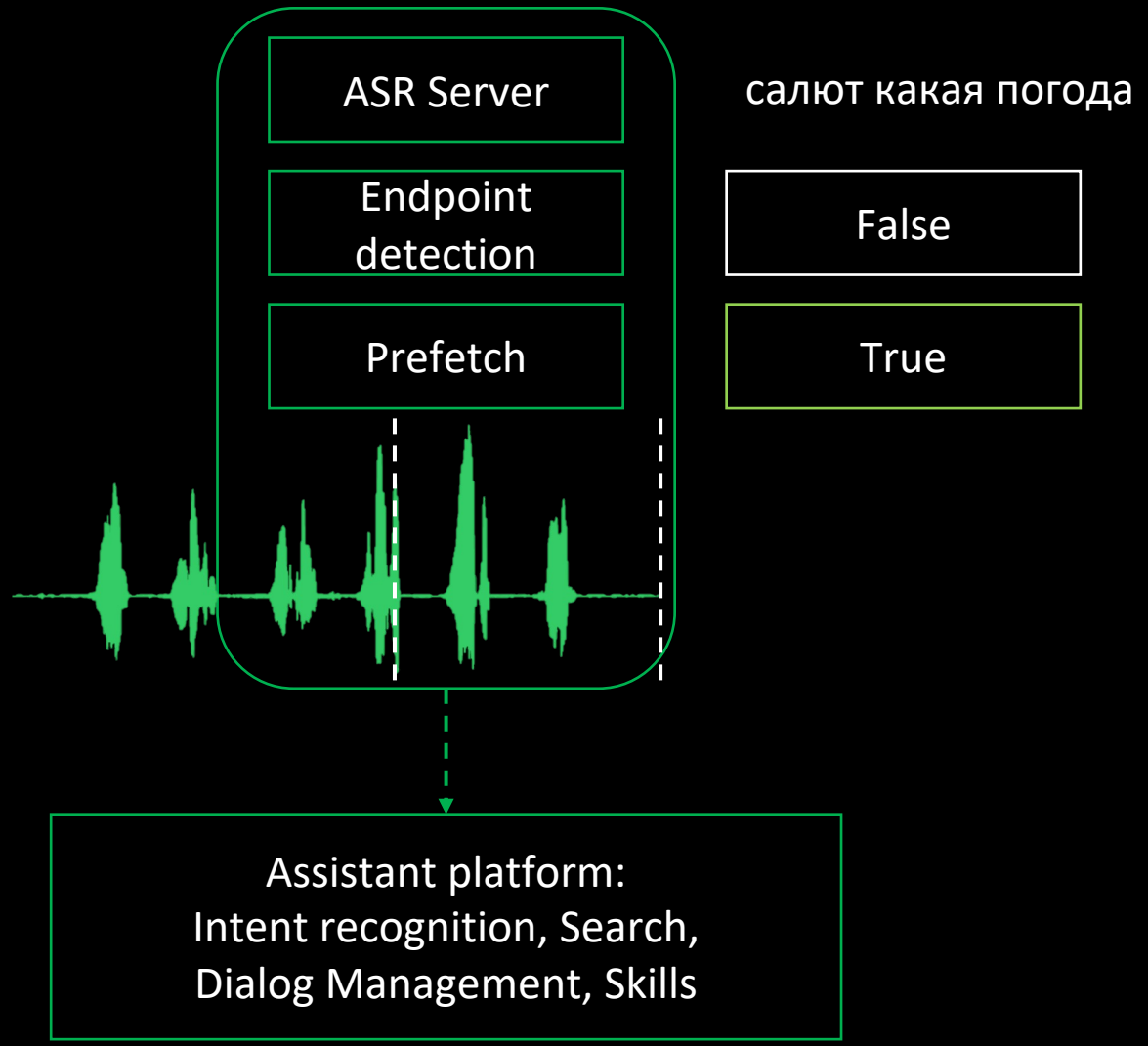
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

