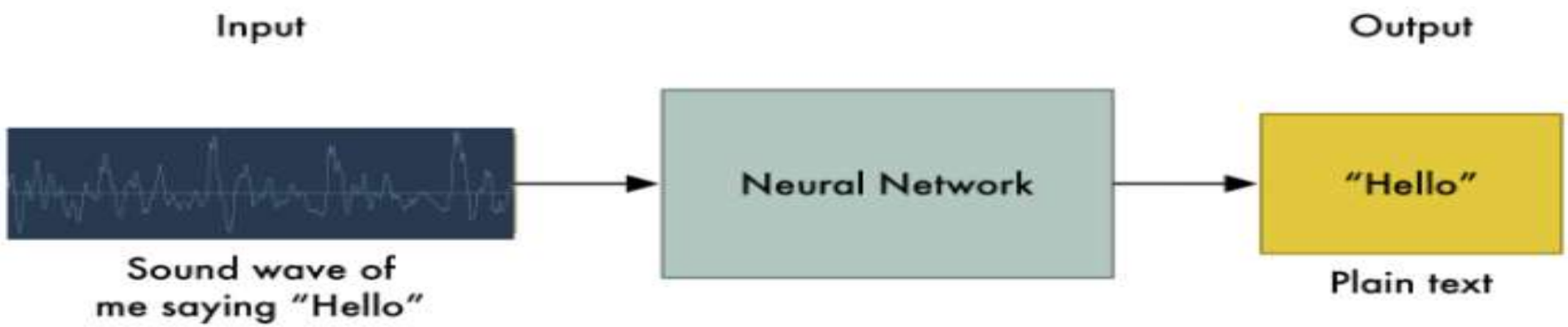
Применение методов deep learning к задаче распознавания речи.

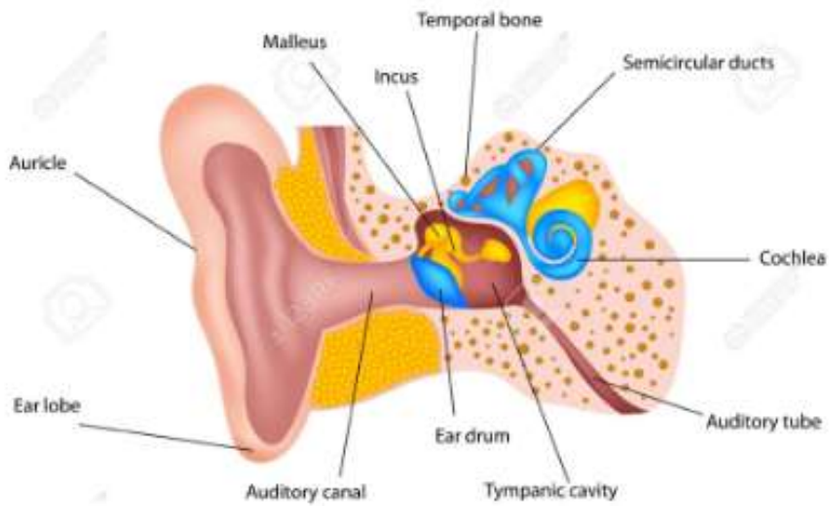
Халилова Сефае

Задача: Распознавание речи(ASR, TTS)

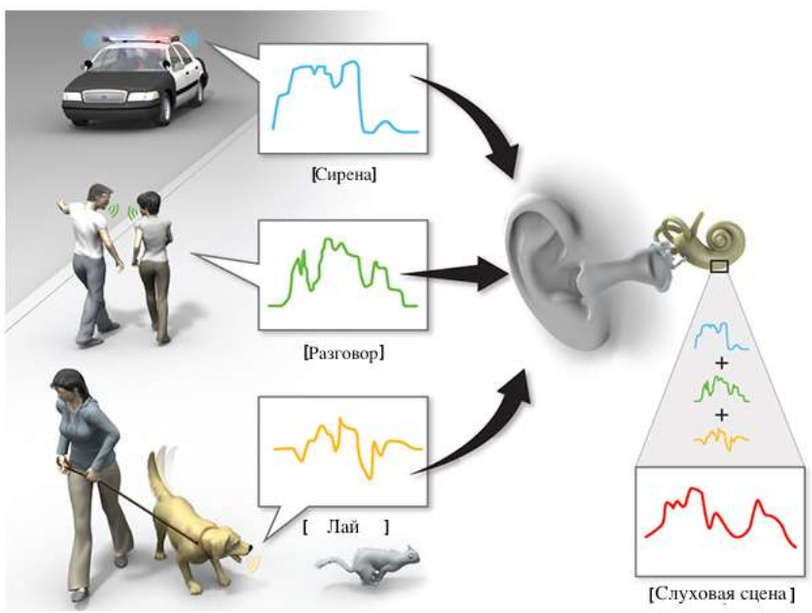
- На входе: аудио, содержащее речь
- На выходе: текст



Как работает человеческое ухо?



