Обработка речевых сигналов

Блок 2. Автоматическое распознавание речи

Максим Кореневский
Старший научный сотрудник ООО «ЦРТ»,
к.ф.-м.н.

Настоящий блок лекций подготовлен при поддержке «ЦРТ | Группа компаний»

ЦРТ | ГРУППА КОМПАНИЙ

Блок 2. Автоматическое распознавание речи (Automatic Speech Recognition, ASR)

Контакты

Максим Кореневский

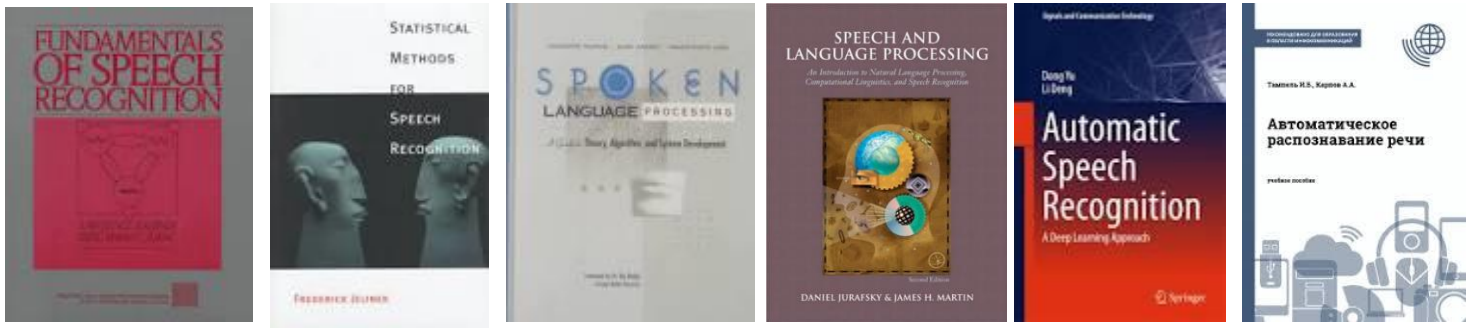
E-mail: maxim_korenevski@mail.ru

ЦРТ | ГРУППА КОМПАНИЙ

https://www.speechpro.ru/

Литература по автоматическому распознаванию речи

- L.Rabiner Fundamentals of Speech Recognition, 1993
- F.Jelinek Statistical Methods for Speech Recognition, 1997
- X.Huang, A.Acero, H.-W.Hon Spoken Language Processing, 2001
- D.Jurafsky, J.H.Martin Speech and Language Processing, 2000 (1st ed.), 2009 (2nd ed.), 2021 (3rd ed. draft)
- D.Yu, L.Deng Automatic Speech Recognition: A Deep Learning Approach, 2014
- И.Б.Тампель, А.А.Карпов Автоматическое распознавание речи, 2016



Содержание текущего блока лекций

- Часть 1. Введение в автоматическое распознавание речи
- Часть 2. Структура традиционной системы распознавания
- Часть 3. Системы распознавания речи на основе GMM-HMM
- Часть 4. Традиционные системы распознавания речи на основе нейронных сетей
- Часть 5. End-to-end подходы к распознаванию речи
- Часть 6. Semi-supervised и self-supervised системы

Часть 1. Введение в ASR

План лекции

- Что такое речь?
- Типы систем распознавания речи и сценарии их использования
- Метрики оценки качества систем распознавания речи
- Трудности при создании систем распознавания речи
- Акустические признаки речи
- Система распознавания речи на основе сравнения с эталоном

План лекции

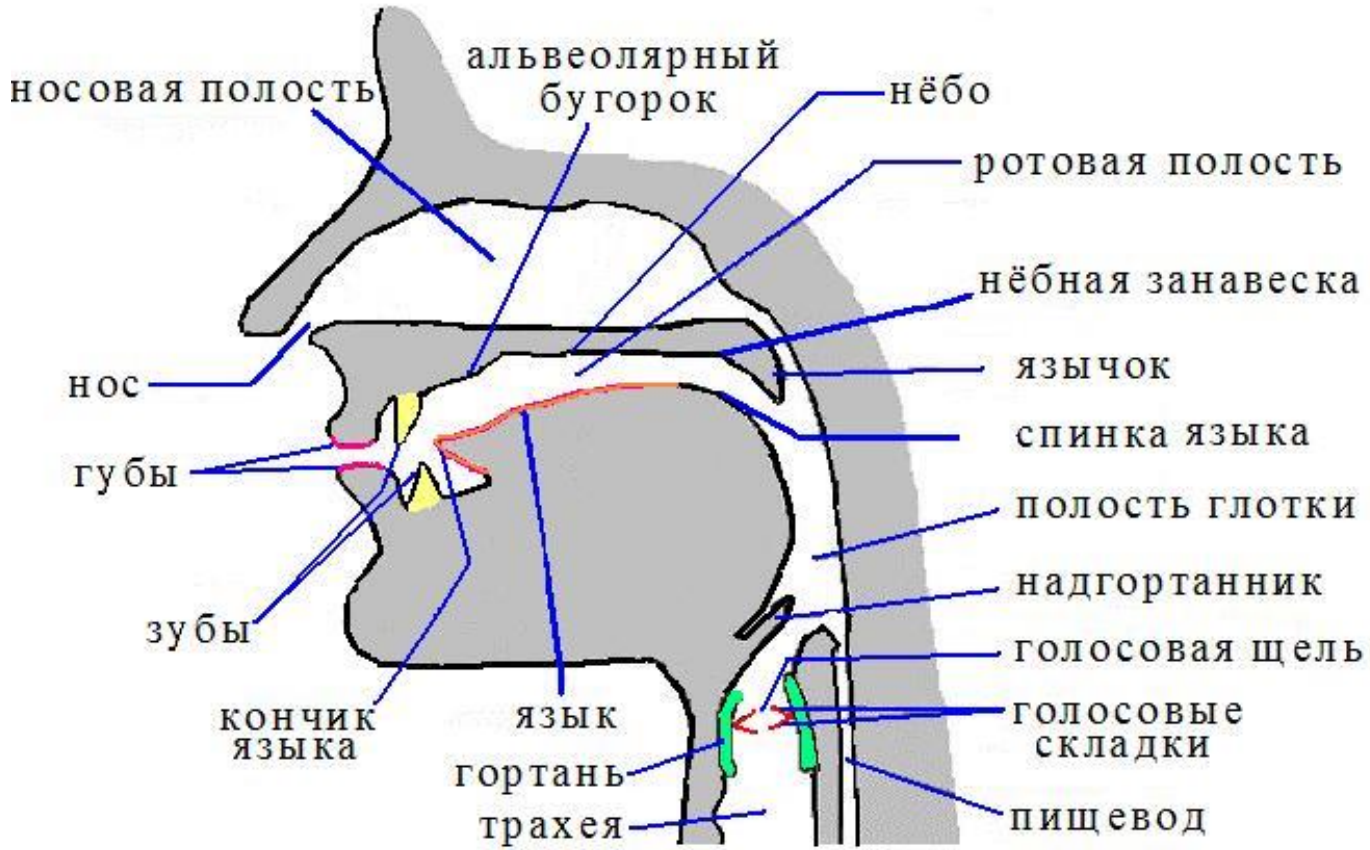
- Что такое речь?
- Типы систем распознавания речи и сценарии их использования
- Метрики оценки качества систем распознавания речи
- Трудности при создании систем распознавания речи
- Акустические признаки речи
- Система распознавания речи на основе сравнения с эталоном

Что такое речь?