User perceived latency: prefetch recognition

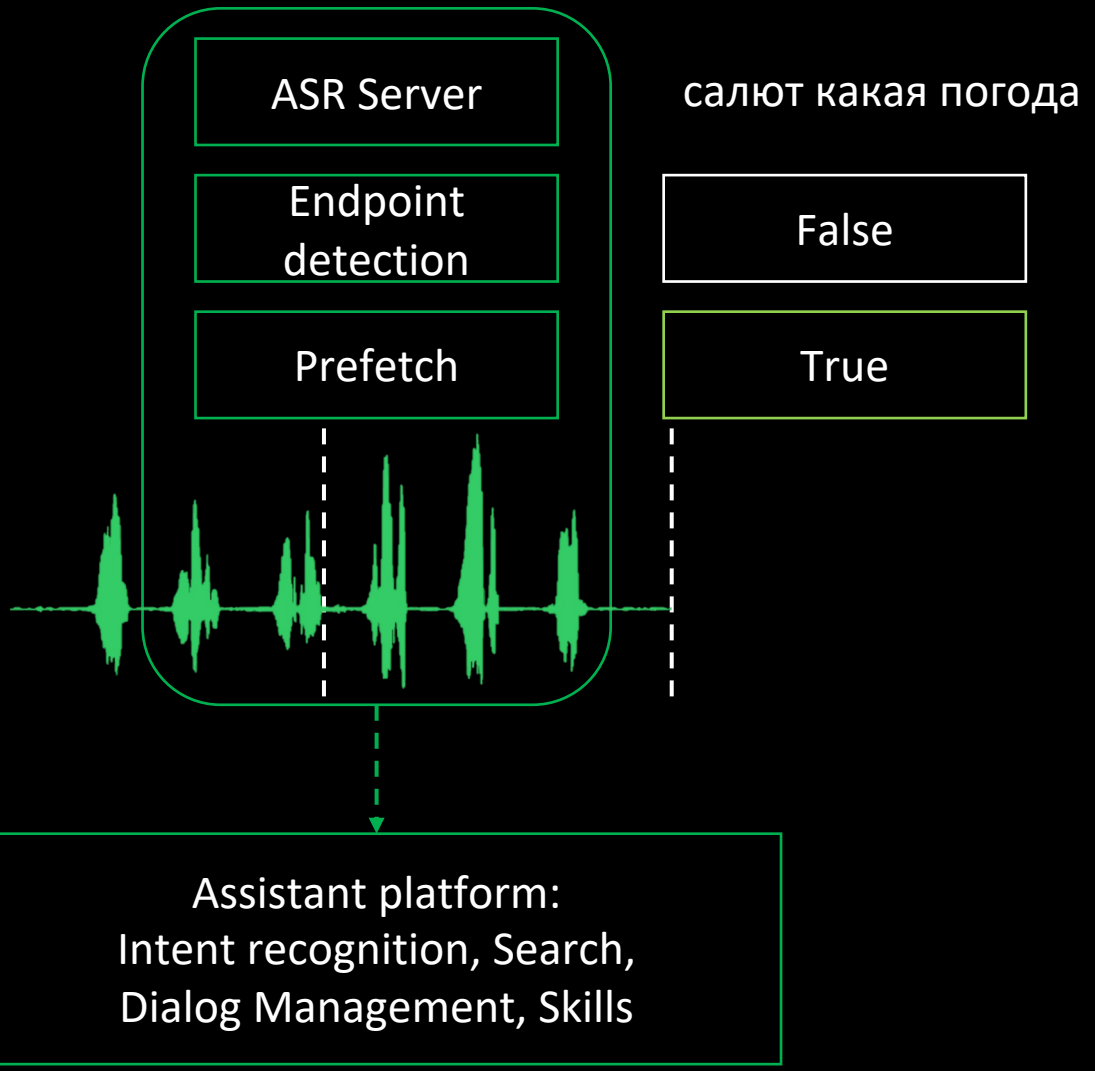


Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

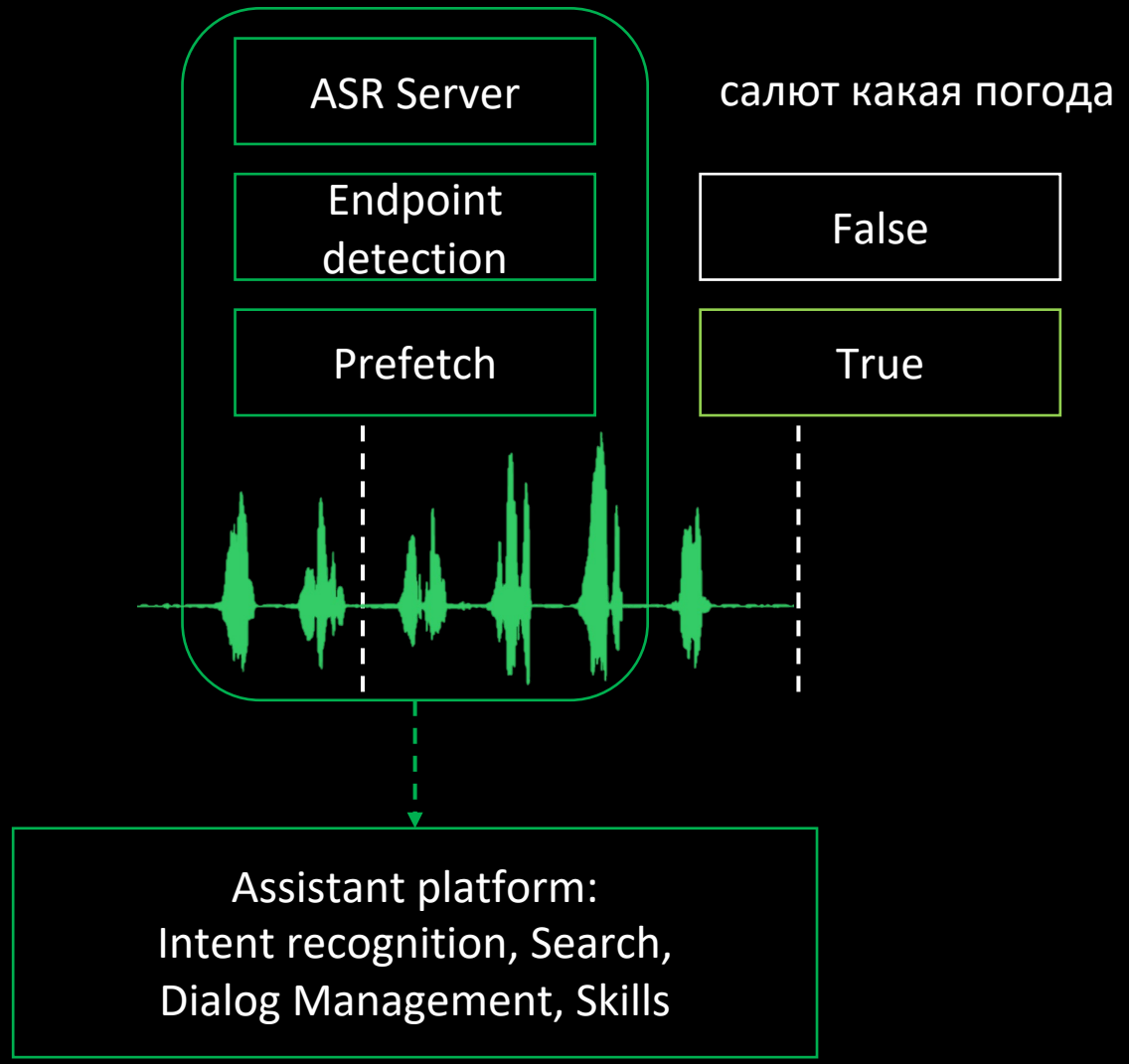
User perceived latency: prefetch recognition