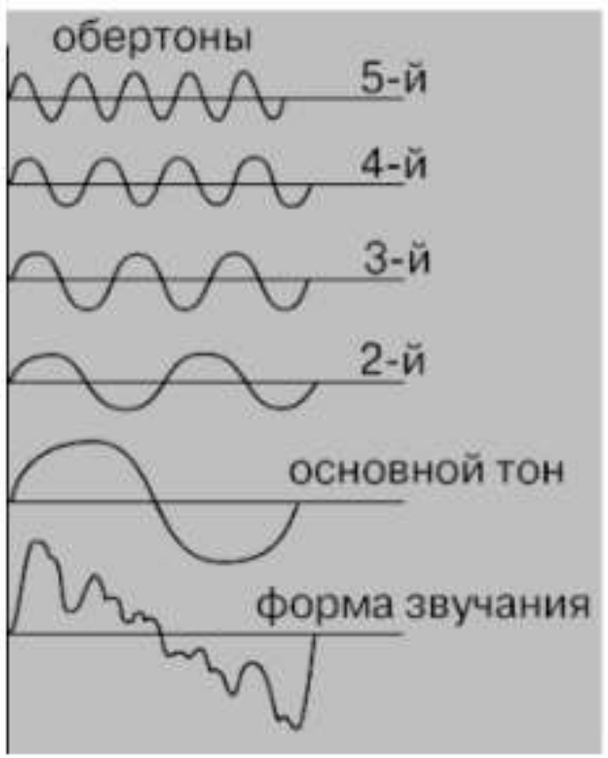
- улитка состоит из большого количества резонаторов;
- каждый из этих резонаторов отзывается на колебания определенной частоты и возбуждает соответствующие нервные окончания, входящие в состав слухового нерва;
- нервы идут к разным нейронам

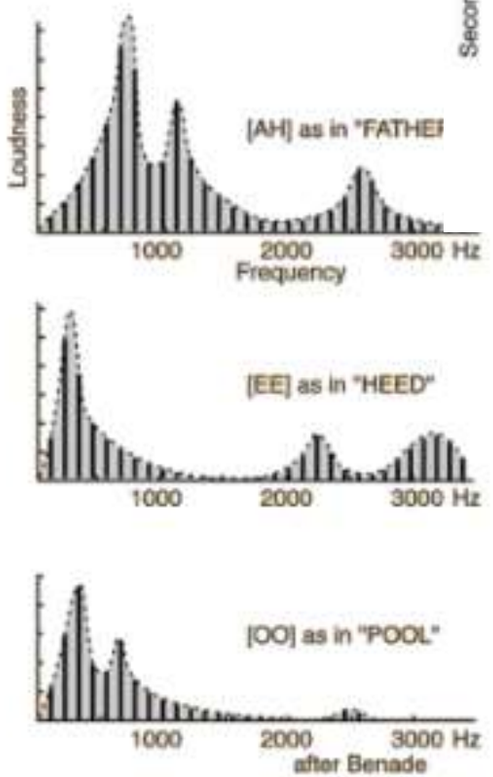


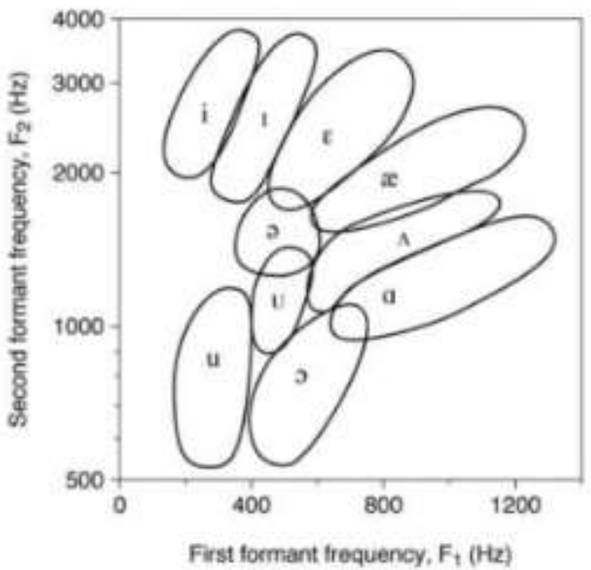
Fast Fourier Transform (FFT, Быстрое преобразование Фурье)

- алгоритм, который преобразует данные из пространственно-временной области в частотную область

- частота колебаний связок – основная частота голоса, "высота тона"(60-400 Гц);
- преобразования голосового тракта обеспечивают обертона;
- FFT намного проще понять, чем исходный звуковой сигнал!