- Лингвистика: средство человеческого общения, последовательность произнесенных слов с интонационной и смысловой нагрузкой
- Фонетика: последовательность звуков, генерируемых органами артикуляции
- Физика/акустика: звук (продольная волна сжатия/разрежения воздуха), производимый органами артикуляции и воспринимаемый органами слуха
- Математика: функция, реализация нестационарного случайного процесса
- Информатика: оцифрованный звуковой сигнал, записанный на микрофон

Речеобразование

Органы речеобразования (голосовой тракт):



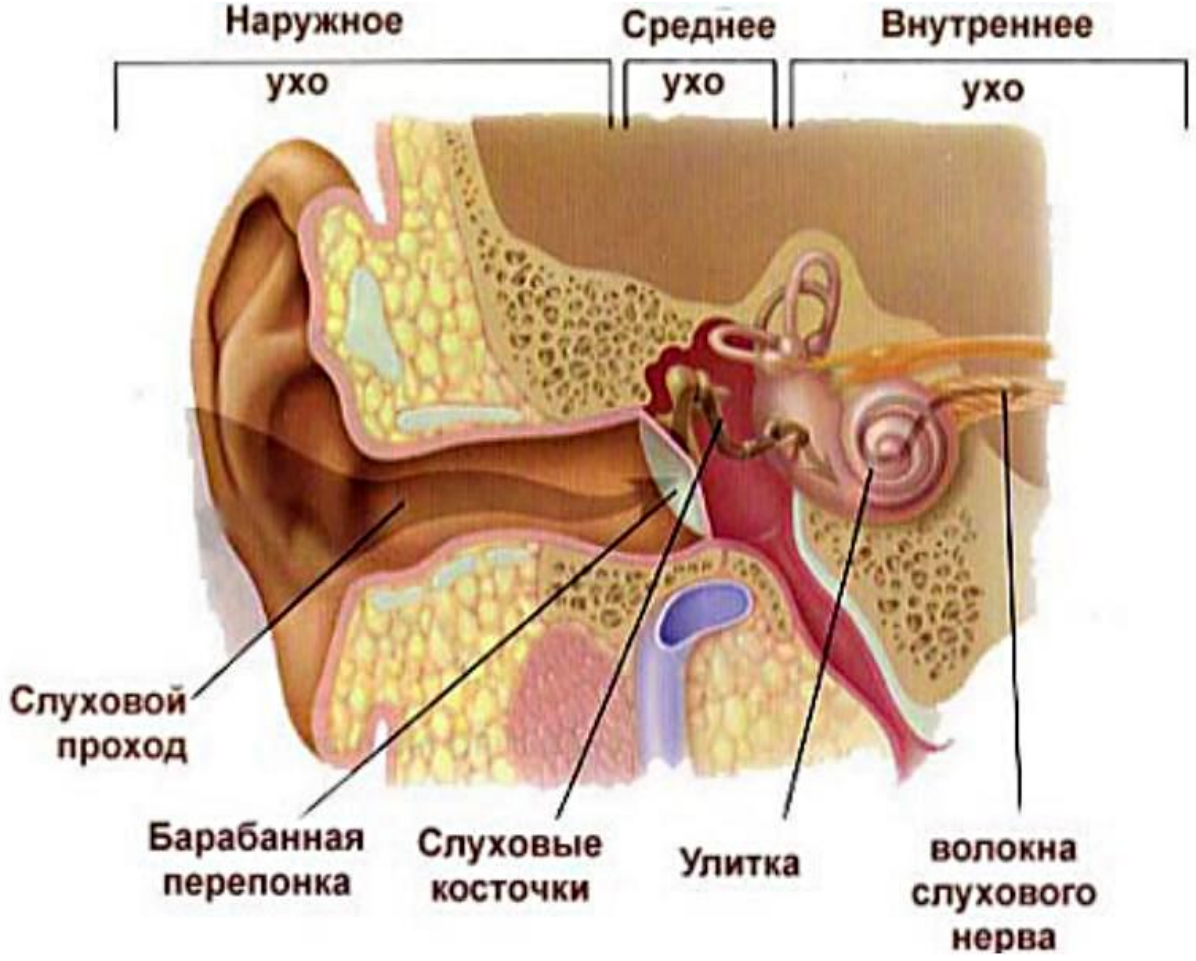
Речеобразование

Характеристики голосового тракта:

- Основной тон (fundamental frequency, pitch, F0) - частота колебаний голосовых связок
- Когда голосовые связки задействованы - звонкая (voiced) речь
- Для женщин F0 = 100-300 Гц, для мужчин F0 = 50-150 Гц
- Обертона - частоты, кратные F0
- Форманты (F1, F2, ...) - резонансные частоты голосового тракта
- Форманты присутствуют как в voiced, так и в unvoiced-речи.

Слуховое восприятие

Слуховая система человека:



Слуховое восприятие

Громкость звука:

- Диапазон воспринимаемых частот: 20 Гц - 20000 Гц
- Ниже - инфразвук, выше – ультразвук
- Звук одной фиксированной частоты ω – тональный сигнал (тон)

$$f(t) = a\cos\omega t + b\sin\omega t = A\cos(\omega t + \varphi) = \Re\{Fe^{i\omega t}\}$$

$A = |F|$ - амплитуда, $\varphi = \arg(F)$ - фаза

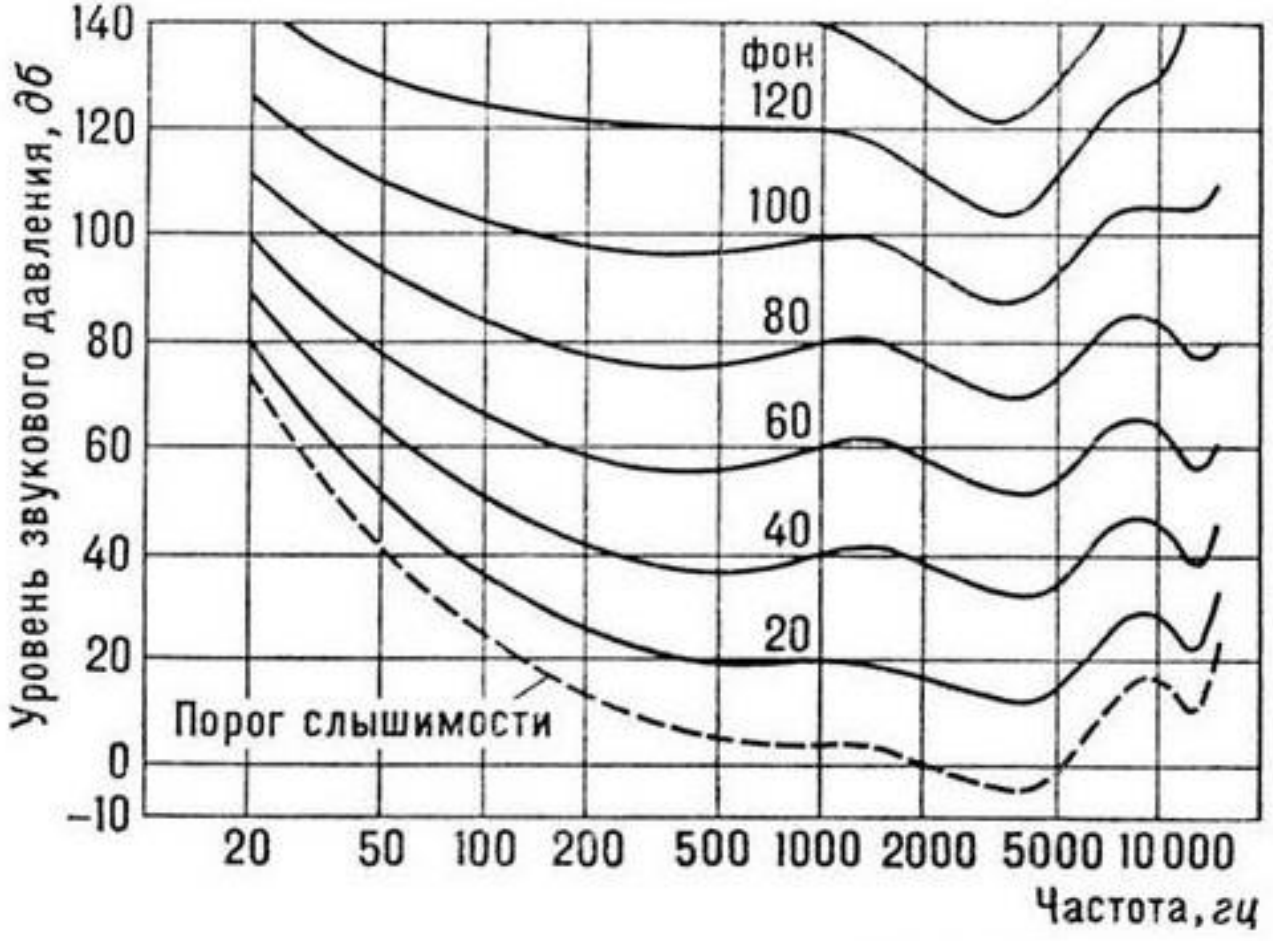
- Фаза слабо влияет на восприятие речи
- Громкость звука (измеряется в децибелах, дБ, dB или фонах): $10\log_{10}\frac{P}{P_0}$

