



Распознавание речи. Обзор современных задач и алгоритмов

Арсений Горин

МГТУ, 21 апреля 2016

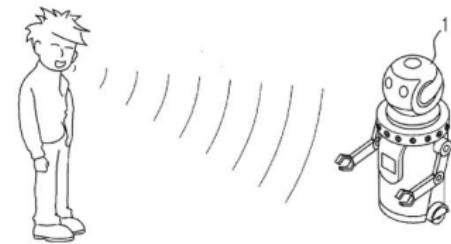
Введение

Распознавание речи - наука о том, как научить компьютер понимать речь

- транскрипции - из сигнала в текст
- извлечение информации - семантика фраз
- диалог человек-машина
- поиск и каталогизация аудио документов

Результат более 50 лет исследований

- цифровая обработка сигналов
- лингвистика и фонетика
- теория вероятности, статистика и машинное обучение
- дискретная математика и оптимизация
- психоакустика и психология



Содержание

1 общий обзор речевых технологий

- речевой сигнал и его возникновение
- исторический экскурс
- современные задачи и проблемы

2 основы обработки речевого сигнала

- спектр и форманты
- банк фильтров и кепстральные признаки

3 статистические модели распознавания речи

- общий принцип работы
- акустическая и языковая модели
- алгоритмы поиска и обучения
- акустические модели на нейронных сетях

4 полезные ссылки на программы и литературу

Содержание

1 общий обзор речевых технологий

- речевой сигнал и его возникновение
- исторический экскурс
- современные задачи и проблемы

2 основы обработки речевого сигнала

- спектр и форманты
- банк фильтров и кепстральные признаки

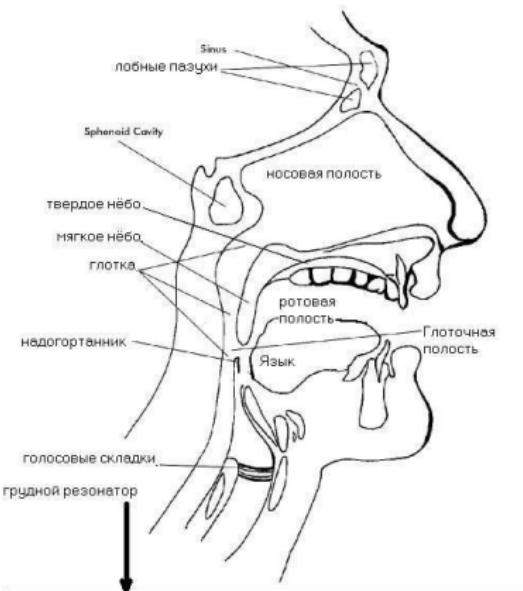
3 статистические модели распознавания речи

- общий принцип работы
- акустическая и языковая модели
- алгоритмы поиска и обучения
- акустические модели на нейронных сетях

4 полезные ссылки на программы и литературу

Речевой аппарат человека

Речевой сигнал формируется различными органами (инструментами) вокального тракта человека. Можно выделить 3 основных категории:



- **резонаторы:** легкие и голосовые связки
 - амплитуда
 - высота и тембр голоса
 - формируют различие между звонкими и глухими звуками
- **частотные фильтры:** носовая и ротовая полости, гортань
 - формируют окраску звуков
 - основные признаки различия гласных
- **модуляторы:** небо, язык, зубы, губы
 - формирование согласных
 - также вносят вклад в звучание гласных

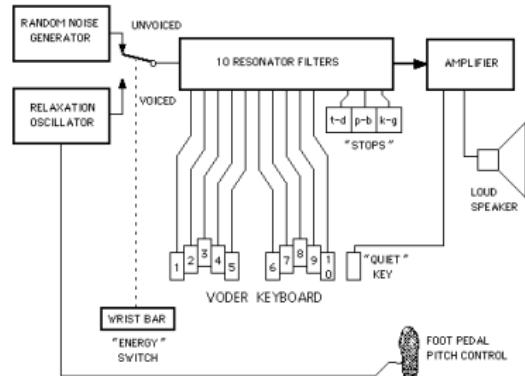
Voder - первый речевой синтезатор

Homer Dudley - Bell Labs (1939)

На международной ярмарке в Нью-Йорке представлен первый синтезатор речи

Состав

- генератор колебаний
- генератор шума (газовая трубка): фрикативные согласные (/ф/, /х/)
- педальный регулятор тона (pitch): “настроение” фразы
- 10 полосовых фильтров, регулируемых клавишами
- дополнительные клавиши - модель взрывных согласных (/п/, /д/)



Системы распознавания речи: 50-60е годы

- 1952 - Bell Labs “Audrey”: дикторозависимая система распознавания цифр (точность примерно 90%)
- 1962 - IBM “Shoebox”: словарь из 16 слов
- методы:
 - сравнение сегментов по правилам, написанным вручную (pattern matching)
 - акустические признаки на основе Linear Predictive Coding (Итакура-Сайто, 1967-1968)
 - динамическое программирование (Винтсюк, Витерби, 1967)
- недостатки:
 - ограниченные словари
 - дикторозависимость
 - работа в условиях, близких к идеальным



Системы распознавания речи: 70-80е годы

- DARPA (Defense Advanced Research Projects Agency):
 - заказы на системы распознавания речи, диалоговые и интеллектуальные системы поиска
 - исследовательские программы и соревнования по точности систем распознавания среди ведущих университетов и лабораторий
- 1976 - CMU “Harpy”: словарь 1100 слов, **распознавание слитной речи**
- 1976 - IBM “Tangora”: пока некоммерческая система, 20000 слов
- методы:
 - статистические модели (**Hidden Markov Models**).
Ф. Джелинек (1985): “Каждый раз, когда я увольняю лингвиста, точность распознавания речи улучшается”
 - улучшены акустические признаки (**MFCC, PLP**)
 - улучшены алгоритмы поиска (**beam search**)
 - адаптация моделей (диктор, шум, и т.д.)
- 1987 - кукла **Julie Talking Doll**: один из первых коммерческих продуктов, использовавший речевые технологии