Голосовое распознавание речи происходит в несколько этапов:

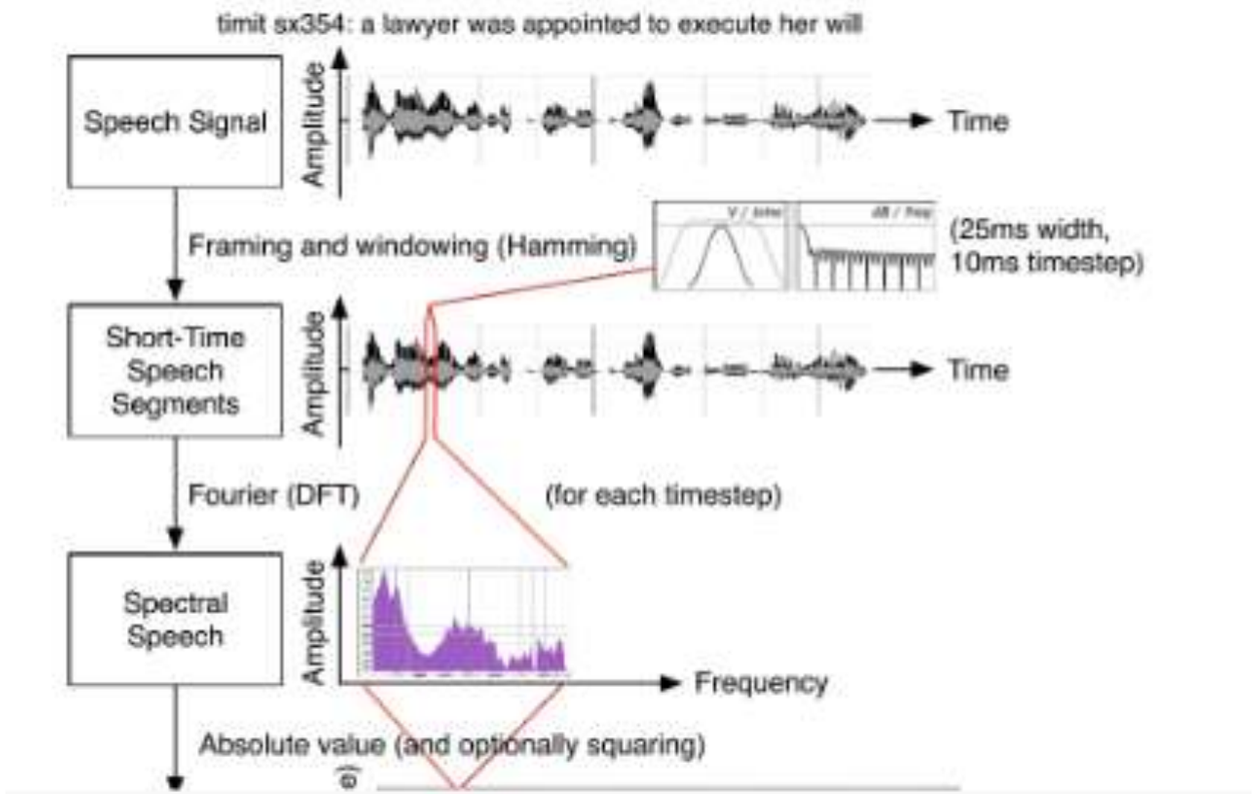
- речь клиента попадет по каналам интерфейса на сервер и разделяется на фреймы (маленькие фрагменты), например, длиной 25 миллисекунд с шагом 10 миллисекунд (таким образом из одной секунды речи получается сто фреймов).