P – энергия звуковой волны,

P_0 – минимальная энергия звука, воспринимаемая человеком (порог слышимости)

Слуховое восприятие

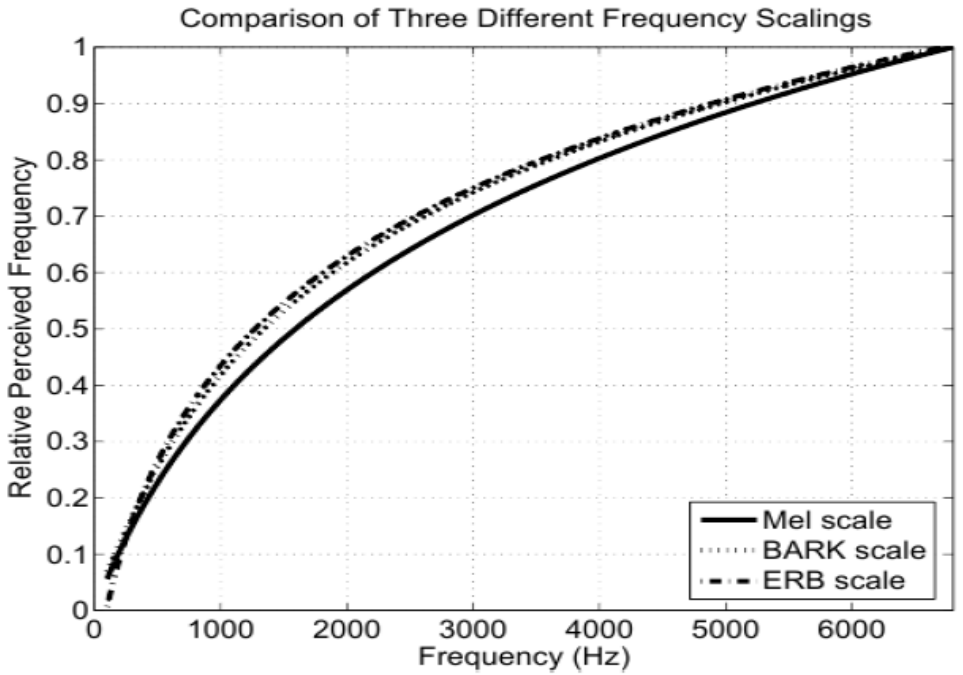
Кривые равной громкости:



Слуховое восприятие

Нелинейность частотной шкалы:

- Изменение ноты на 1-2-3-… «октавы» = увеличение частоты в 2-4-8-… раз
- Везде, кроме низких частот, воспринимаемая «высота» звука зависит от частоты почти ЛОГАРИФМИЧЕСКИ
- $$Mel(f) = 2595\log\left(1 + \frac{f}{700}\right)$$
- На низких частотах – квазилинейна
- Критические полосы слуха: два тона с частотой в пределах полосы неразличимы на слух.
- Ширина полосы тоже растет ~ логарифмически