90-2000е: из институтов в продукты

- 1992-..., DARPA, NIST: ежегодные исследовательские проекты
 - системы распознавания слитной речи с большим словарем (**Large Vocabulary Continuous Speech Recognition**)
 - транскрипции радиопередач и телефонных записей
 - поиск ключевых слов в речевом потоке
- 2003 CALO: Cognitive Assistant that Learns and Organizes
 - разработка интеллектуального организатора для бойцов армии США
 - 300 исследователей, 25 университетов
 - одна из лидирующих команд позднее создала **Siri** (куплено Apple)
- коммерческие системы
 - 1990 - **Dragon dictate (Windows 98, Mac)**: распознавание раздельных слов с большим словарем
 - 1997-2016 - **Dragon Naturally Speaking**: слитная речь.
Перешел к компании Nuance Communications

21 век: большие данные и старые проблемы

За последние 10 лет больше речевых сервисов, чем за все годы

- **голосовой поиск:** Google Voice Search, Яндекс, Microsoft Bing, Baidu
- **интеллектуальные диалоговые системы**
(Apple Siri, Google Now, Microsoft Cortana, Samsung S Voice)
- **сервисы транскрипции аудио данных** (Nuance, IBM Watson, ...)
- **извлечение информации из аудио потока и голосовая аналитика**

Основные направления улучшения точности

- огромное количество данных для **обучения** систем
 - тысячи и десятки тысяч часов аудио с транскрипциям
 - текстовые данные для языковых моделей
- **новые алгоритмы и модели**, ранее неприменимые ввиду недостаточной мощности вычислительных ресурсов (многослойные нейронные сети)
- **поддержка большего количества языков**
(Google: 41, Nuance: 86, планы facebook: 70 языков)

Разве еще не все проблемы решены?

Несмотря на коммерческий успех, точность современных речевых систем далека от человека

- **системы “настраиваются”** на определенный контекст за счет ограничения словаря и модели языка
("OK Google: бб, зеленый кирпич на 2м вперед и тепловоз в Сыктывкар")
- **вариации темпа и стиля речи** также являются помехой
(смешанный темп, произнесение по слогам, крик)
- **различие акустических признаков дикторов**
(например, система обучена на взрослых и тестируется на детях)
- **спонтанная речь с перекрытиями** (дебаты, споры)
- **распознавание речи в условиях помех** (шум, реверберация)
- **недостаток данных для обучения моделей**, например редких языков

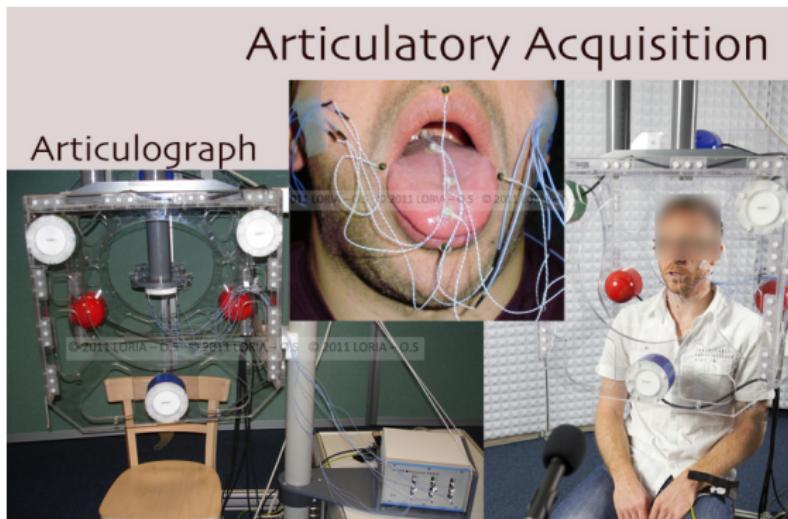
Часто требуется создание, настроенной на решение конкретной задачи
(простая модель для голосового управления работом работает точнее, чем google, настроенный на распознавание Web запросов)

Примеры (не/плохо)решенных задач

- системы поиска ключевых слов для редких языков
 - IARPA Babel 2015-2016
 - телефонные записи на 26 малоресурсных языках
(Казахский, Тамил, Литовский, Курдский, Монгольский и т.д.)
 - трудности:
 - отсутствие лингвистических знаний и данных
 - спонтанная речь (нарушение грамматических правил, темп)
 - зашумленный канал
 - ошибки составляют порядка 40-70% в зависимости от языка
- распознавание зашумленной речи с реверберацией
 - IARPA ASPIRE 2015 challenge
 - трудности: сильно искаженный сигнал
 - ошибки: около 30%
 - аналогичная система дает 18% для телефонного канала без реверберации
- распознавание конференций, записанных на удаленный микрофон
 - AMI corpus (Edinburgh university)
 - трудности: спонтанная речь (перекрытия, меняющийся темп), микрофон
 - ошибки: 40-50% (20%, если дикторы используют гарнитуру)
 - например, транскрипции радиопередач делаются с примерно 10% ошибок

Другие задачи: артикулография

- синхронная запись аудио потока и динамики речевого аппарата
- небольшие датчики закрепляются внутри ротовой полости, иногда также используют фронтальную камеру
- задачи
 - изучение произношения звуков и пар звуков, анализ дефектов речи,
 - синтез немой речи (silent speech)



Другие задачи: оценка произношения

Использование методов распознавания речи для оценки правильности произношения

- модель обучается на носителях языка
- из тестовой фразы извлекаются **метрики качества** произношения фонем: насколько похоже на носителя языка, длительность, интонация
- по выборке студентов преподаватель обучает автоматическую **систему выставления оценок** (какие ошибки критичны, какие не очень)