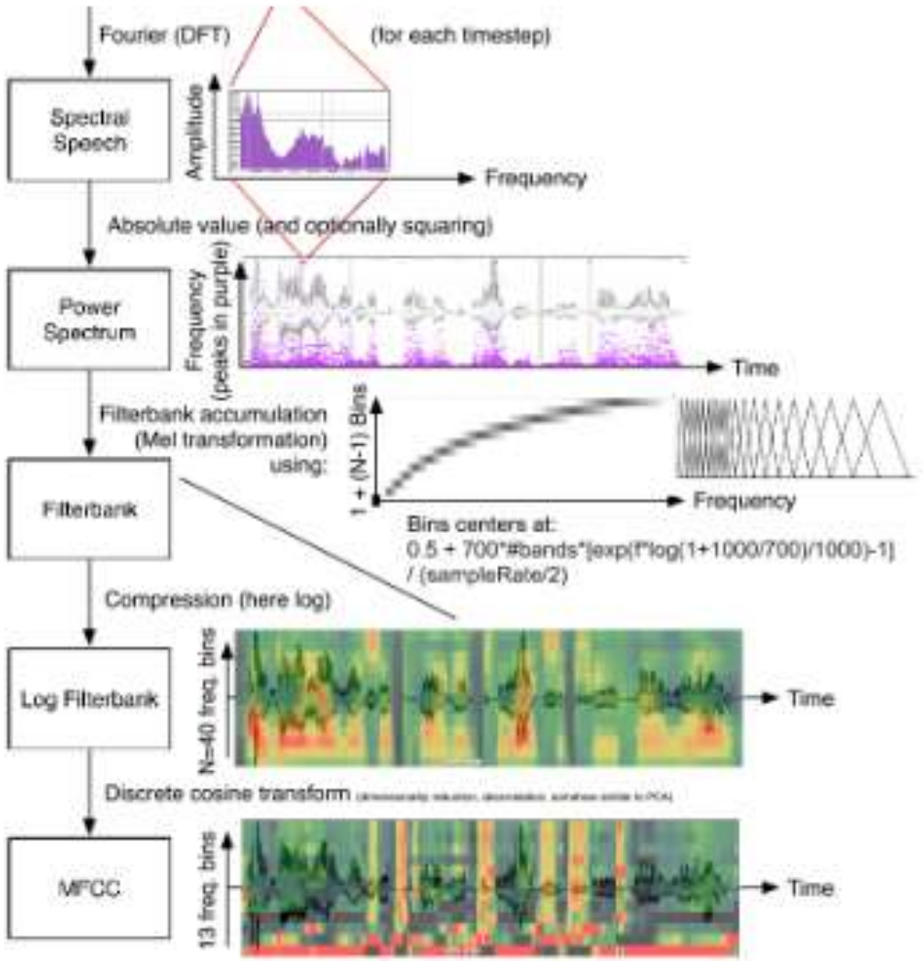
- нейросеть отсеивает шумовые помехи, удаляются фреймы, не несущие звуковой окраски;

- очищенная звуковая дорожка поступает в устройство акустического моделирования, где импульсы преобразуются в фонемы (минимальные единицы языка);

- фонемы поступают в лингвистическую программную модель, где происходит анализ потока и из них выстраиваются законченные фразы;

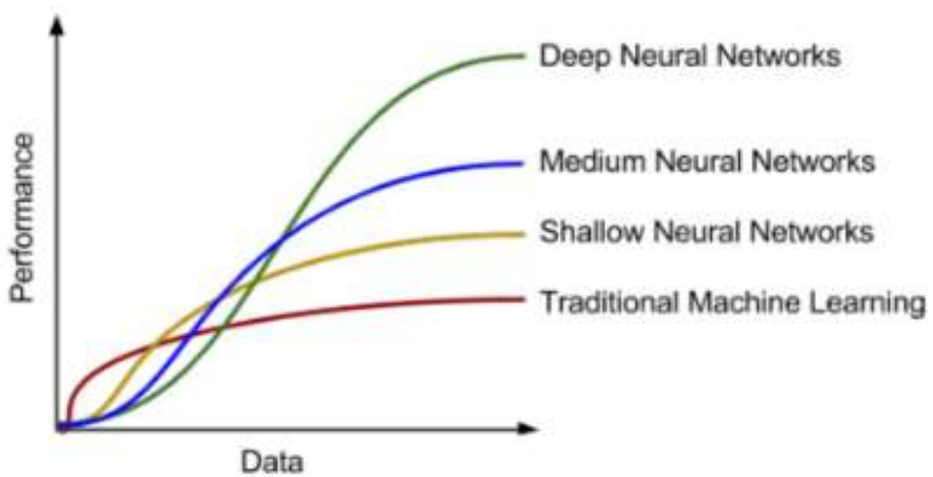
Предобработка речи





Speech Recognition

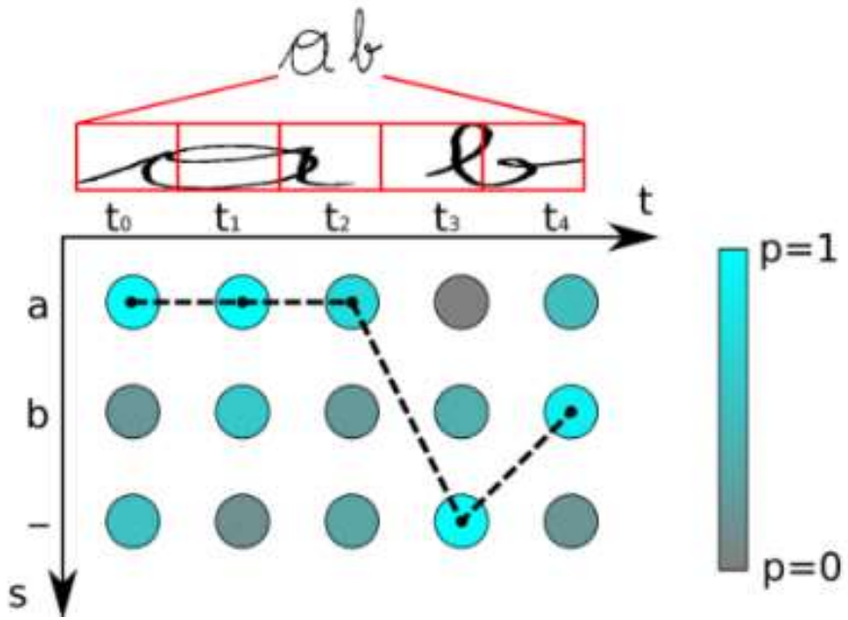
- Речь -> ... -> Фонемы -> Словарь
  - TIMIT Database (en) - 3h, 1993
  - CMUDict (en) - 100k, 1993
- Разбили сложную задачу на две более простых... но они требуют разметки фонем, составления словаря и алгоритма поиска
- Слово "IPA"= [aɪ pʰiː eɪ]