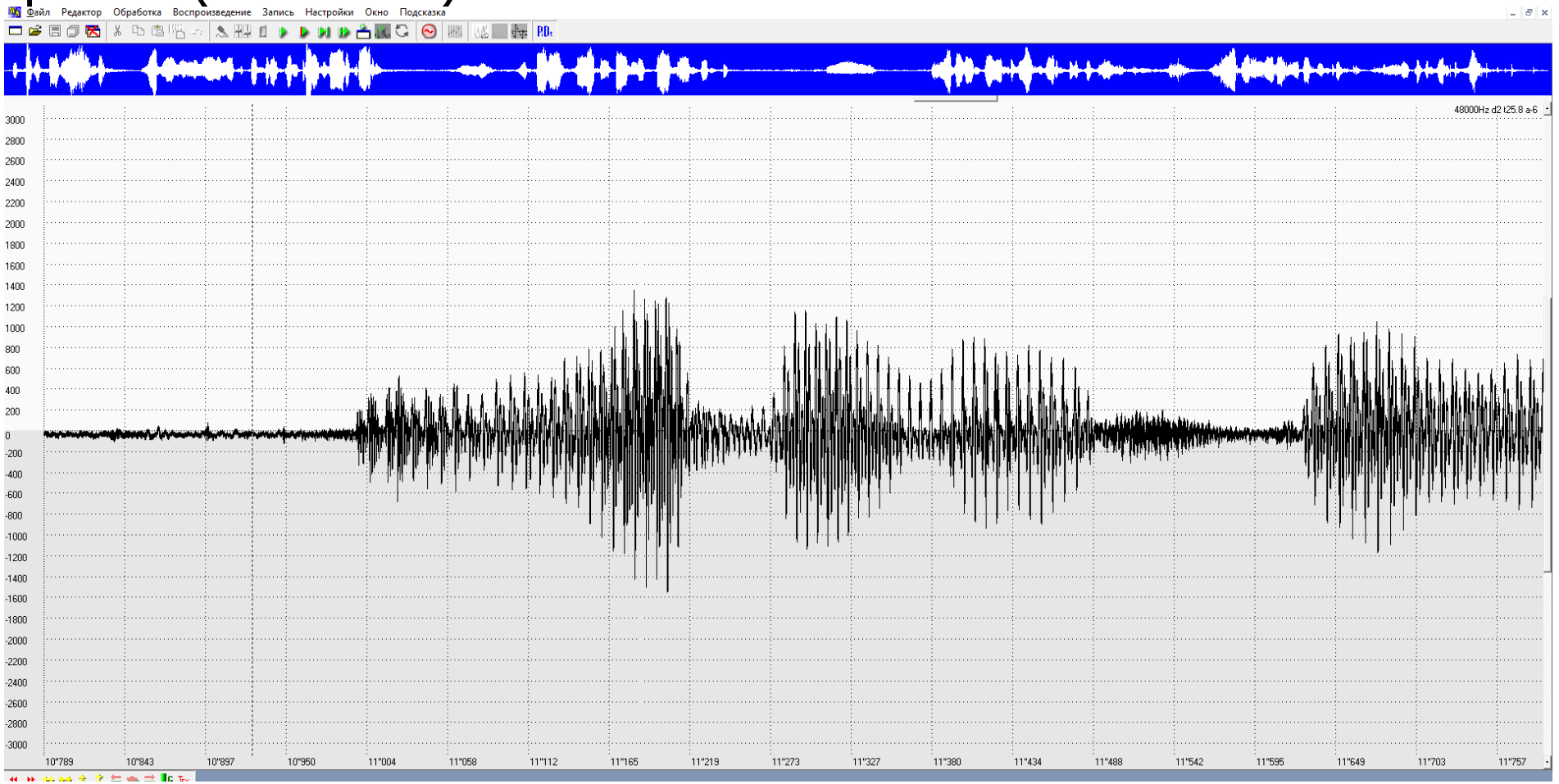
Математика

Речь с точки зрения математики:

- Речь - реализация нестационарного случайного процесса
- Факторы случайности:
  - внутридикторская и междикторская вариативность
  - турбулентность воздушного потока
  - флуктуации среды передачи звука и устройств записи
- Интервал квазистационарности - 10-25 мс
- Обычно речь обрабатывают кадрами (фреймами) по 15-25 мс с шагом ~10 мс

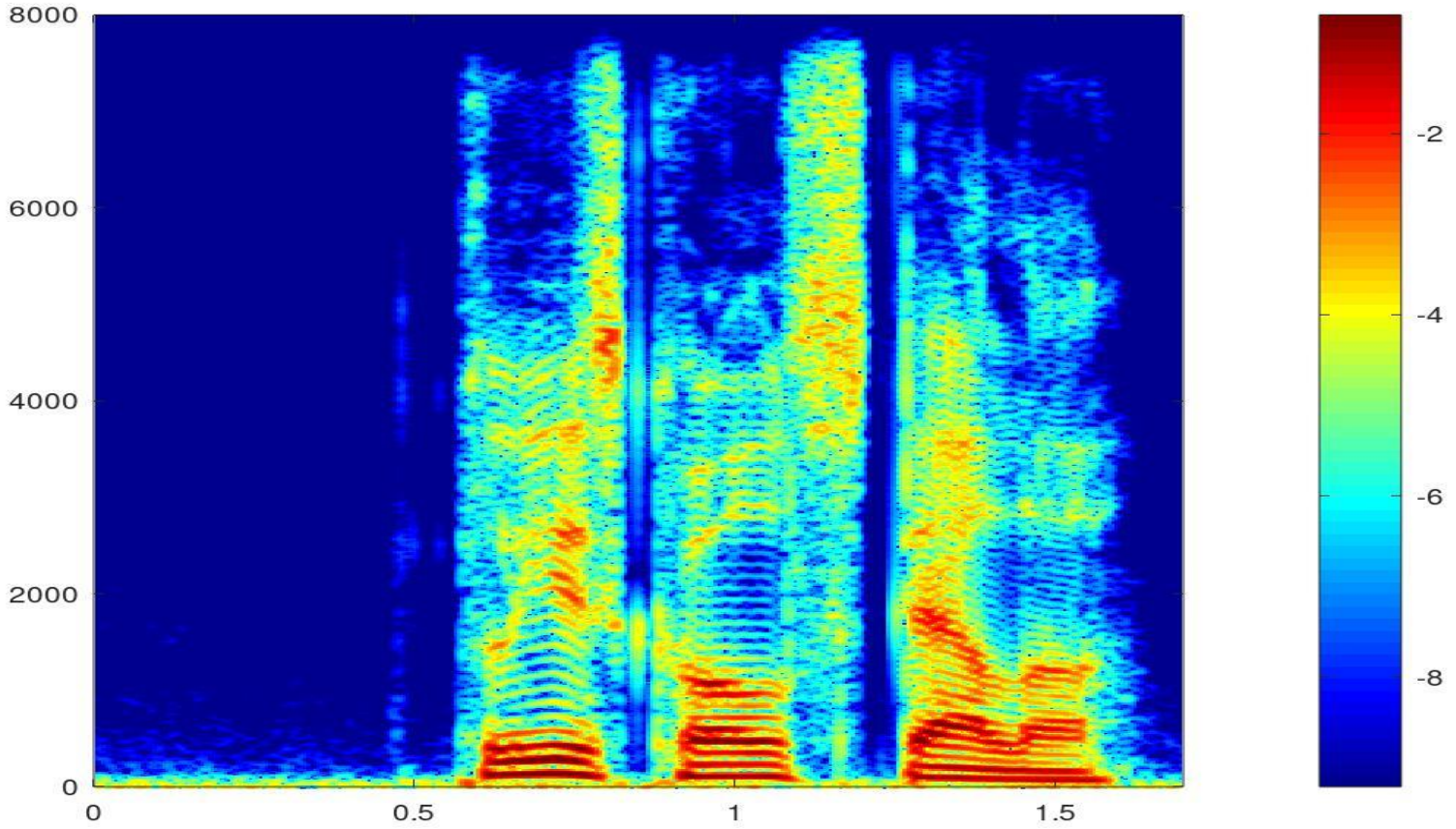
Математика

Осциллограмма (waveform):



Математика

Спектрограмма (частотно-временно́е представление):



План лекции

- Что такое речь?
- Типы систем распознавания речи и сценарии их использования
- Метрики оценки качества систем распознавания речи
- Трудности при создании систем распознавания речи
- Акустические признаки речи
- Система распознавания речи на основе сравнения с эталоном

Задачи распознавания речи

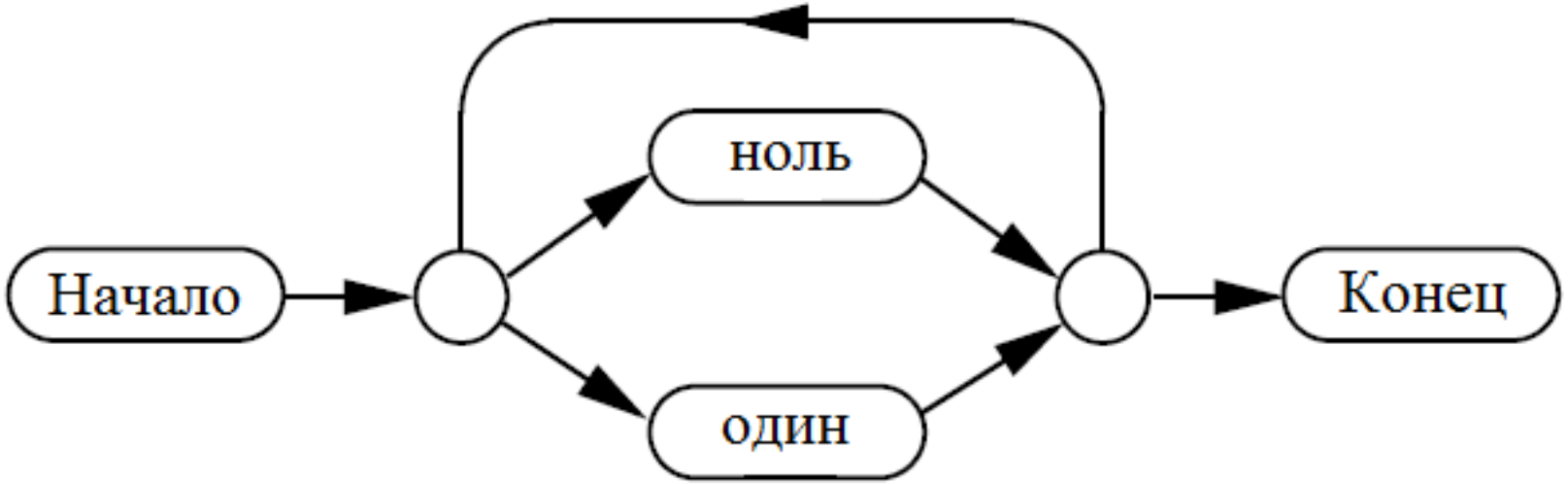
Распознавание фиксированного набора слов/фраз