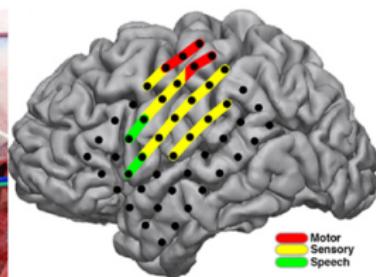
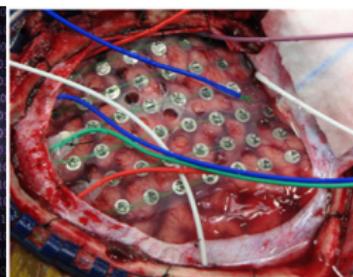
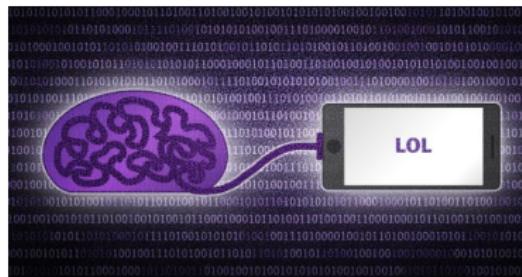
Ссылки:

- <http://pronunciationeval.blogspot.ru> - проект Google summer of code на основе системы Sphinx
- <http://www.englishcentral.com> - бесплатное приложение (Android, iOS). Можно потестировать

Другие задачи: проект мозг-в-текст

Karlsruhe Institute of Technology (KIT), 2015 (brain-to-text)

- данные с 7 пациентов больных эпилепсией
- сенсоры размещаются на коре головного мозга (cerebral cortex): электрокортикография (electrocorticography)
- модель обучается по сигналам мозга при произнесении фраз
- при распознавании сигналы с мозга используют для предсказания фонем, далее используется модель языка



Содержание

1 общий обзор речевых технологий

- речевой сигнал и его возникновение
- исторический экскурс
- современные задачи и проблемы

2 основы обработки речевого сигнала

- спектр и форманты
- банк фильтров и кепстральные признаки

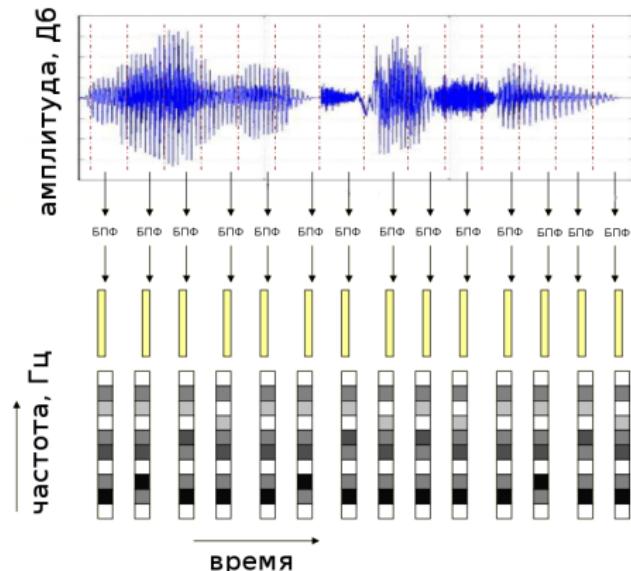
3 статистические модели распознавания речи

- общий принцип работы
- акустическая и языковая модели
- алгоритмы поиска и обучения
- акустические модели на нейронных сетях

4 полезные ссылки на программы и литературу

Спектральный анализ

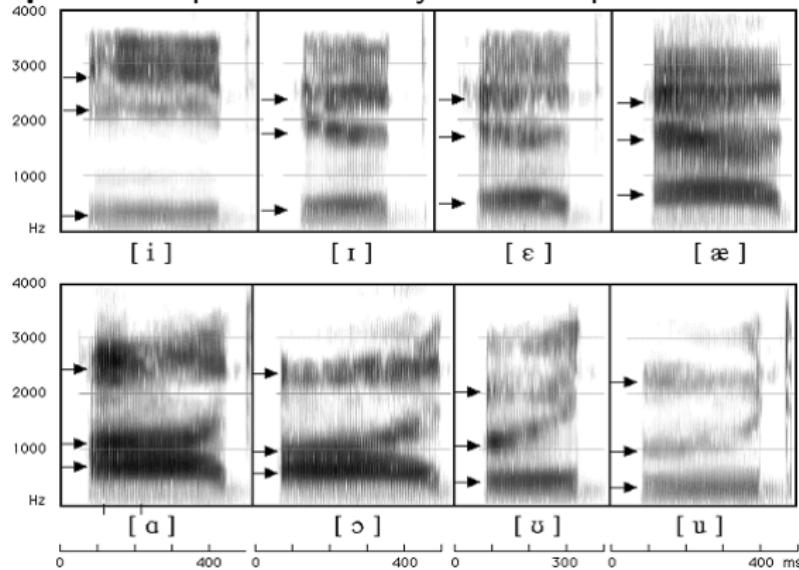
- оконное преобразование Фурье: $X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n] \omega[n-m] e^{-j\omega n}$
- m -размер окна; x - амплитуда в точке n , $\omega[n]$ - оконная функция
- спектrogramma $\{x(t)\}(m, \omega) = |X(m, \omega)|^2$