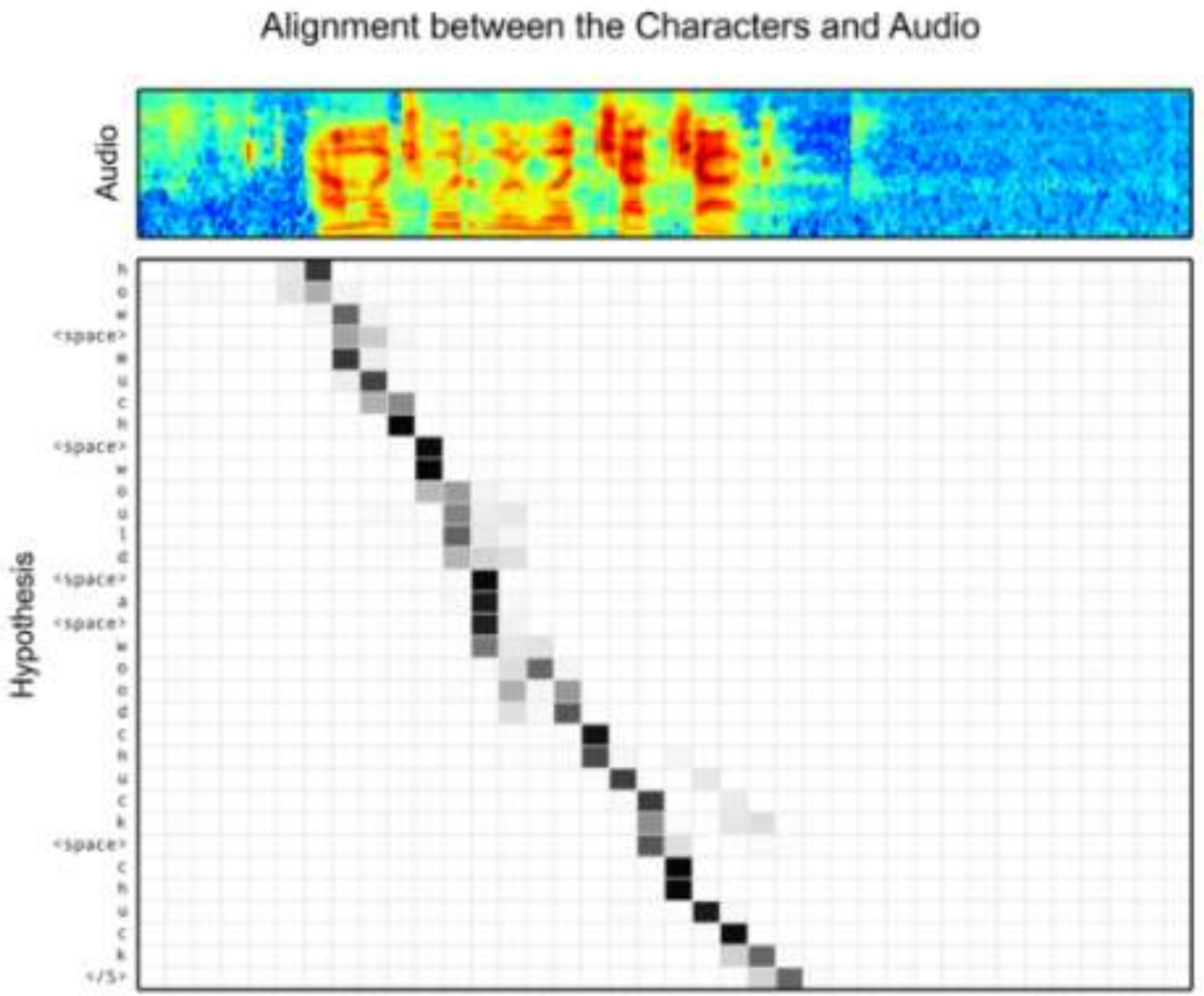
- Речь -> ... -> Буквы -> Текст
- Открытые датасеты:
  - WSJ (en) - 81h, 1993
  - Switchboard (en) - 240h, 1993
  - VoxForge (ru) - 17h, 2009

