НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ МИФИ

ЛЕКЦИЯ 11. СИНТЕЗ И РАСПОЗНАВАНИЕ РЕЧИ

Демидов Д.В.

Обработка аудиовизуальной информации.
Бакалавры, 6 семестр. Магистры, 9 семестр

План лекции

- Синтез речи
  - Фонемы и морфемы
  - Методы синтеза
- Распознавание речи
  - Поиск звуковых фрагментов
  - Голосовая биометрия
  - 4-х факторная авторизация
  - Распознавание голоса

Синтез речи

Основные задачи

Некоторые технологии

Эуфония – механическая говорящая машина Фабера (1845)

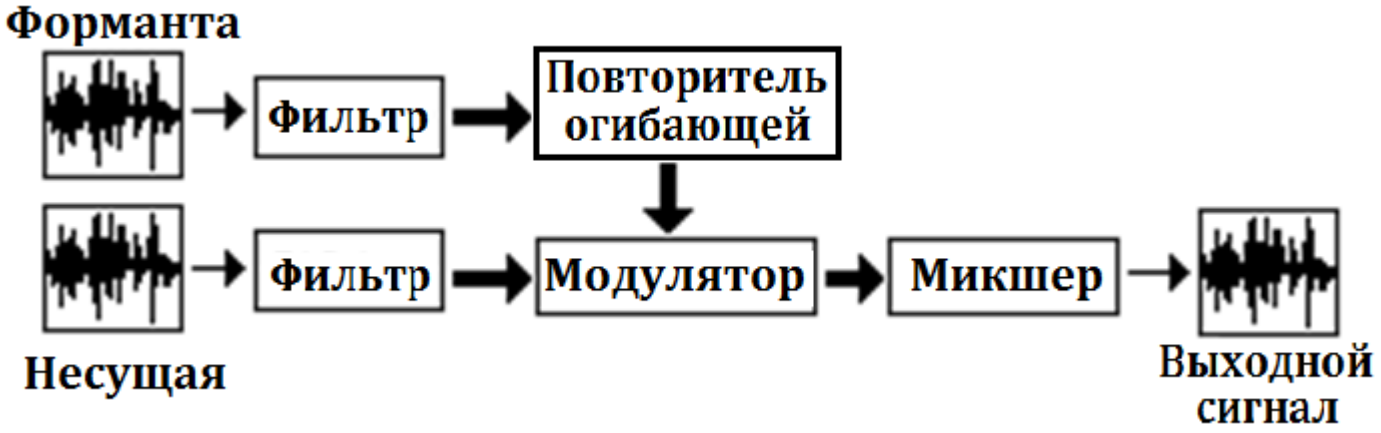
- Воздушный мех, приводимый в движение ножной педалью - «лёгкие».
- Вытесняемый из меха воздух при помощи ряда клавиш направляется в различные по объёму трубки - разные положения голосовой щели и полости рта.

Вокодер – voice coder

- ***Вокодер*** – синтезатор речи на основе произвольного сигнала с богатым спектром.
- Изначально вокодеры были разработаны в целях экономии частотных ресурсов радиолинии системы связи при передаче речевых сообщений.
- Вместо собственно речевого сигнала передают только значения его определённых параметров, которые на приёмной стороне управляют синтезатором речи.