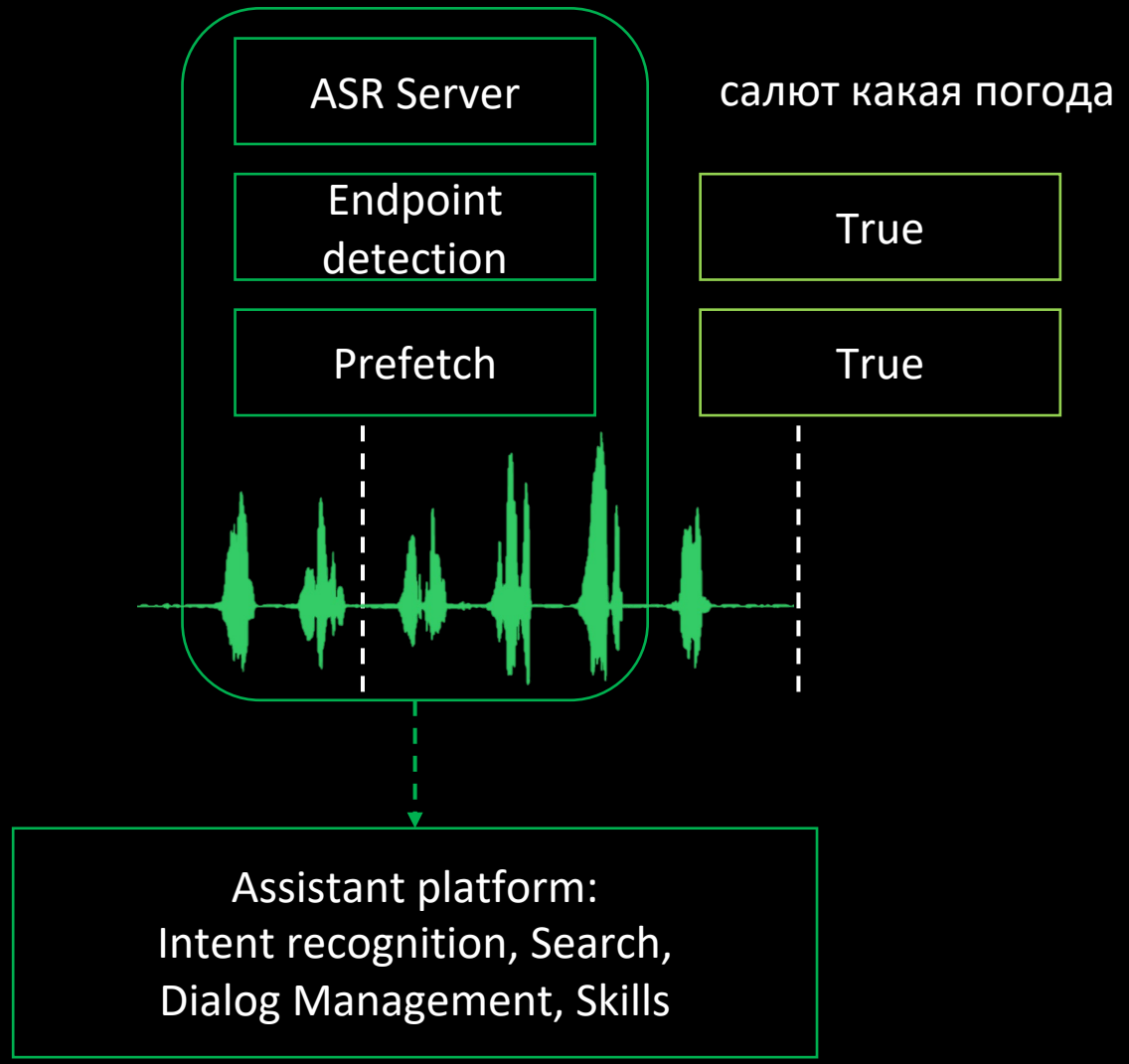
User perceived latency: prefetch recognition