Авторазметка (Alignment)

Forward-backward algorithm



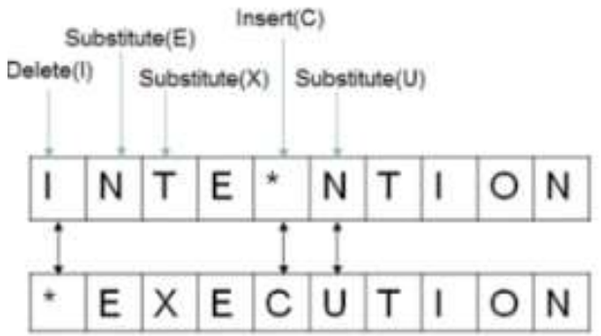
Каждой возможной букве или звуку сопоставляется вероятность, а потом нужно построить путь по матрице вероятностей, минимизирующий некоторую метрику...



CTC (Connectionists Temporal Classification)

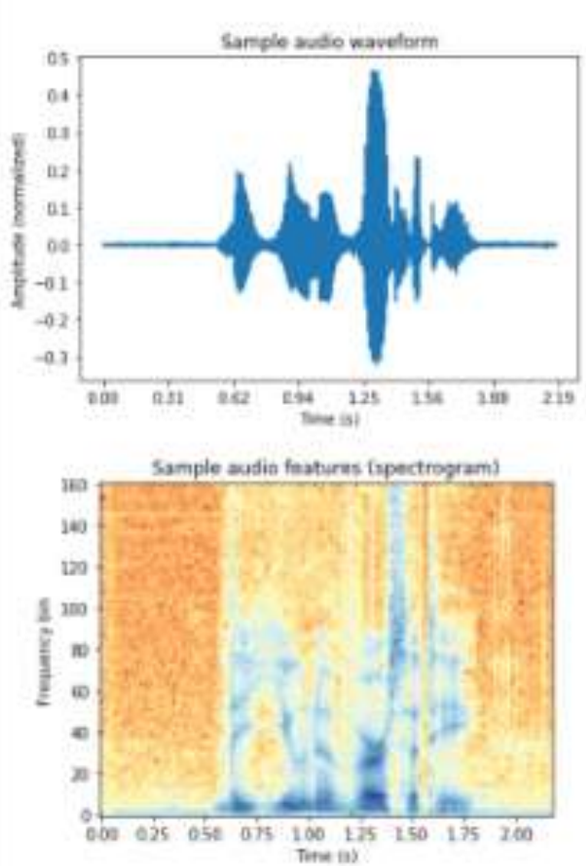
		E	L	E	P	H	A	N	T
	0	1	2	3	4	5	6	7	8
R	1	1	2	3	4	5	6	7	8
E	2	1	2	2	3	4	5	6	7
L	3	2	1	2	3	4	5	6	7
E	4	3	2	1	2	3	4	5	6
V	5	4	3	2	2	3	4	5	6
A	6	5	4	3	3	3	3	4	5
N	7	6	5	4	4	4	4	3	4
T	8	7	6	5	5	5	5	4	3

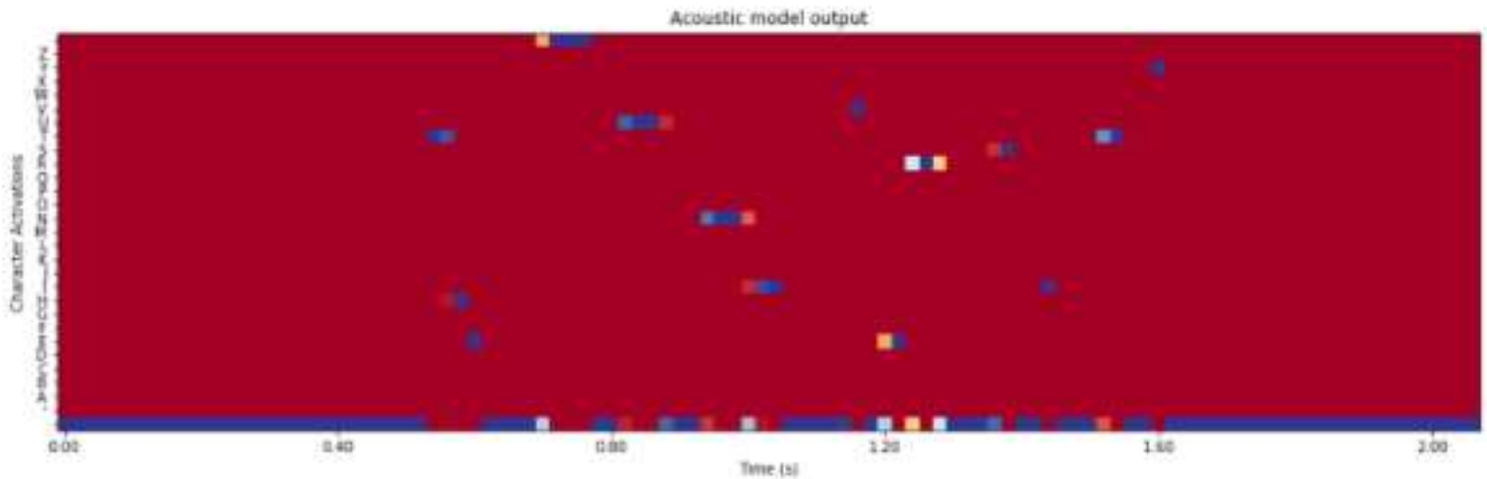
Штрафует нейросеть за несоответствия в тех местах, где произошли ошибки.