Слово/фраза	Оценка (score)
Здравствуйте!	30
До свидания	5
Как тебя зовут?	95
Меню, пожалуйста	10
…	

Задачи распознавания речи

Распознавание по грамматике:

- Грамматика определяет допустимые последовательности слов
- В грамматике могут быть ветвления и циклы (петли)
- Можно сопоставлять определенным путям конкретные действия
- Существуют стандарты описания грамматик (например SRGS)



Задачи распознавания речи

Распознавание слитной речи:

- Не накладывается никаких ограничений на последовательность слов
- Приходится учитывать ограничения, существующие в самом языке
- Большой размер «словаря распознавания»

Однажды в студеную зимнюю пору….
Однажды в студеную зимнюю гору…
Однажды в студеную зиму набору…
Однажды в суденышке мимо забора…
…..

Задачи распознавания речи

Поиск ключевых слов:

– Я хочу купить билеты из Москвы в Санкт-Петербург на Сапсан

Билеты: с 1.13 с. по 1.78 с., уверенность 0.93
Поезд: с 2.22 с. по 2.80 с., уверенность 0.13
Вокзал: с 3.40 с. по 3.96 с., уверенность 0.32
Сапсан: с 3.72 с. по 4.38 с., уверенность 0.98
….

Приложения систем распознавания речи

Диктовка:

- документы
- электронные письма
- заметки и т.д.

Приложения систем распознавания речи

Расшифровка:

- стенограммы
- лекции
- телефонные переговоры

Приложения систем распознавания речи

Системы поиска ключевых слов

- Акустический поиск
  - Малый словарь, работает online
  - Приложения: следственные действия, борьба с терроризмом, голосовое управление, системы «умный дом», контроль качества обслуживания в офисах продаж
- Индексированный поиск
  - Произвольный словарь, работает с большими корпусами речевых данных, создает «индекс» для быстрого поиска
  - Приложения: поиск в базах речевых документов (фильмы, переговоры, лекции и т.д.)

Приложения систем распознавания речи

Распознавание по грамматикам

- IVR-системы
  - Банки
  - Контакт-центры
  - Киоски голосового самообслуживания
- Контроль переговоров, соблюдение регламента
  - Употребление определенных речевых конструкций
  - Контроль использования ненормативной лексики
  - и т.д.

Приложения систем распознавания речи

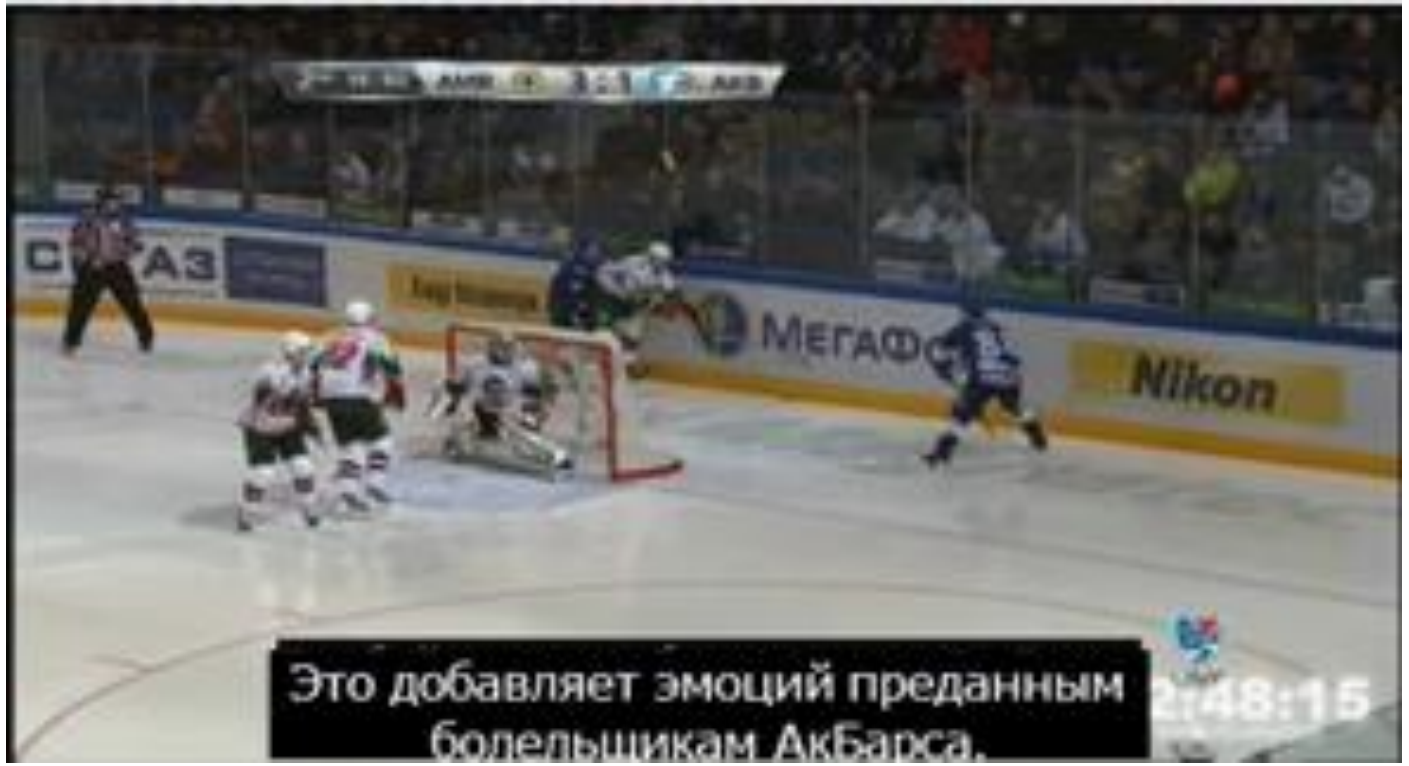
NLU-системы (Natural Language Understanding)

- Классификация речевых сообщений по тематике
- Извлечение смысла речевого сообщения
- Диалоговые системы
- Голосовые помощники (Google Siri, Amazon Alexa, Яндекс Алиса и т.п.)
- и т.д.