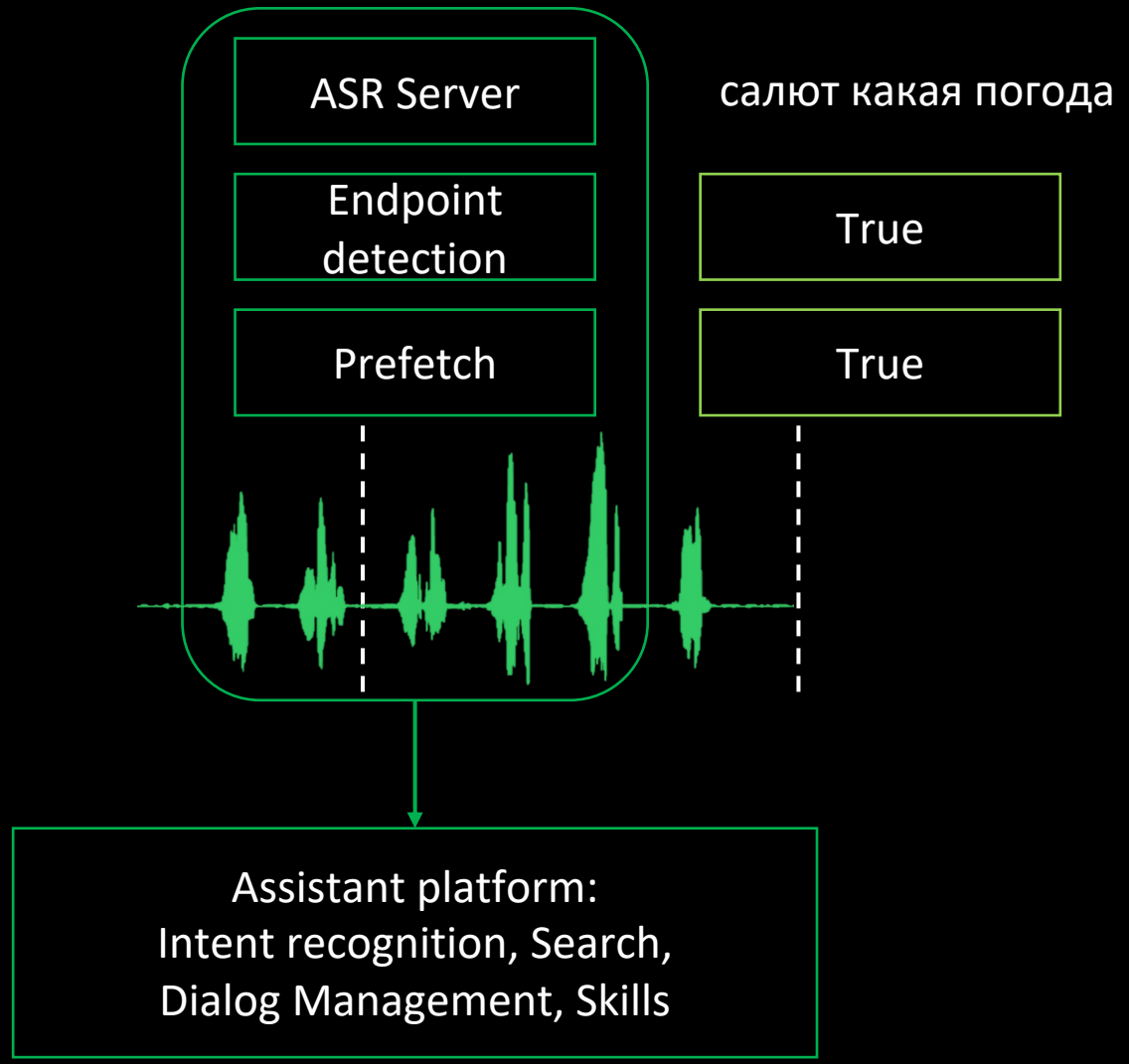
User perceived latency: prefetch recognition