Примеры спектрограмм. Форманты. Фонемы

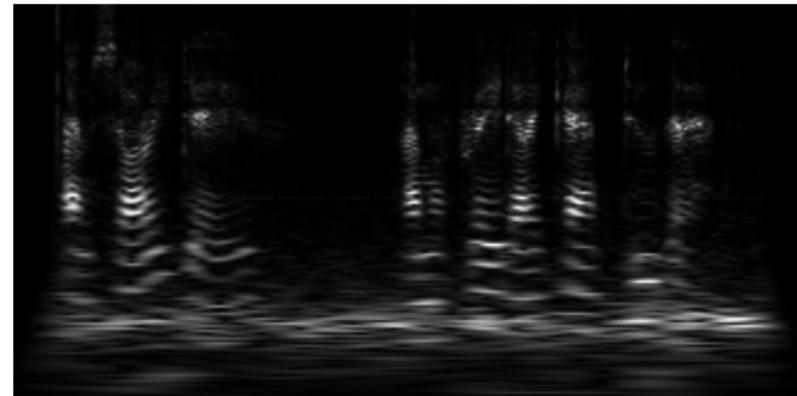
- **форманты** - характерные кривые спектрограммы (в основном для гласных)
- **описывают частоту тона** (темпер звука)
- **не всегда просто определить**: зависят от контекста и диктора

Пример англ. **фонем** - простейших акустически различимых единиц языка

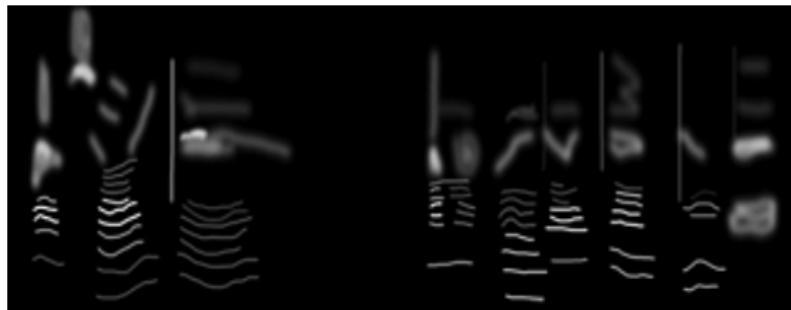


Форманты. Еще пример

<http://arss.sourceforge.net>

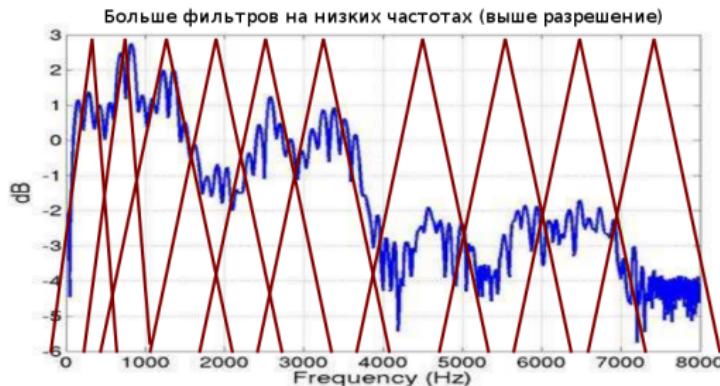


Ps



Банк фильтров

- спектральные признаки имеют слишком большую размерность и сильно зависят от диктора, аудио фона, контекста (**не робастны**)
- в реальных задачах используется **банк фильтров** в шкале Мела и кепстральные признаки
 - частота в шкале Мела основана на экспериментах с человеческим восприятием звука $Mel(f) = 2595 \log_{10}(1 + \frac{f}{700})$
 - амплитуды оконной АЧХ умножаются на треугольные фильтры, значения внутри каждого фильтра складываются
 - признаки банка - вектор размерности, равной количеству фильтров



Коротко о кепстральных признаках

Признаки банка фильтров

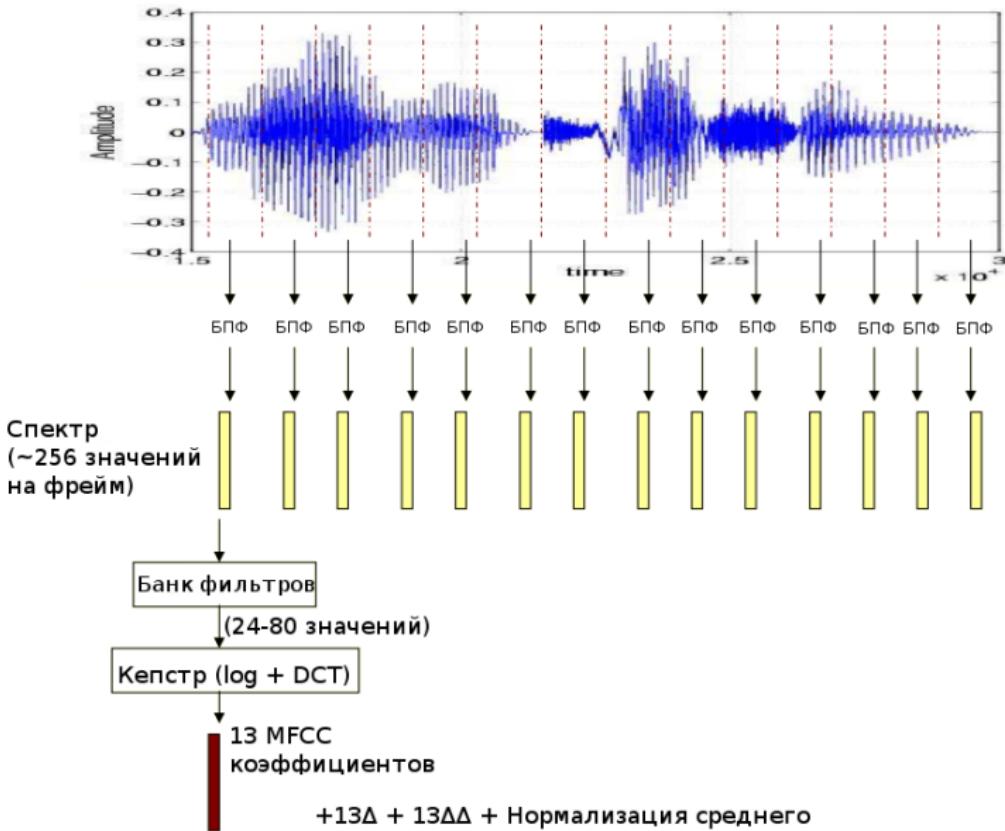
- все еще сильно коррелируют между собой
- сложно использовать в некоторых статистических моделях
(например в Гауссовых смесях с диагональной ковариационной матрицей)