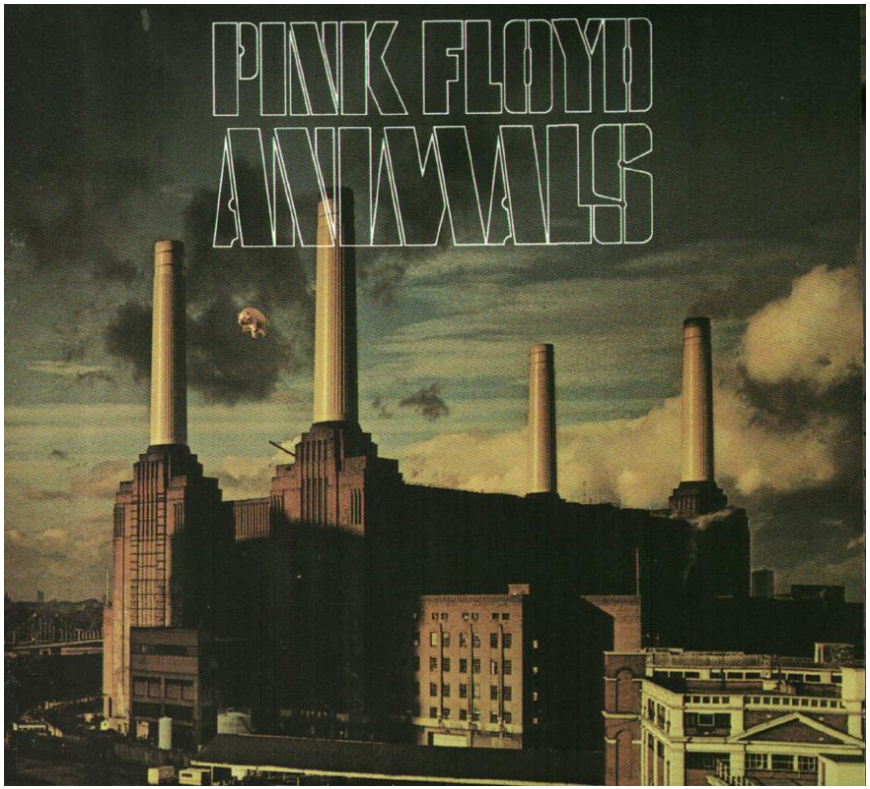
Синтезатор речи в вокодере

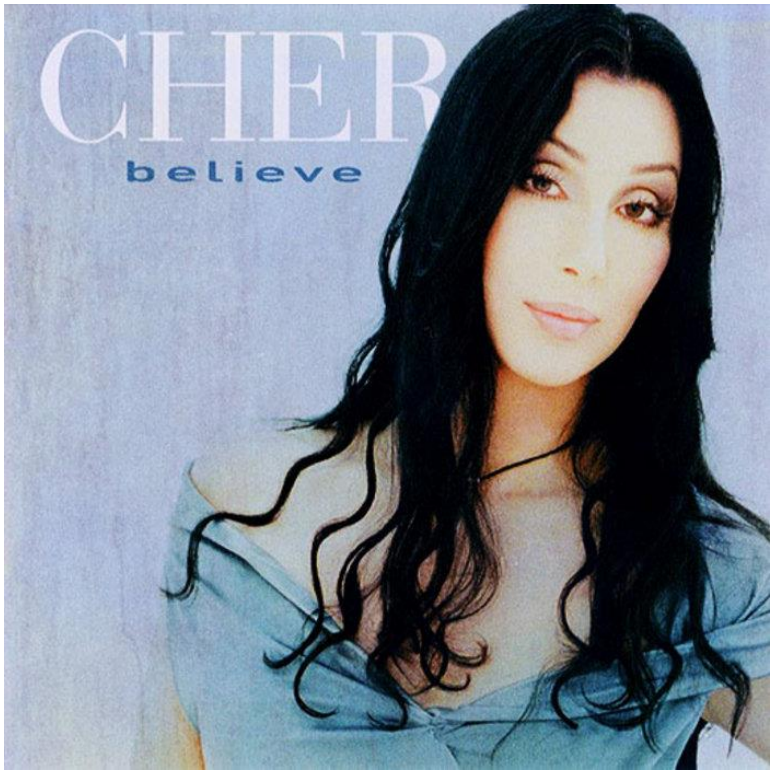
- Основу синтезатора речи составляют три элемента:
  - Генератор тонального сигнала для формирования гласных звуков;
  - Генератор шума для формирования согласных;
  - Система формантных фильтров для воссоздания индивидуальных особенностей голоса.