User perceived latency: prefetch recognition



User perceived latency: prefetch recognition



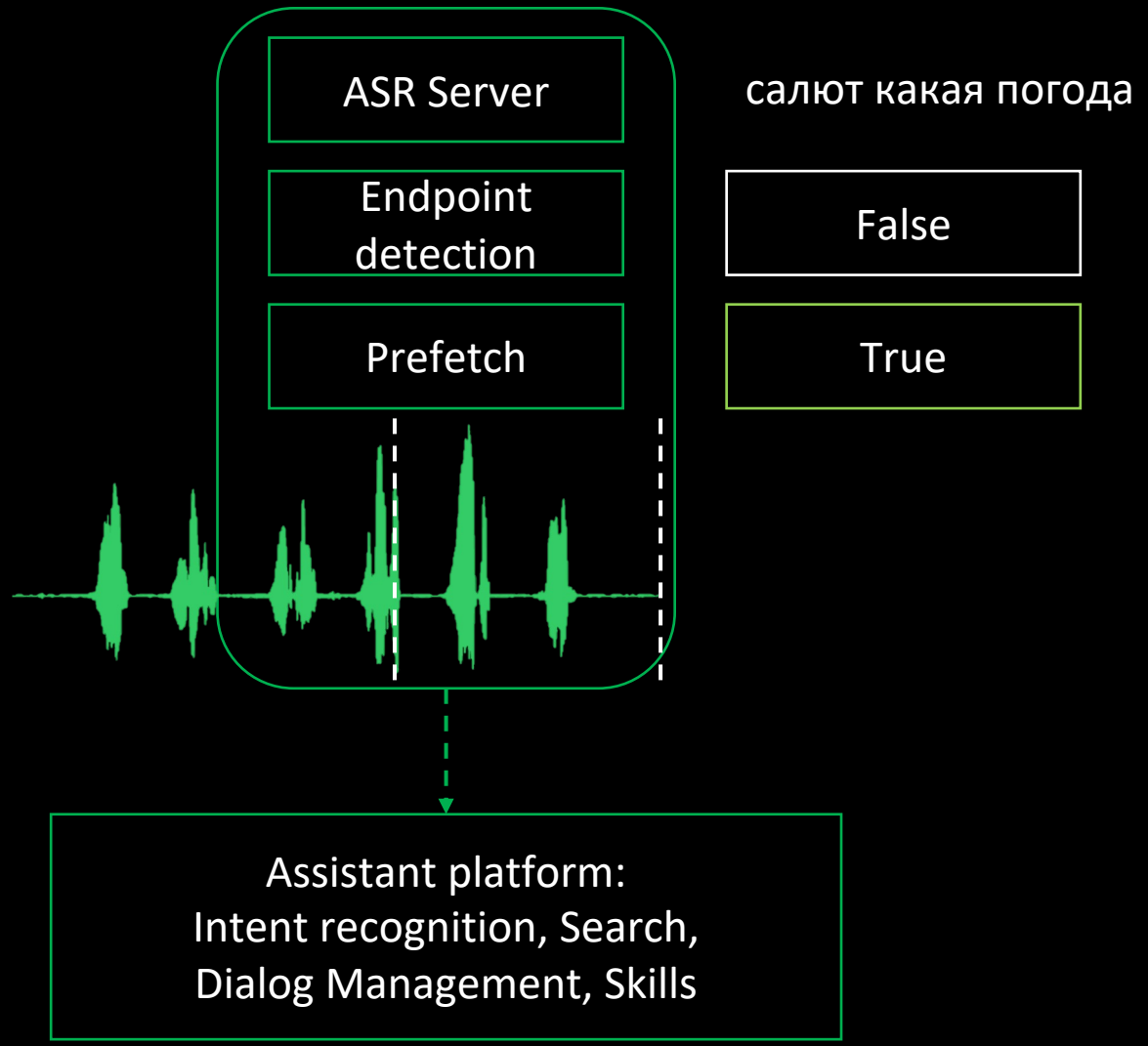
User perceived latency: prefetch recognition

Основная задача в рамках Voice Assistant Pipeline
- Оценка завершенности частично распознанного запроса

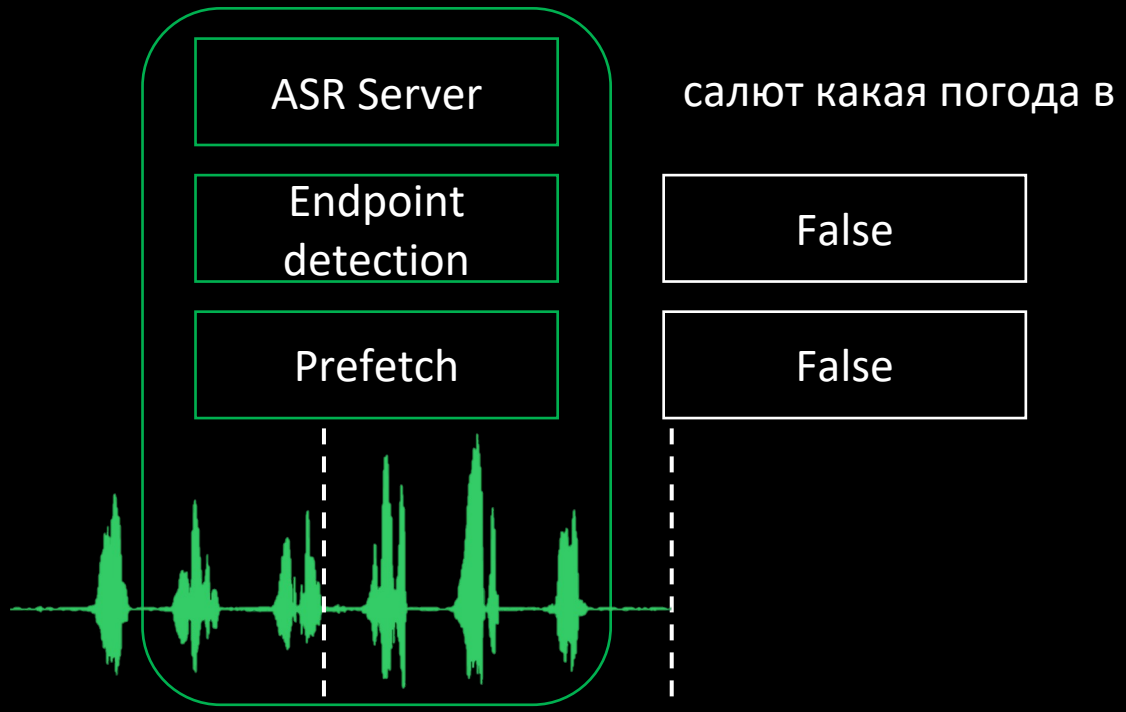
Преимущества:
- Снижаем user perceived latency за счет уменьшения latency на уровне Assistant Platform