Кепстр (*игра слов: spectrum-cepstrum, frequency analysis \Rightarrow quefrency analysis*)

- отражает информацию об изменении в спектре
- иногда называют “спектр спектра”
- на практике используют дискретное косинусное преобразование DCT на логарифме энергий банка фильтров

- $DCT_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right]$
- $k = 0, \dots, N - 1$
- N обычно 12-13

Коротко о кепстральных признаках



Содержание

1 общий обзор речевых технологий

- речевой сигнал и его возникновение
- исторический экскурс
- современные задачи и проблемы

2 основы обработки речевого сигнала

- спектр и форманты
- банк фильтров и кепстральные признаки

3 статистические модели распознавания речи

- общий принцип работы
- акустическая и языковая модели
- алгоритмы поиска и обучения
- акустические модели на нейронных сетях

4 полезные ссылки на программы и литературу

Формализация задачи

Зная последовательность акустических признаков $\mathcal{O} = (\mathbf{o}_1, \dots, \mathbf{o}_T)$,
найти наиболее вероятную последовательность слов $\mathcal{W}^* = (w_1, \dots, w_n)$
из множества возможных фраз некоторого языка \mathcal{L}
(например, всего Русского, или языка возможных команд управления роботом)

$$\mathcal{W}^* = \arg \max_{\mathcal{W} \in \mathcal{L}} P(\mathcal{W} | \mathcal{O})$$

Условная вероятность $P(\mathcal{W} | \mathcal{O})$ удобно представляется
по **формуле Байеса**

$$P(\mathcal{W} | \mathcal{O}) = \frac{P(\mathcal{O} | \mathcal{W}) P(\mathcal{W})}{P(\mathcal{O})} \propto P(\mathcal{O} | \mathcal{W}) P(\mathcal{W})$$

- 
- $P(\mathcal{O} | \mathcal{W})$ - **акустическая модель**
(функция правдоподобия)
 - $P(\mathcal{W})$ - **языковая модель**

Основные компоненты системы распознавания

последовательность слов

- языковая модель

последовательность звуков (фонем)

- словарь произношений

последовательность фонем
в каждом окне сигнала

- состояния скрытой Марковской модели

модель наблюдаемой величины
(распределения признаков)

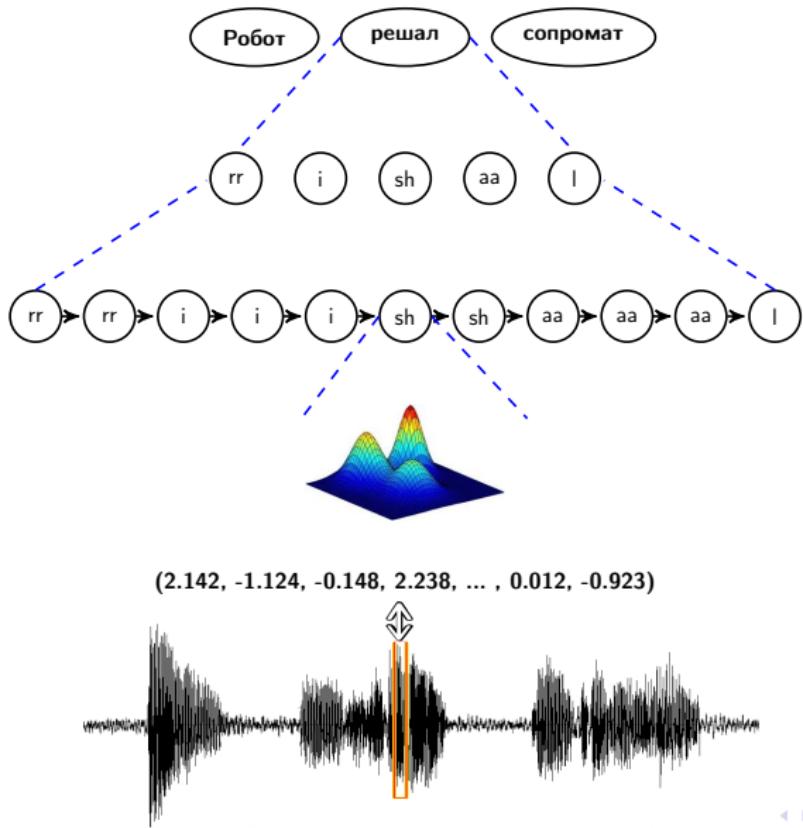
- модель смеси Гауссиан

- нейронная сеть

вектор акустических признаков окна

- MFCC, PLP, PNCC, LPC, и т.д.

аудио сигнал



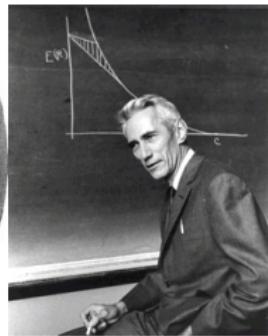
Языковая модель (language model)

- **A. A. Марков. 1913 г.**

“Примѣръ статистическаго изслѣдованія
надъ текстомъ “Евгенія Онѣгина” ”

- **C. E. Shannon. 1951 г.**

“Prediction and entropy of printed English.”