Вокодер в музыке

- Pink Floyd – Dogs, Sheep ‘1977 (6:33 «the Lord is my shepherd…»)
- Electric Light Orchestra – Mr. Blue Sky ‘1977
- Cher – Believe ‘1998 (0:33)
- Daft Punk – Get lucky ‘2013 (эффекты)





Фонема

- **Фонема** – звук (греч.), абстрактная единица языка, соответствующая звуку речи как конкретной единице, в которой фонема материально реализуется.
- **Фоны** (звуки речи) бесконечно разнообразны: один человек никогда не произносит одинаково один и тот же звук.
- **Аллофон** – вариант реализации фонемы, обусловленный конкретным фонетическим окружением. Таких окружений гораздо меньше, чем число возможных звуков, поэтому удобно представлять фонему основными аллофонами, а не всеми вариантами фонов.
- Все варианты произношения звука, позволяющие правильно опознавать и различать слова с этим звуком, будут являться реализацией одной и той же фонемы.

Назначение фонем

- Фонема исполняет две ключевые функции, которые характеризуются наличием тесной связи друг с другом:
  - **конститутивная функция** (constitute) состоит в предоставлении фонемного инвентаря, своеобразного строительного материала для конструирования морфем и иных вышестоящих единиц языка;
  - **дистинктивная функция** (distinct) состоит, в свою очередь, в обеспечении различения отдельных морфем.
- Для записи транскрипций слов используется международный фонетический алфавит.

Морфема

- **Морфема** – наименьшая единица языка, значимая часть слова
- Деление морфем на части приводит к выделению незначимых элементов – фонем
- **Классы** морфем:
  - Корни
  - Аффиксы
    - Префиксы (приставки),
    - Постфиксы (после корня)
      - Суффиксы,
      - Флексии (окончания)
      - Постфиксы (после корня)
    - Интерфиксы (между двух корней)
    - Некоторые более экзотические

Сложность сегментирования речи

- Устная речь являет собой практически непрерывный звуковой поток, представляющий определенные сложности для сегментирования.
- В большинстве случаев для выделения фонемы необходимо знание языка.
- Выделяемость фонемы некоторым образом сопряжена со смыслом и со значением, хотя сама по себе она не является значащей единицей.