Недостатки:
- Дополнительная нагрузка на Assistant Platform

User perceived latency: prefetch recognition

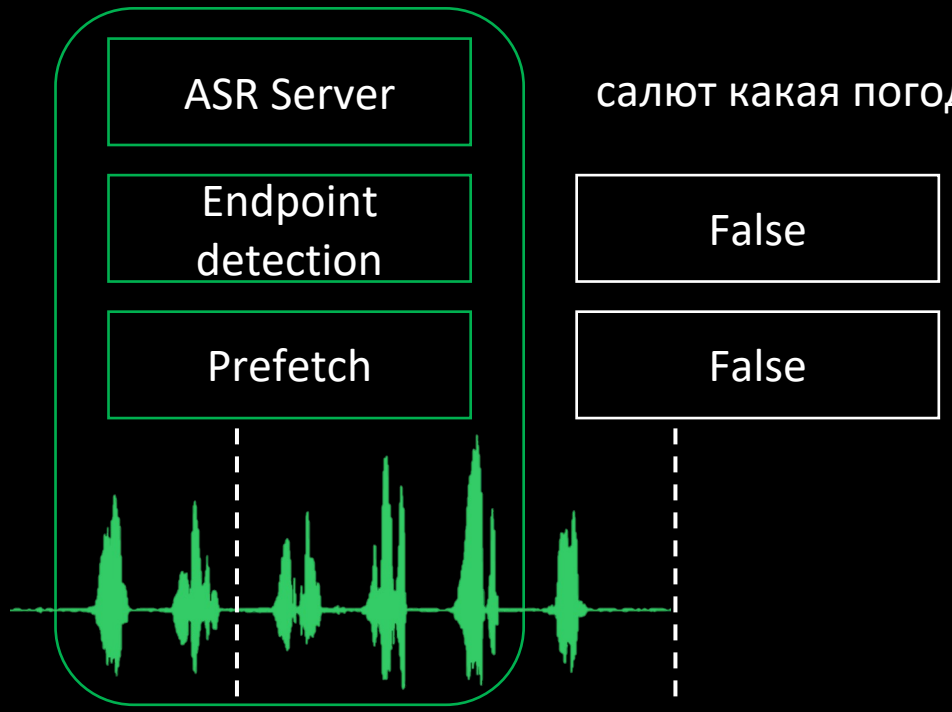


User perceived latency: prefetch recognition



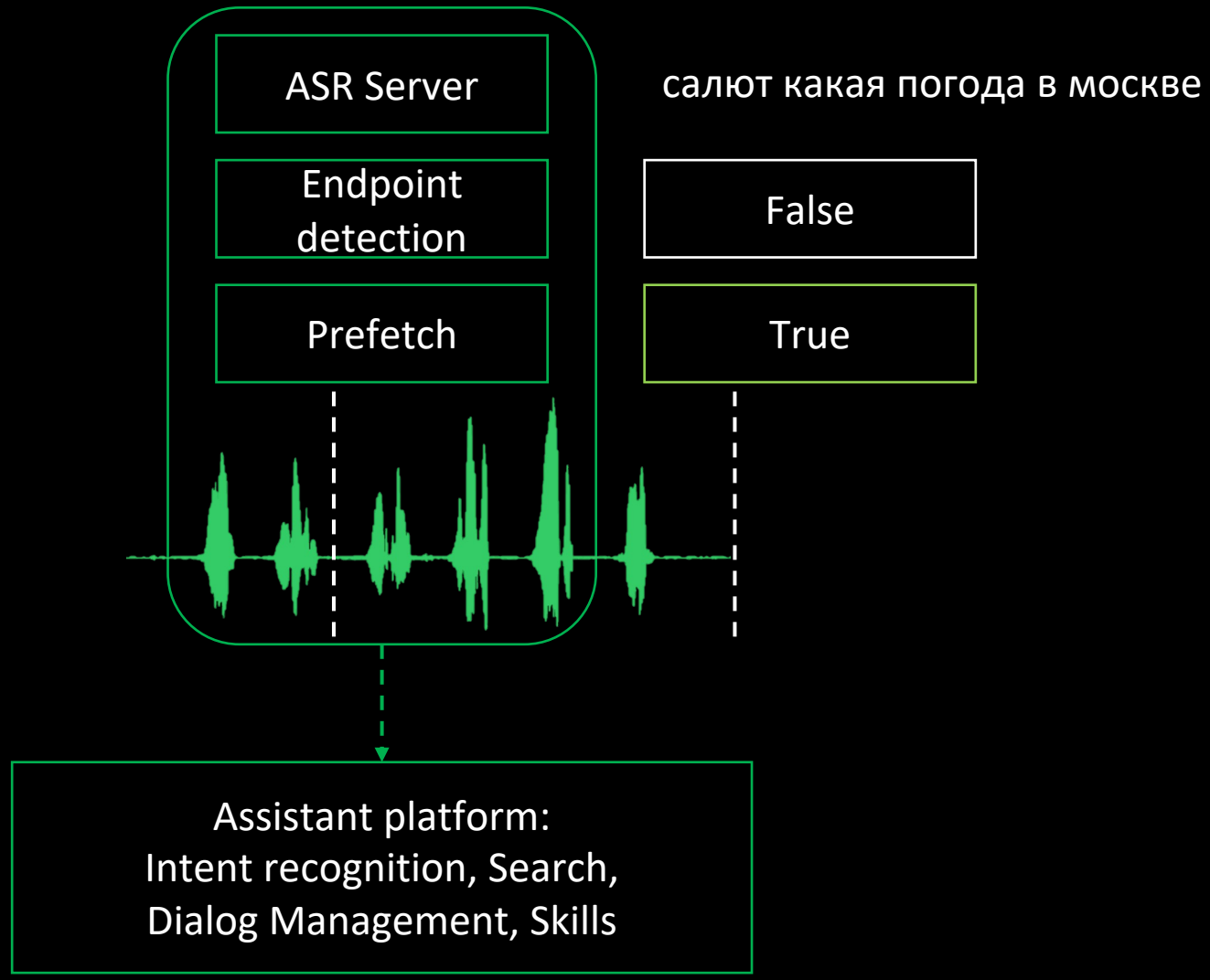
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