- **N-граммы (часто 3-граммы):**

какова вероятность увидеть слово w_i в контексте (w_0, \dots, w_{i-1})

$$P(w_i | w_{i-1}, \dots, w_0) \approx P(w_i | w_{i-1}, w_{i-2})$$

- **обучение** заключается в подсчете последовательностей слов в текстовых данных

- **тематика текстов** зависит от задачи

- **LM задает множество возможных гипотез** системы распознавания, а также позволяет уменьшить ошибки

Словарь (lexicon, dictionary)

- **задает произношения** для известных системе распознавания слов
- **создается лингвистом, либо системой графемы-в-фонемы** (grapheme-to-phoneme)
- слово может иметь несколько произношений
- системы с малым узкоспециализированным словарем работают гораздо точнее (меньше возможности делать ошибки)

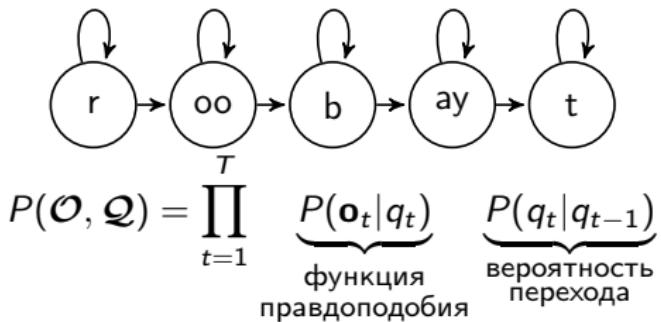
Пример словаря CMU Sphinx (186k слов)

ёлку	j oo l k u
ёжик	j oo zh ay k
...	...
чучело	ch uu ch ae l a
<hr/> <sil>	<hr/> SIL
++HES++	+HES+

Акустическая модель

Скрытая Марковская Модель:

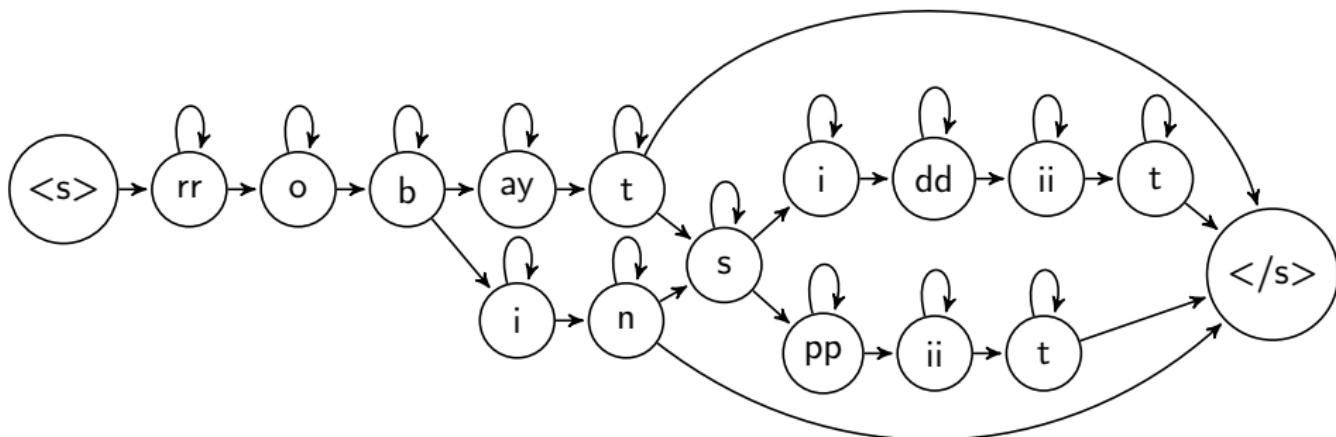
- наблюдаемые величины (акустические признаки) $\mathcal{O} = (\mathbf{o}_1, \dots, \mathbf{o}_t)$
- скрытые состояния (предполагаемые фонемы) $\mathcal{Q} = (q_1, \dots, q_t)$
- вместо фонем чаще используют контекстно-зависимые трифоны



- функция правдоподобия: на какую фонему больше всего похожи признаки \mathbf{o}_t в фрейме t
- сравниваем с моделью: Гауссова смесь, Нейронная сеть
- длительность фонемы задается моделью перехода

Все вместе (упрощенная модель)

- зададим простой **язык**: (`{'робот', 'робин'}`, `{'_', 'сидит', 'спит'}`)
- добавим **произношения**
- изобразим схематично в виде **Марковской модели**



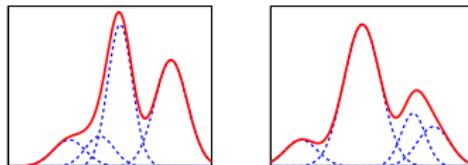
остается добавить веса и функции правдоподобия...

Функция правдоподобия - GMM

Задача: по вектору акустических признаков \mathbf{o}_t посчитать вероятности фонем (или трифонов)