Методы синтеза речи

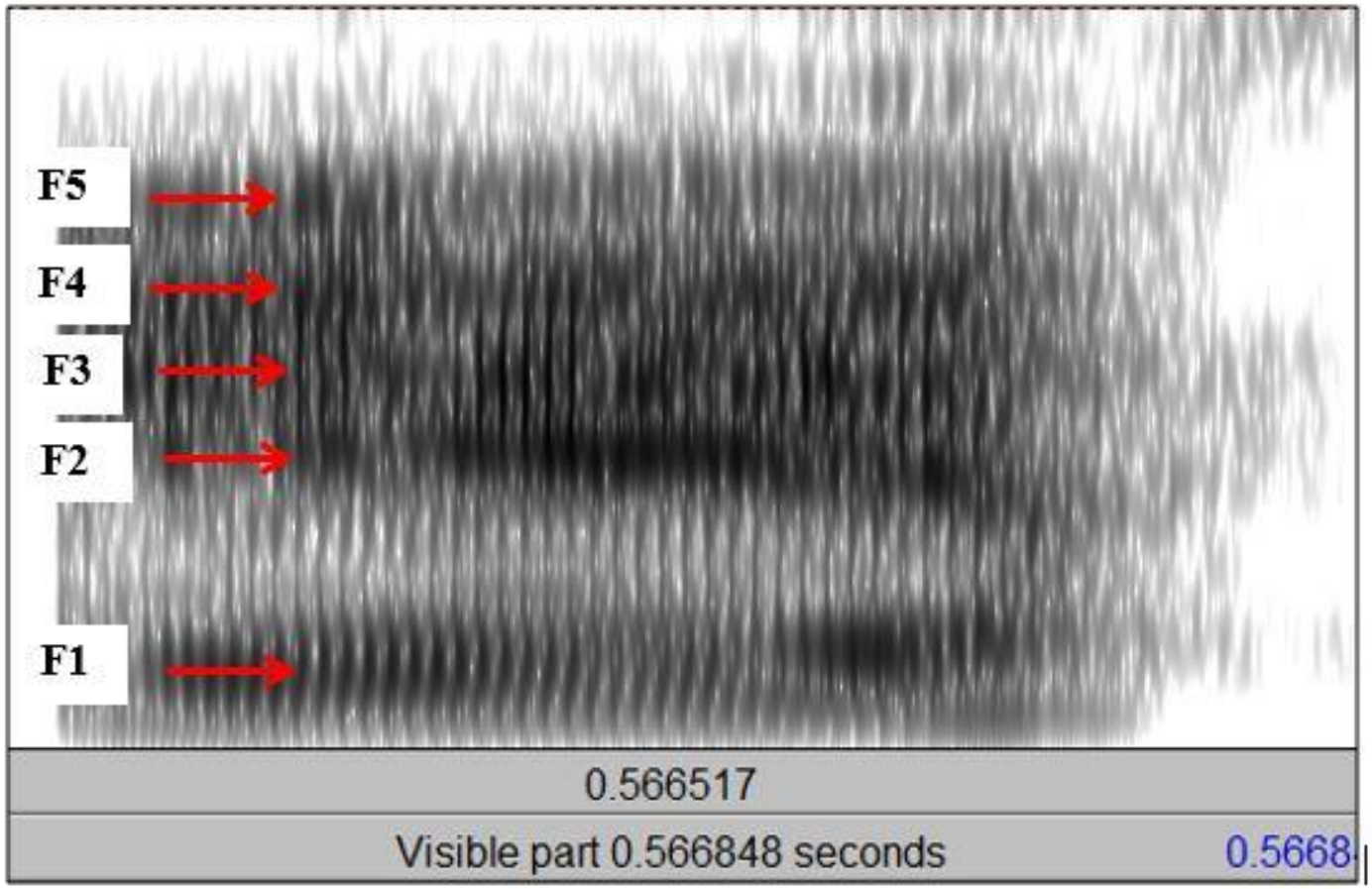
Таблица формант для гласных

- Чтобы синтезировать речевой сигнал, соответствующий определённой фонеме, необходимо настроить центральную частоту каждого полосового фильтра системы на соответствующую частоту форманты.
- Таблица частот формант для некоторых фонем

Фонема	Первая форманта, Гц	Вторая форманта, Гц	Третья форманта, Гц
«и»	270	2300	3000
«е»	400	2000	2550
«а»	660	1700	2400
«у»	640	1200	2400

Спектрограмма гласной [е]

- Видны пять формант – частот с наибольшей энергией

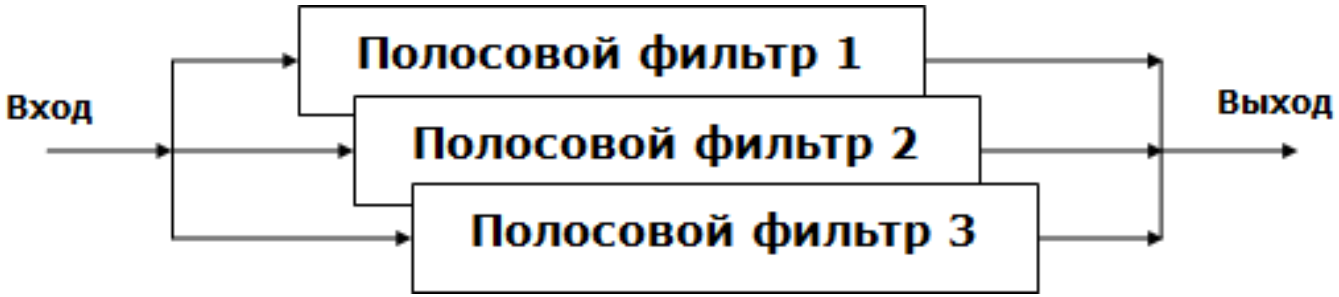


Формантный фильтр

- В основу структуры формантного фильтра заложена упрощённая модель голосового тракта.
- В соответствии с моделью, голосовой тракт представляет собой резонатор с несколькими пиками АЧХ, частоты которых определяют вид произносимой фонемы. Эти пики АЧХ получили название формант.
- Пример спектра фонемы «А»:

Формантный фильтр (2)