Эта метрика называется Edit Distance, или расстояние Левенштейна. Это метрика, позволяющая определить «схожесть» двух строк — минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

Рассмотрим пример






```
                THE ___ U   N I   VE R   S  I  T Y
Error:              IRR                    III EEE SSS
Decoder: THE UNIVERSITY
Correct: THEIR UNIVERSITIES
```

Синтез речи

Задачу тоже приходится делить на части:

- Генерация спектрограммы (predictor)
- Генерация звука (vocoder)

Датасеты:

- LJ Speech (24часа)

Давайте рассмотрим классические методы синтеза.

- Конкатенативный синтез речи:

основан на предварительной записи коротких аудио-фрагментов, которые затем объединяются для создания связной речи. Она получается очень чистая и ясная, но абсолютно лишена эмоциональной и интонационной составляющих, то есть звучит неестественно. Применение конкатенативного TTS ограничено из-за больших требований к данным и времени разработки.

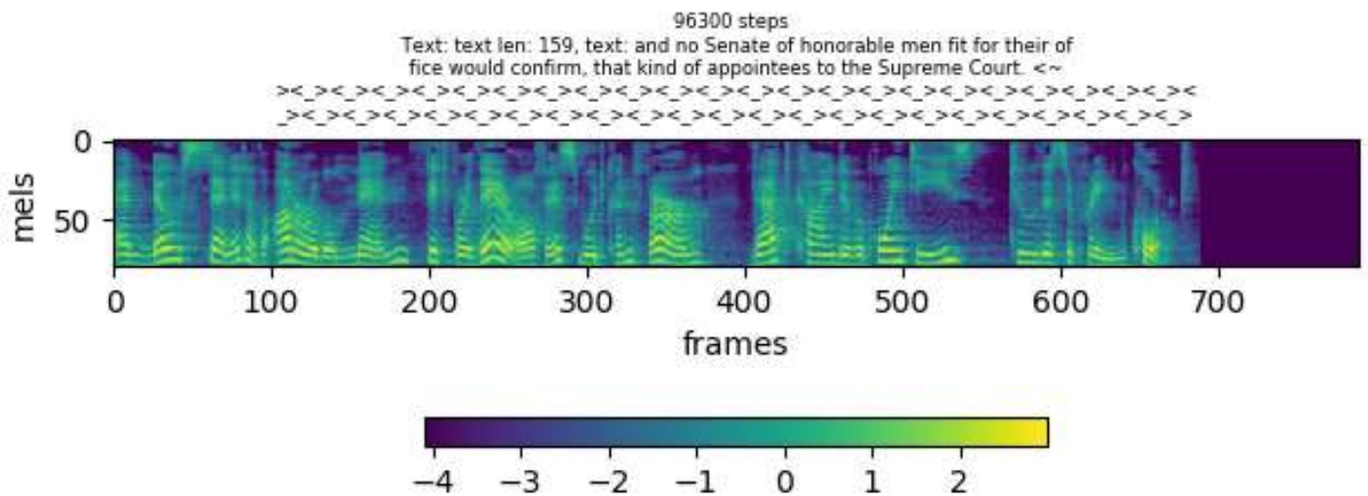
- Параметрический синтез речи:

исследует саму природу данных. Он генерирует речь с помощью комбинирования таких параметров, как частота, спектр амплитуд и т.д.

Этапы параметрического синтеза

- Сначала из текста извлекаются лингвистические признаки, например, фонемы, продолжительность и т.д.
- Затем для вокодера (системы, генерирующей wave-формы) извлекаются признаки, которые представляют соответствующий речевой сигнал: мел-спектрограмма.
- Эти, настроенные вручную, параметры наряду с лингвистическими особенностями передаются в модель вокодера, а тот выполняет множество сложных преобразований для генерирования звуковой волны.