Приложения систем распознавания речи

Прочие приложения

- Автоматическая подготовка субтитров
- Разметка и аннотирование медиа-баз



План лекции

- Что такое речь?
- Типы систем распознавания речи и сценарии их использования
- Метрики оценки качества систем распознавания речи
- Трудности при создании систем распознавания речи
- Акустические признаки речи
- Система распознавания речи на основе сравнения с эталоном

Оценка качества/сравнение систем распознавания

Распознавание по грамматикам:

- Оценивается точное распознавание всей фразы/последовательности слов
- Естественная мера качества: SER (string/sentence error rate) – доля неверно распознанных фраз
- Вычисляется в процентах:

$$SER = \frac{\#\text{неверно распознанных фраз}}{\#\text{распознаваемых фраз}} * 100\%$$

Оценка качества/сравнение систем распознавания

Распознавание слитной речи:

- Выравнивание (по Левенштейну):
  - Эталон: Мой дядя, самых честных правил, когда не в шутку занемог…
  - Распознано: Мой дятел самых честь не правил когда в шутку за не мог
  - Замены (substitutions), Вставки (insertions), Удаления (deletions)
- Важно: выравнивание должно минимизировать суммарное количество ошибок
- WER (Word Error Rate) – пословная ошибка распознавания, измеряется в процентах
- Accuracy – точность распознавания

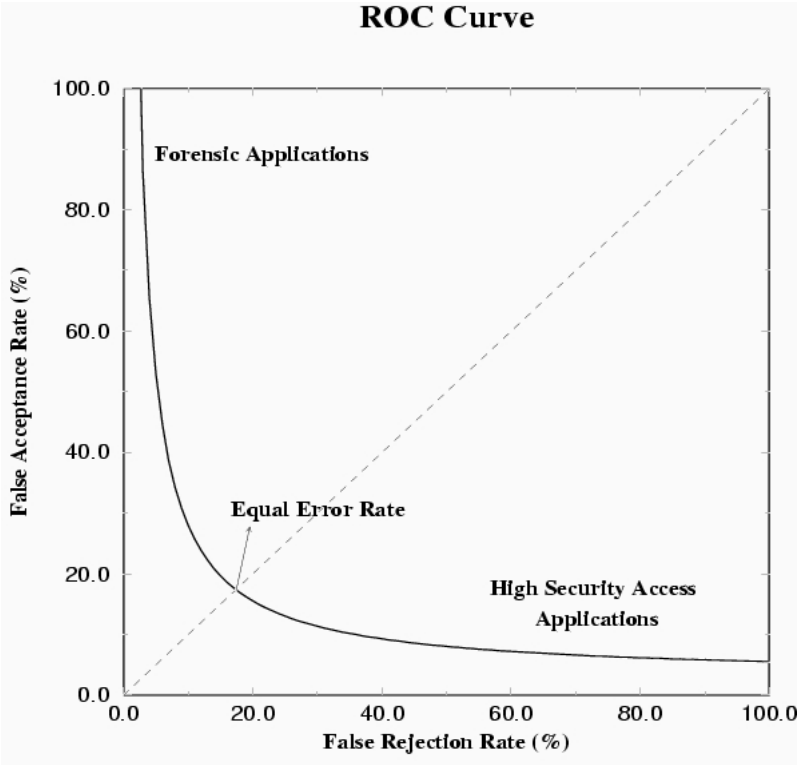
$$WER = \frac{\#\text{замен} + \#\text{вставок} + \#\text{удалений}}{\#\text{слов в эталоне}} * 100\%$$

$$Accuracy = 100\% - WER$$

Оценка качества/сравнение систем распознавания

Поиск ключевых слов:

- Метрики FR(false rejection) и FA (false acceptance/false alarm)
  - Эталон: Мой дядя, самых честных правил, когда не в шутку занемог…
  - Слова для поиска: дядя, тетя, когда, утка
  - Найдено: дядя, утка; Верно найдено: дядя
  - Ложный пропуск (false rejection): когда
  - Ложное срабатывание (false acceptance): утка
  - $FR = \frac{\#\text{ ложных пропусков}}{\#\text{ ключевых слов во фразе}} * 100\%,$
  - $FA = \frac{\#\text{ ложных срабатываний}}{\#\text{ НЕключевых слов во фразе}} * 100\%.$



План лекции

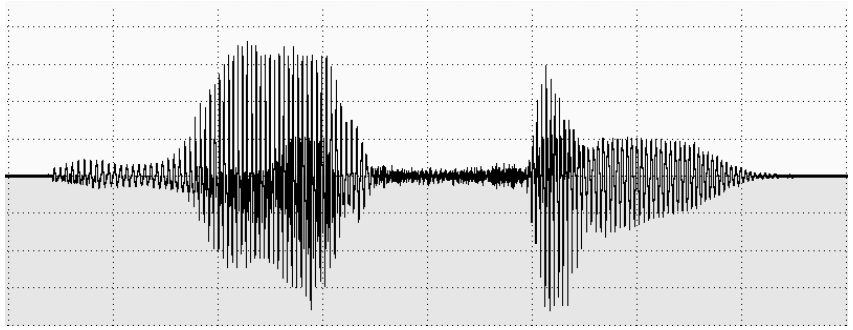
- Что такое речь?
- Типы систем распознавания речи и сценарии их использования
- Метрики оценки качества систем распознавания речи
- Трудности при создании систем распознавания речи
- Акустические признаки речи
- Система распознавания речи на основе сравнения с эталоном

Трудности разработки систем распознавания речи

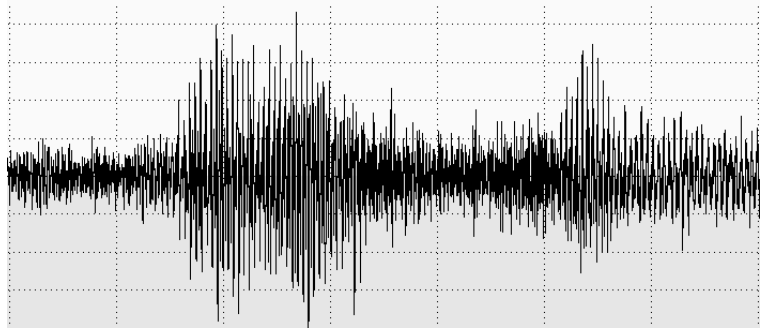
Разнообразие условий записи:

- Окружающие шумы и помехи
- Влияние канала и среды передачи звука (микрофон, стационарный/мобильный телефон)
- Реверберация (переотражения от стен помещения и предметов)
- Частота дискретизации (8000, 11025, 16000, 22050, 44100 Гц)
- Квантование и кодирование
- Клиппирование

Чистая речь («восемь»)



С шумом кафе (SNR=0dB)



Трудности разработки систем распознавания речи

Междикторская и внутридикторская вариативность:

- Разнообразие голосов (пол, возраст, социальное положение, образование)
- Различные региональные акценты («оканье» и т.п.)
- Дефекты речи (картавость, шепелявость и т.д.)
- Эмоциональное состояние (безразличие, гнев, радость, возбуждение …)
- Физическое состояние (усталость, простуженность/охриплость …)

Трудности разработки систем распознавания речи

Разнообразие стилей речи:

- Распознавание последовательностей цифр – 10 слов
- Распознавание имен и фамилий – сотни слов
- Распознавание новостей – тысячи слов
- Распознавание общей лексики – сотни тысяч слов
- Размер «эффективного» словаря зависит от языка:
  - Для английского языка 99% текстов покрываются 65 тыс. слов
  - Для русского языка 99% текстов покрываются ~500 тыс. **словоформ**
- Размер словаря растет из-за богатой морфологии языка, в частности
  - Флективности (изменение окончаний по падежам/родам/числам)
  - Агглютинативности (добавление разных префиксов/аффиксов уточняет значение)