- Формантный фильтр создаёт формантные области в спектре входного сигнала с помощью нескольких параллельно соединённых полосовых или фазовых фильтров.
- Количество звеньев в схеме определяет порядок формантного фильтра.
- Схема формантного фильтра третьего порядка:



Определения синтеза речи

- ***Синтез речи*** - это
  - Формирование речевого сигнала по тексту.
  - Искусственное производство человеческой речи.
  - Восстановление формы речевого сигнала по его параметрам.
- Применение:
  - Информирование человека (в т.ч. в системе голосового управления)
  - Акустический диалог человека и компьютера

Способы синтеза речи

- Параметрический синтез
- Компиляционный синтез (конкатенативный, компилятивный) синтез
- Полный синтез по правилам
  - Формантный
  - Артикуляторный
  - На основе записанных отрезков речи

Параметрический синтез

- При параметрическом синтезе звуковой сигнал представлен определённым числом непрерывно изменяющихся параметров.
- Для формирования гласных звуков используется генератор тонального сигнала, для согласных - генератор шума.
- Метод обычно применяют для записи голоса в музыкальных композициях, и чаще речь идет даже не о чистом синтезе голоса, а, скорее, о модуляции.
- Является последней фазой в вокодерных системах.
- Параметрический синтез целесообразно применять в тех случаях, когда набор сообщений ограничен и изменяется не слишком часто.

Компиляционный синтез

- Метод компиляционного синтеза основывается на составлении текстов из заранее записанного "словаря" элементов.
- Размер элемента системы должен быть не менее слова.
- Обычно запас элементов ограничивается несколькими сотнями слов, а содержание синтезируемых текстов - объёмом словаря.
- Используется в различных справочных службах (продажа билетов, прогноз погоды) и технике, требующей оснащения системами речевого ответа: говорящие часы, навигаторы и др.

Полный синтез по правилам

- Полный синтез речи по правилам может воспроизводить речь по заранее неизвестному тексту. Базируется на запрограммированных лингвистических и акустических алгоритмах. Элементы человеческой речи не используются.
- Реализуется путём моделирования речевого тракта, применения аналоговой или цифровой техники. Причём в процессе синтеза значения параметров и правила соединения фонем вводят последовательно через определённый временной интервал, например 5—10 мс.
- Подходы:
  - ***Формантный метод*** – базируется на формантах - частотных резонансах речевой акустической системы. Моделируется работа речевого тракта человека, работающего как набор резонаторов. Универсальная и перспективная технология, но понимание результата синтеза требует подготовки.
  - ***Артикуляторный метод*** – пытается доработать недостатки формантного путем добавления в модель фонетических особенностей произнесения отдельных звуков.

Синтез речи по правилам на основе отрезков речи

- Базисные единицы речи:
  - Микросегменты;
  - Аллофоны;
  - Дифоны – участки речевого сигнала, включающие в себя переходы между звуками;
  - Полуслоги – сегменты, содержащие половину согласного и половину примыкающего к нему гласного;
  - Слоги;
  - Единицы произвольного размера.
- Есть возможность синтеза речи по не заданному заранее тексту (чтение книги на лету).
- Трудно управлять интонационными характеристиками речи, так как характеристики отдельных слов могут изменяться в зависимости от контекста или типа фразы.
- Качество синтезированной речи несопоставимо с качеством речи естественной (на границах сшивки элементов могут возникать искажения).

Распознавание речи

Основные задачи

Некоторые технологии

Основные задачи

- Системы Interactive Voice Response (IVR) в колл-центрах
  - Биометрическая идентификация
  - Автоматическая маршрутизация звонка
  - Аналитика речи, поиск пауз, тормозов в интерфейсе, аналитика эмоционального состояния
  - Сбор оценок операторов колл-центра
- Персональные помощники
  - Распознавание поисковых запросов
  - Распознавание команд управления

Голосовая биометрия

- 4 фактора:
  - "кто вы"
  - "что у вас есть"
  - "что вы знаете"
  - "что вы делаете"
- Пассивный режим, пассивные системы – не зависят от текста, не проявляют себя (слушают).
- Активный режим, активные системы – зависят от текста, взаимодействуют с пользователем.