Предобработка данных



Мел-спектрограмма аудиосигнала речи, приведенная к диапазону [-4;4].

https://keithito.com/LJ-Speech-Dataset/

Архитектура

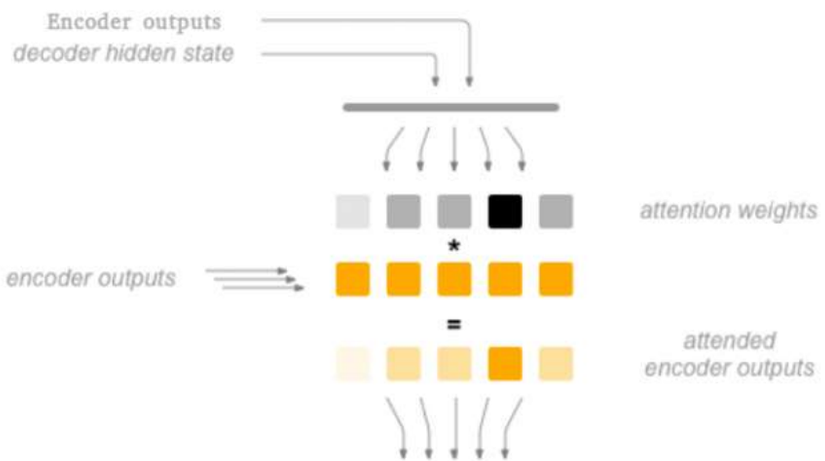
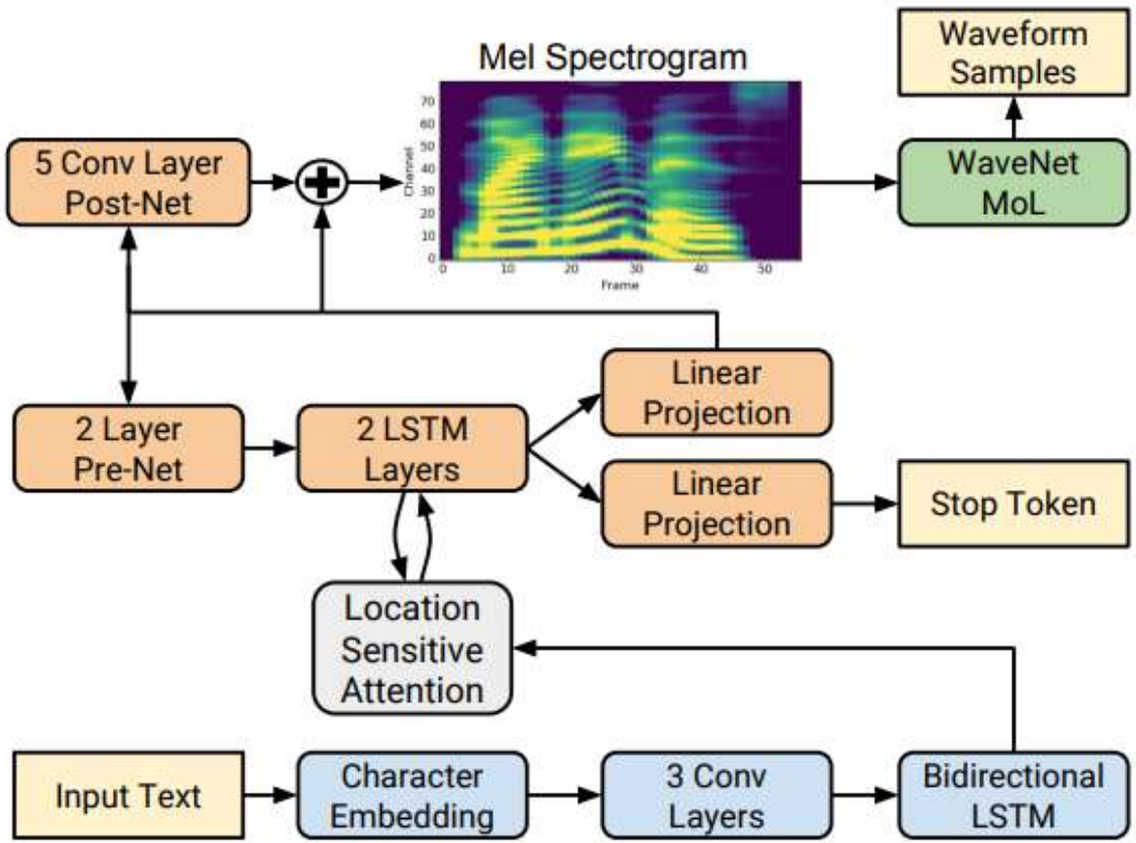


Схема работы механизма внимания



Архитектура сети Tacotron 2.