Трудности разработки систем распознавания речи

Размеры словаря:

- Подготовленная (продуманная) речь
- Чтение текста
- Спонтанная речь
  - Различный темп
  - «Проглатывание» окончаний слов
  - Повторы слов, куски слов, «само-исправления»
  - Паузы хезитации («ээээ», «мм»)
  - Слова-паразиты и междометия, нарушающие естественный порядок слов

План лекции

- Что такое речь?
- Типы систем распознавания речи и сценарии их использования
- Метрики оценки качества систем распознавания речи
- Трудности при создании систем распознавания речи
- Акустические признаки речи
- Система распознавания речи на основе сравнения с эталоном

Акустические признаки речи

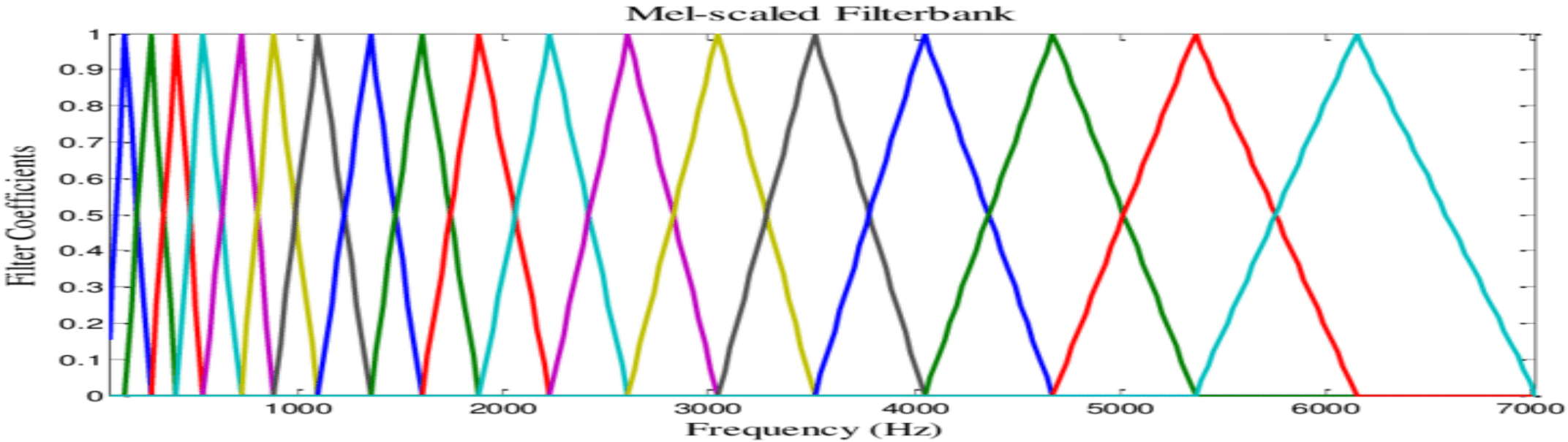
Требования к признакам:

- Должны характеризовать текущий произносимый звук
- Для разных произнесений одного и того же звука должны быть близки
- Для произнесений различных звуков должны сильно отличаться
- Должны иметь достаточно небольшую размерность
- **Желательно**: должны быть устойчивы к
  - Смене говорящего
  - Изменению громкости и темпа речи
  - Расстоянию до микрофона
  - Шумам и реверберации

Акустические признаки речи

Мел-частотные кепстральные коэффициенты:

- Вычислить кратковременный Фурье-спектр (short-time Fourier Transform, STFT)
- Найти спектр мощности
- Взвесить треугольными mel-фильтрами



- Вычислить логарифм. То, что получилось, часто называют log-mel-fbanks
- Взять дискретное косинус-преобразование (DCT)

План лекции

- Что такое речь?
- Типы систем распознавания речи и сценарии их использования
- Метрики оценки качества систем распознавания речи
- Трудности при создании систем распознавания речи
- Акустические признаки речи
- Система распознавания речи на основе сравнения с эталоном

Сравнение с эталоном

Постановка задачи. Общая идея.

- Хочется распознавать небольшое количество фиксированных слов/фраз
- Разработчик системы записывает по несколько экземпляров каждого слова - эталоны
- В test-time система «сравнивает» записанный звук с каждым из эталонов
- Слово, соответствующее ближайшему эталону, – результат распознавания!
- Главный вопрос: а как сравнивать две фонограммы?
  - Вычислим признаки для каждой из фонограмм (например MFCC – 13 мерные векторы)
  - Векторы можно сравнивать друг с другом, например с помощью Евклидова расстояния

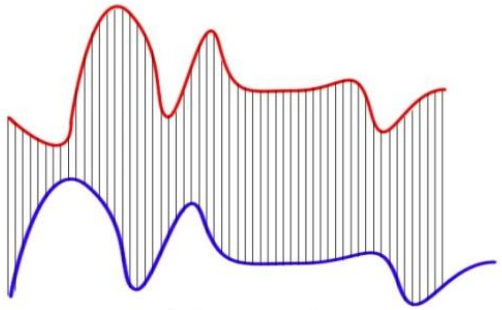
$$d(X,Y) = \sqrt{\sum_i (x_i - y_i)^2}$$

  - Но как сравнить две последовательности векторов разной длины?

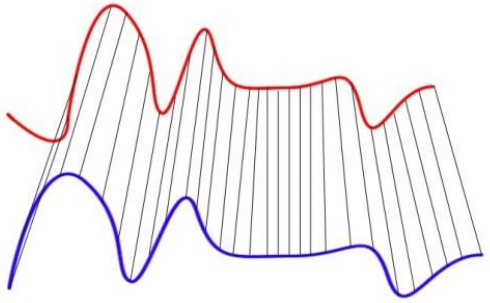
Сравнение с эталоном

Алгоритм Dynamic Time Warping (DTW)

- Идея: «деформировать» шкалу времени для каждой из фонограмм, так чтобы минимизировать суммарное отклонение признаков двух фонограмм



Euclidean Matching



Dynamic Time Warping Matching

- Формально: найти такие последовательности индексов, $\{i_k\}$: $i_1 = 1,\ i_N = T_1,\ \ i_k \le i_{k+1} \le i_k + 1$ и $\{j_k\}$: $j_1 = 1,\ j_N = T_2,\ \ j_k \le j_{k+1} \le j_k + 1$ (выравнивание), что

$$D = \sum_{k=1}^{N} d(X^{i_k}, Y^{j_k}) \to \min$$

Сравнение с эталоном

Алгоритм DTW (продолжение)

- Алгоритм динамического программирования (Bellman).
- Введем вспомогательную функцию $D(i,j)$ - расстояние от первых i кадров последовательности X до первых j кадров последовательности Y.
- Для нее справедлив простой рекурсивный способ вычисления:
  - $D(1,1) = d(X^1, Y^1)$, $D(1,j) = D(1,j-1) + d(X^1, Y^j)$, $j > 1$, $D(i,1) = D(i-1,1) + d(X^i, Y^1)$, $i > 1$
  - $D(i,j) = \min(D(i-1,j), D(i,j-1), D(i-1,j-1)) + d(X^i, Y^j)$, $i,j > 1$
  - $D = D(T_1, T_2)$.
- Иллюстрация:

	-2	10	-10	15	-13	20	-5	14	2
3	5	12	25	37	53	70	78	89	90
-13	16	28	15	43	37	70	78	105	104
14	32	20	39	16	43	43	62	62	74
-7	37	37	23	38	22	49	45	66	71
9	48	38	42	29	44	33	47	50	57
-2	48	50	46	46	40	55	36	52	**54**

Сравнение с эталоном

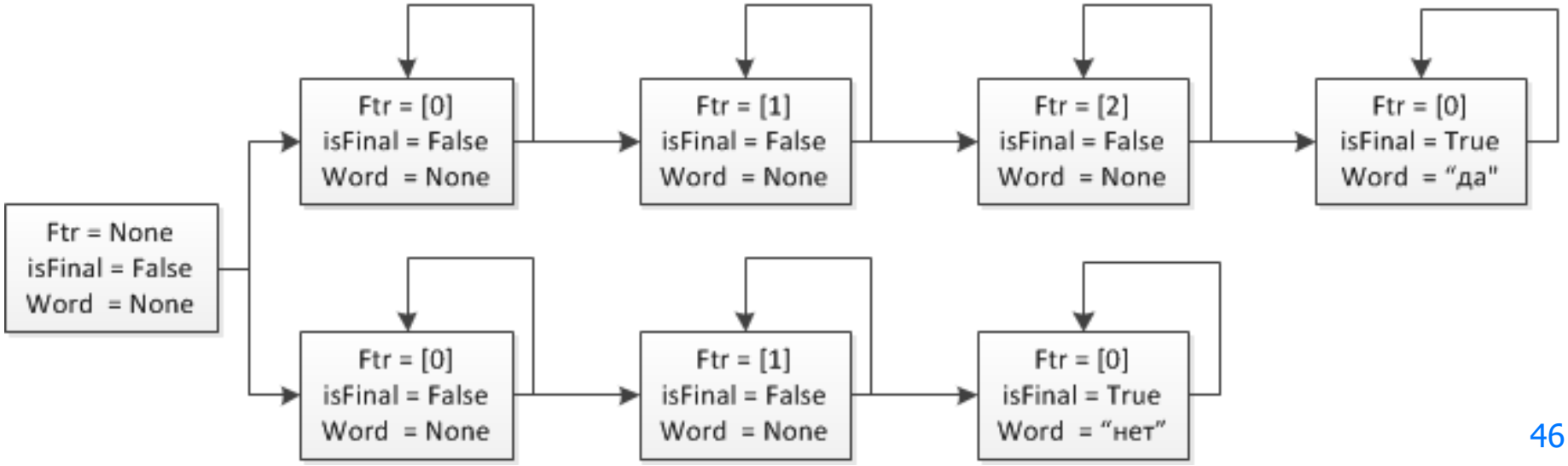
Алгоритм Token Passing

- Построение (направленного) графа распознавания:
  - Каждый кадр (вектор признаков) каждого эталона свяжем с одним узлом графа (состояние)
  - Дополним состояние меткой конца фразы (isFinal), словом, соответствующим эталону (только в последнем кадре) и списком следующих состояний.
  - Введем фиктивный стартовый узел (просто для удобства)
- Пример:

  2 эталона

  «да» (0, 1, 2, 0)

  «нет» (0, 1, 0)



Сравнение с эталоном

Алгоритм Token Passing (продолжение)

- Есть тестовая фонограмма – надо найти ближайший эталон
- Токен – структура, связанная с состоянием графа и хранящее текущее расстояние, накопленное при проходе по эталону до этого состояния.
- В каждый момент храним много токенов, при переходе на следующий кадр данный токен либо остаётся в том же состоянии, либо перемещается в следующее и обновляет накопленную дистанцию
- В данном состоянии имеет смысл хранить только токен с лучшей дистанцией! (принцип динамического программирования)
- Когда дошли до конца – сравниваем токены в финальных состояниях каждого эталона: токен с наилучшей дистанцией определяет «выигравшее» слово.

Сравнение с эталоном

Алгоритм Token Passing: псевдокод (python-like)

```
10 создаём стартовый токен, помещаем его в виртуальный узел, помещаем в activeTokens
20 for frame in РаспознаваемыйФайл:
30     for token in activeTokens:
40             for переход in всеВозможныеПереходыИз(token.state):
50                     newToken=создать новый токен в узле, куда указывает переход
50                     скопировать всё из token в newToken
60                     newToken.dist+=расст(frame, КадрЭталонаТамКудаМыПерешли)
70                     nextTokens.append(newToken)
80     закончили обработку кадра
90     проредить токены, оставив в каждом узле графа только токен с лучшей дистанцией
100    activeTokens=nextTokens
110    очистить nextTokens
120 закончили обработку записи
130 Для выдачи результата, перебрать все токены, дошедшие до финальных узлов.
140 Вернуть слово из токена с лучшим расстоянием победил.
```

Сравнение с эталоном

Алгоритм Token Passing: достоинства и недостатки

- Достоинства алгоритма:
  - Компактное представление всех эталонов сразу
  - Удобный и единообразный алгоритм обработки
  - Легко обобщается на более сложные графы (рассмотрим в следующих лекциях)
- Недостаток:
  - Сложность пропорциональна числу узлов графа (по одному токену на узел)
- Pruning – отсечка (выкидывание) малоперспективных токенов на каждом кадре:
  - Beam pruning: удаление всех токенов отличающихся от наилучшего по накопленной дистанции не более, чем на заданную величину (beam width)
  - Histogram pruning: удаление всех токенов, кроме N лучших по накопленной дистанции

Сравнение с эталоном

Достоинства и недостатки DTW-подхода в целом:

- \+ Интуитивность идеи
- \+ Простота реализации
- \+ Допустимо создавать эталоны не слов, а произвольных звуков
- – Необходимость хранить все эталоны
- – Ограниченность набора эталонов в смысле обобщающей способности
- – Невысокая точность
- – Маленький объем словаря

- Выход:
  - Создавать «модели» слов, описывающие все потенциальное множество их эталонов
  - Обучать модели по большим объемам данных
  - Для распознавания использовать только сами модели, без эталонов

Группа компаний ЦРТ

О НАС

В группу компаний ЦРТ входят компании ЦРТ, ЦРТ-инновации и SpeechPro.

ЦРТ – российская компания, разработчик инновационных систем в сфере технологий синтеза и распознавания речи, анализа аудио- и видеоинформации, распознавания лиц, голосовой и мультимодальной биометрии.

ЦРТ-инновации – научно-исследовательская компания, передовой разработчик голосовых и бимодальных биометрических систем. Резидент Фонда «Сколково». Области научно-исследовательской деятельности компании: биометрия по голосу и лицу, распознавание речи, анализ больших данных.

SpeechPro – представительство Группы ЦРТ в Северной Америке с главным офисом в Нью-Йорке. SpeechPro взаимодействует с клиентами и партнерами ЦРТ из США и Канады.

КОНТАКТНАЯ ИНФОРМАЦИЯ

Санкт-Петербург
Адрес: 194044, г. Санкт-Петербург,
ул. Гельсингфорсская, 3-11, лит. Д
Телефон: (+7 812) 325-88-48
Факс: (+7 812) 327-92-97
Эл. почта: stc-spb@speechpro.com

Москва
Адрес: Москва, ул. Земляной Вал, д. 59, стр. 2
Телефон: +7 (495) 669-74-40
Факс: +7 (495) 669-74-44
Эл. почта: stc-msk@speechpro.com

Спасибо за внимание!

Вопросы?