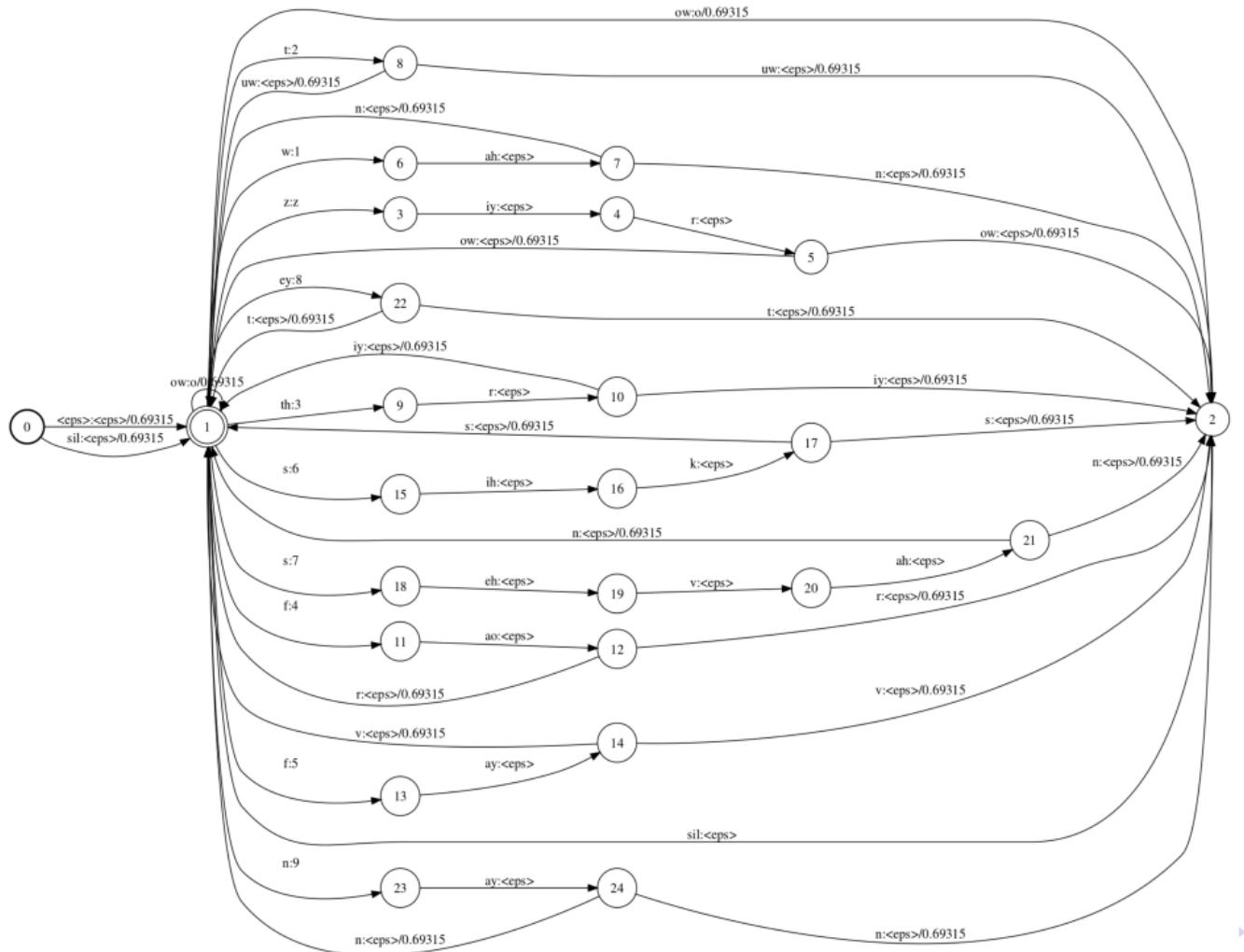
Гауссова смесь (Gaussian Mixture Model, GMM)

- для каждого состояния j (фонемы, трифона, или их частей) определим смесь из нескольких Гауссиан:
 - векторы мат. ожидания μ_{jl}
 - дисперсии σ_{jl}
 - веса Гауссиан ω_{jl}
- $P(\mathbf{o}_t | q_t = j) = \sum_{l=1}^M \omega_{jl} \mathcal{N}(\mathbf{o}_t | \mu_{jl}, \sigma_{jl})$



Процесс распознавания

- CMU Harpy. Эвристический поиск **beam search** 
- в современных системах часто используются **взвешенные конечные автоматы (WFST)**
 - **входные символы:** фонемы
 - **выходные символы:** слова, или пустые (`<eps>`)
 - на дугах размещаются **веса**, вычисляемые по акустической и языковой моделям
- **компактное представление** словаря и языковой модели
- **удобно использовать** вместе с вероятностями от акустической модели
- **эффективные алгоритмы поиска** (OpenFST)



Обучение модели (кратко)

Обучение - процесс оценки параметров моделей по наблюдаемым данным

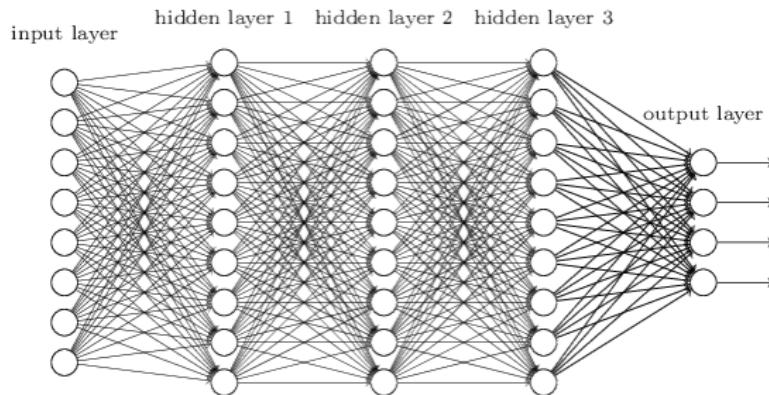
- **словарь**: вручную, либо изъятие правил из существующего словаря для угадывания произношений новых слов (grapheme-to-phoneme)
- **языковая модель**: считаем словосочетания в текстах

Обучение акустических моделей - чуть менее очевидно

- **даны** аудио записи фраз (сегменты < 20с) с транскрипциями (слова)
- **неизвестны** параметры модели и четкие границы слов и фонем
- **используется** итеративный алгоритм (**Expectation Maximization**)
 - начинаем с предположения о равномерной сегментации
 - обучаем модель (сперва грубую)
 - используем модель для уточнения границ (Forced Alignment)
 - повторяем предыдущие 2 шага несколько раз
- **при обучении решаются сразу 2 задачи: выравнивание аудио и фонемной последовательностей и оценка параметров HMM-GMM**

Акустические модели на нейронных сетях

В настоящее время популярны модели, использующие **многослойные нейронные сети** (Deep Neural Networks, DNN) вместо Гауссовых смесей.