Зачем нужна голосовая биометрия?

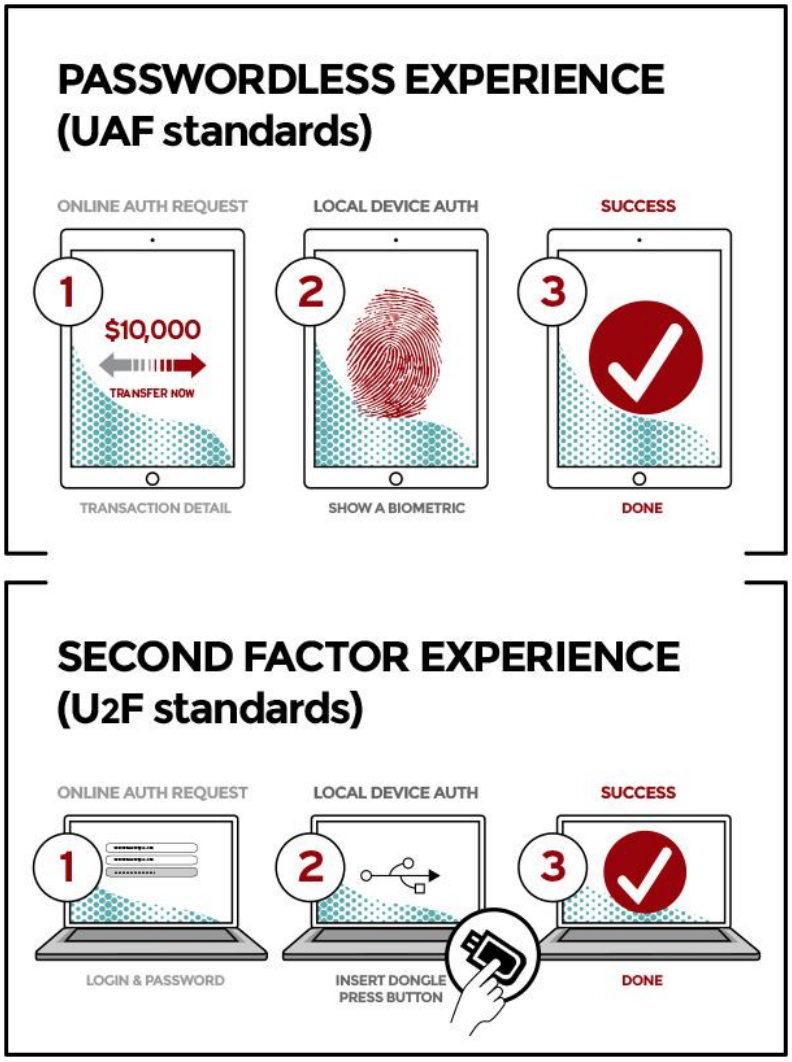
- Сокращение времени на аутентификацию пользователя с 23 секунд в ручном режиме в центре обработки вызовов (Call Center) до 5 секунд в автоматическом.
- Повышение лояльности пользователей (и, как следствие, доходов от них) в результате отказа от необходимости запоминать всем известные ответы на "секретные" вопросы, помнить PIN-код для входа в систему или отвечать на вопросы назойливого сотрудника банка (ваши ФИО, дата вашего рождения, номер карты и т.п.).
- Снижение числа сотрудников центра обработки вызовов за счет автоматической обработки многих простых вопросов (время работы офиса в праздники, ближайший офис или банкомат, тарифы и т.п.).
- Снижение числа мошеннических операций.
- Снижение времени на ожидание правильного сотрудника, который поможет ответить звонящему.
- Рост продуктивности работников компании и центра обработки вызовов.