User perceived latency: prefetch recognition

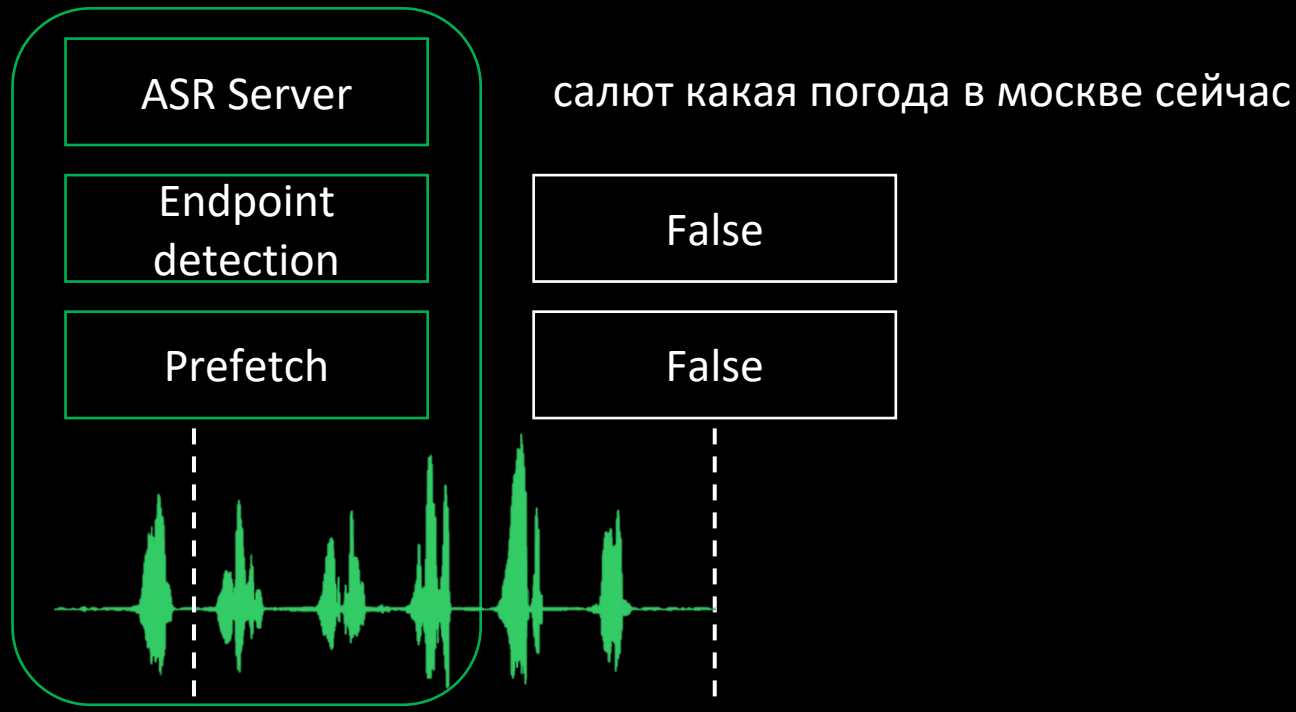


Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

User perceived latency: prefetch recognition

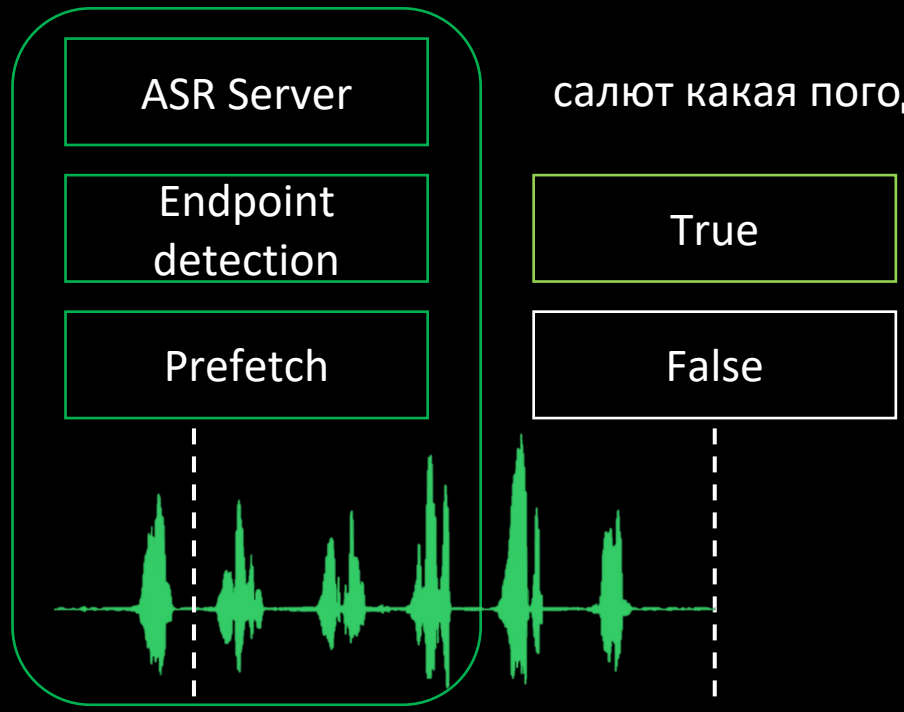


User perceived latency: prefetch recognition



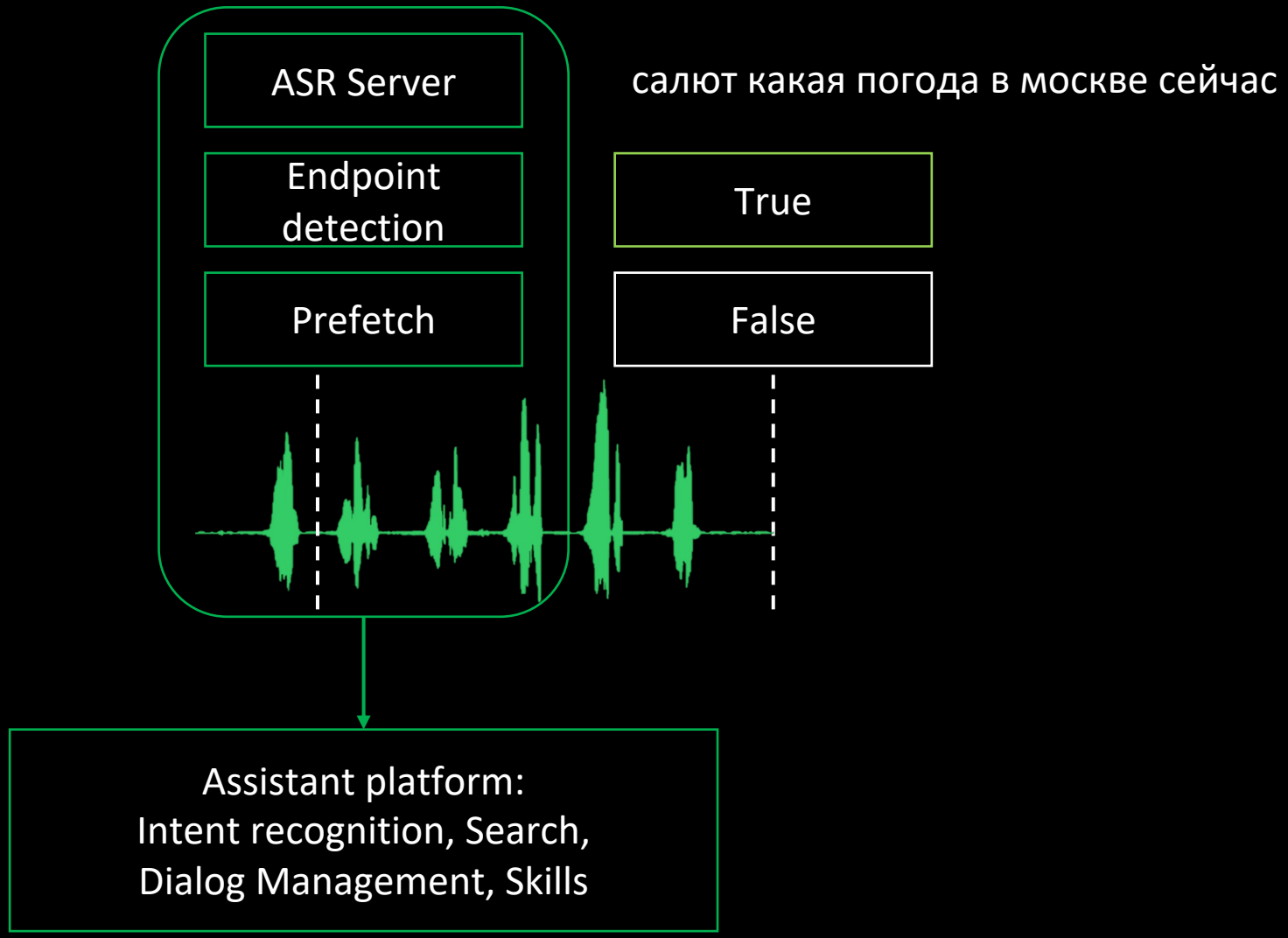
Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

User perceived latency: prefetch recognition



Assistant platform:
Intent recognition, Search,
Dialog Management, Skills

User perceived latency: prefetch recognition



Итоги:

- Voice assistant pipeline
- User perceived latency
  - Keyword detection
    - Подходы: on-device, cascade system, complex cascade system
    - Метрики: FAR, FRR, FAh
  - Endpoint detection
    - Подходы: audio / text only, audio & text
    - Метрики: WERR, EEPR, latency
  - Prefetch recognition