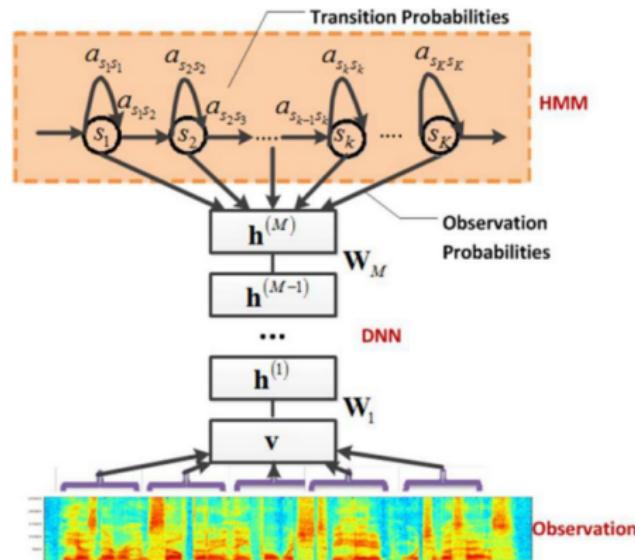
- **много параметров**
(4-8 слоев, 1-4k нейронов на слой, 10-100M параметров в сумме)
- **модель обучается** методом обратного распространения ошибки
- **используют выровненные последовательности** аудио-фонемы
(полученные с HMM-GMM)
- используют разные целевые функции и методы оптимизации

Акустические модели на нейронных сетях

- + гораздо более точные (особенно для больших данных)
- обычно требуют настройки множества параметров
(зависит от данных, трудно найти оптимальное число параметров)
- обучение достаточно ресурсоемкое (проще с GPU)
- множество архитектур сетей, пока что не ясно, какая универсальна
- крайне легко переобучаются

Гибридные HMM-DNN

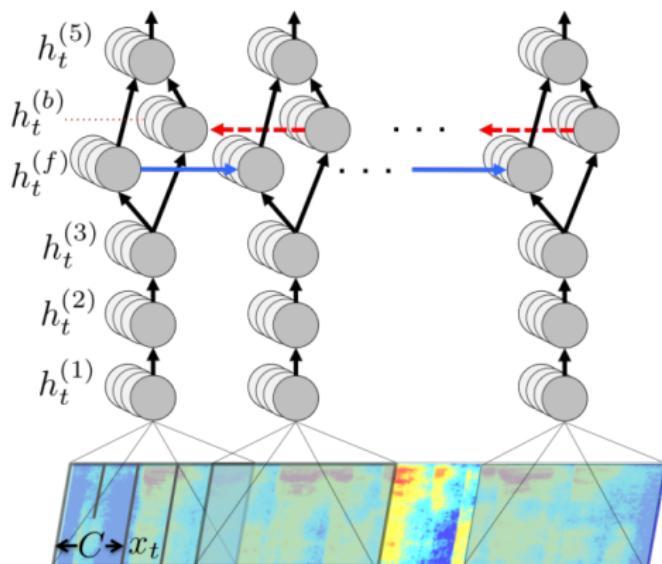
- **на входе** склеенный контекст признаков (часто более 20 фреймов)
- **на выходе** вероятности фонем
- **вероятности используются в поиске** так же, как функция правдоподобия Гауссовых смесей



Dahl, George E., et al. "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition." *Audio, Speech, and Language Processing, IEEE Transactions on* 20.1 (2012): 30-42.

Рекуррентные сети RNN, LSTM

- кроме акустических признаков, **на вход подается выход нейросети из прошлого**
- “пробегают” по спектрограмме, храня контекст всей фразы
- активно используются в других задачах, связанных с последовательностями (языковые модели, машинный перевод, и т.д.)



Содержание

1 общий обзор речевых технологий

- речевой сигнал и его возникновение
- исторический экскурс
- современные задачи и проблемы

2 основы обработки речевого сигнала

- спектр и форманты
- банк фильтров и кепстральные признаки

3 статистические модели распознавания речи

- общий принцип работы
- акустическая и языковая модели
- алгоритмы поиска и обучения
- акустические модели на нейронных сетях

4 полезные ссылки на программы и литературу

Что почитать?

- Jurafsky D., Martin J.H. (2008) Speech and language processing, 2nd edition. Prentice Hall.
- Hinton, G., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6), 82-97.
- <https://habrahabr.ru/company/yandex/blog/198556> - Распознавание речи от Яндекса. Под капотом у Yandex.SpeechKit

А может сразу что-то позапускать?

- **CMU Sphinx** (<http://cmusphinx.sourceforge.net>)
 - быстрый старт, много tutorial'ов
 - готовые бесплатные модели для Русского языка
 - java декодер - удобно встраивать в мобильные платформы
 - <http://cmusphinx.sourceforge.net/wiki/tutorialam> - пошаговая инструкция
 - <http://nshmyrev.blogspot.ru/2010/04/testing-asr-with-voxforge-database.html> - обучение модели на бесплатном дата сете voxforge
 - <https://sourceforge.net/p/cmusphinx/discussion/help> - с вопросами сюда
- **Kaldi** (<http://kaldi.sourceforge.net>)
 - старт медленнее, но **можно строить более точные системы**
 - большой бесплатный **корпус Английских аудиокниг** и модели для Kaldi (около 1000 часов): <http://www.openslr.org/12>
 - <http://kaldi.sourceforge.net/tutorial.html> - инструкция
 - <https://sourceforge.net/p/kaldi/discussion> - за помощью сюда
- **другие бесплатные полезные инструменты, которые полезно знать**
 - **SRILM** - обучение языковых моделей (<http://www.speech.sri.com>)
 - **praat** - анализ аудио сигнала: спектр, форманты, и т.д. (<http://www.fon.hum.uva.nl/praat>)
 - **sox** - "Swiss Army knife of sound processing" (<http://sox.sourceforge.net>)
 - крайне желательно: **python, bash, linux**