Поставщики решений голосовой биометрии

- Некоторые поставщики голосовой биометрии:
  - Agnitio
  - Auraya Systems
  - Authentify
  - KeyLemon
  - Nuance
  - ValidSoft
  - Verint Systems
  - VoiceTrust
  - VoiceVault
  - ЦРТ - Центр речевых технологий (https://www.speechpro.ru/)

Стандарты аутентификации FIDO Alliance

- FIDO – Fast IDentity Online
- UAF – Universal Authentication Framework
- U2F – Universal 2nd Factor (U2F) authentication
- https://security.stackexchange.com/questions/71590/what-are-the-differences-between-the-u2f-and-uaf-fido-authentication-standards



Распознавание говорящего по свободной речи

- Пассивный режим
- Идентификация мошенников в режиме тихого прослушивания звонящего/говорящего и идентифицирующие его речь, ничем себя не выдавая.
- Поэтому пассивные системы проще в использовании, но и требуют больших ресурсов для своей реализации.

Распознавание говорящего по заранее определённым фразам

- Активный режим (система «выдаёт» себя, управляя диалогом)
- Требуют большего участия пользователя
- Сложно идентифицировать мошенника

Голосовой отпечаток

- *Голосовой отпечаток* – некая уникальная для человека запись, характеризующая голос в целом. Длительность записи 1-2 минуты.
- Несколько различных голосовых отпечатков формируют профиль голоса.
- При создании отпечатка могут использоваться опорные точки в речи (переходы между звуками), высота звуков, акцентированные звуки, темп речи; косвенно учитываются физиологические особенности звукового тракта, горла, глотки, носа; особенности произношения слов и звуков, физические характеристики голоса.
- Распознавание на основе отпечатка может длиться 5-15 секунд.

Интеллектуальные голосовые помощники

- Интеллектуальные голосовые помощники:
  - Алиса (Yandex)
  - iOS Siri (Apple/Nuance)
  - Google Assistant
  - Amazon Alexa

Речевые API

- Yandex Speech Kit

  https://tech.yandex.ru/speechkit/
  - Распознавание речи
  - Анализ речи (биометрическая информация)
  - Синтез речи
- Google Speech API

  https://cloud.google.com/speech/
  - Распознавание речи (110 диалектов языков)
  - Фильтрация неуместной лексики
  - Контекстно-зависимое распознавание

Yandex SK

- Пример удачного распознавания:

```
<recognitionResults success="1">
    <variant confidence="0.69">твой номер 212-85-06</variant>
    <variant confidence="0.7">твой номер 213-85-06</variant>
</recognitionResults>
```

- Пример неудачного распознавания:

```
<recognitionResults success="0"/>
```

Yandex SK (2)

- Оценка биометрических параметров: пол, возраст, язык
- Возрастная группа задается буквой латинского алфавита:
  - 'c' — ребенок (до 14 лет). Для этой группы пол не указывается;
  - 'y' — подросток (14-20 лет);
  - 'a' — взрослый (20-55 лет);
  - 's' — пожилой (старше 55 лет).
- Пол задается буквой 'm' (male) или 'f' (female).

Yandex SK (3)

- Пример результатов распознавания биометрии

```
{
tag: 'gender', class: 'female',
confidence: 0.3632163107395172
}
{
tag: 'language', class: 'ru',
confidence: 0.8913142085075378 // --> Наиболее
   вероятный язык речи - русский.
}
```

Yandex SK (4)

- Настройки синтеза фразы:
  - Диктор
  - Эмоции в голосе
- Например:

```
tts.speak(
    'Меня зовут Вася',
    {
        speaker: 'zahar',
        emotion: 'neutral',
        stopCallback: function () {...}
    }
)
```

Google Speech API

- Пример распознавания:


```
{
        "result":[
        {
                "alternative":[
                {
                        "transcript":"this is a test",
                        "confidence":0.97321892
                },
                {
                        "transcript":"this is a test for"
                }
                ],
                "final":true
        }
        ],
        "result_index":0
}
```

Что почитать

- Вокодер https://eomi.ru/electronic/vocoder/
- Учебник по фонетике русского языка http://www.speech.nw.ru/Manual/menu.html
- Синтезатор речи с открытым исходным кодом RHVoice http://tiflo.info/rhvoice/
- Алиса. Как Яндекс учит искусственный интеллект разговаривать с людьми https://habrahabr.ru/company/yandex/blog/339638/
  - Распознавание речи от Яндекса. Под капотом у Yandex.SpeechKit https://habrahabr.ru/company/yandex/blog/198556/