Решение задачи распознавания речи

Щуров Алексей Андреевич
M9102
Руководитель: Кленин А.С.

Тяжелые случаи

- Фоновый шум
  Эффект Ломбарда
- Диктор далеко от микрофона
- Несколько дикторов
- Сложный язык

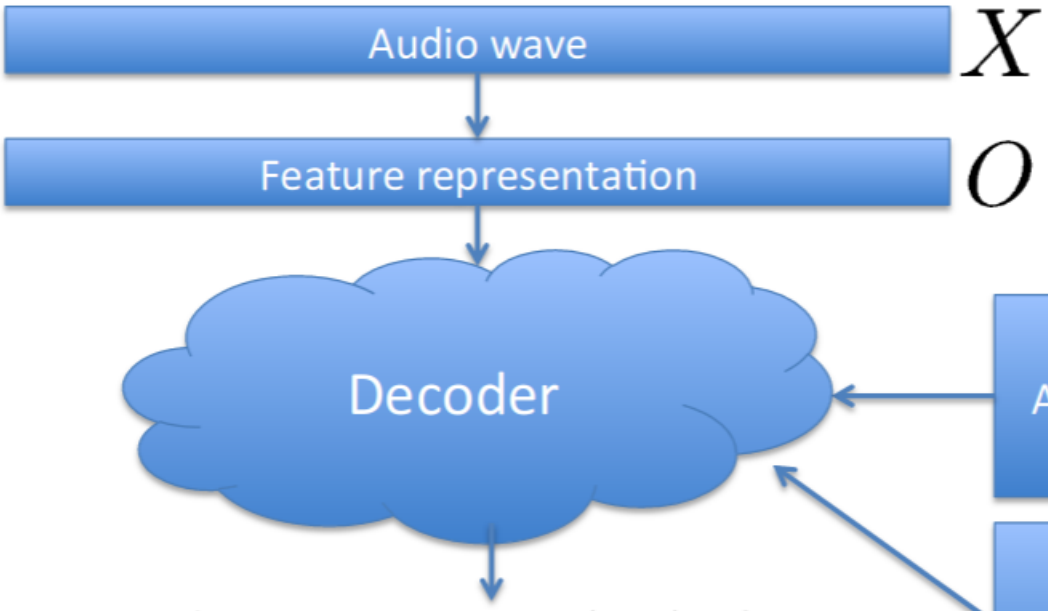
Цели

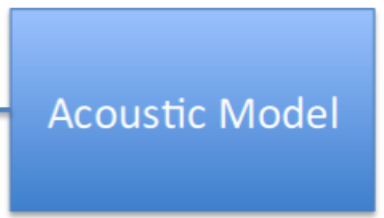
- Достичь сопоставимого с коммерческими системами качества распознавания в сложных условиях

Открытые решения

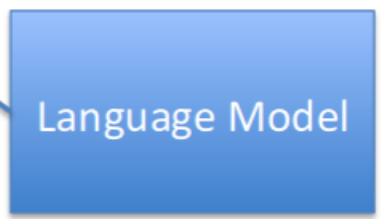
- Mozilla Deep Speech
- Wav2letter
- Kaldi
- Аутсайдеры

Традиционная стратегия распознавания





$P(O|W)$



$P(W)$

$$W^* = \underset{W}{\operatorname{argmax}}\ P(W|X)$$

$$= \underset{W}{\operatorname{argmax}}\ P(O|W)P(W)$$

Gales & Young, 2008
Jurafsky & Martin, 2000

Mozilla Deep Speech

- Baidu research (2014)
- RNN
- 6 слоев

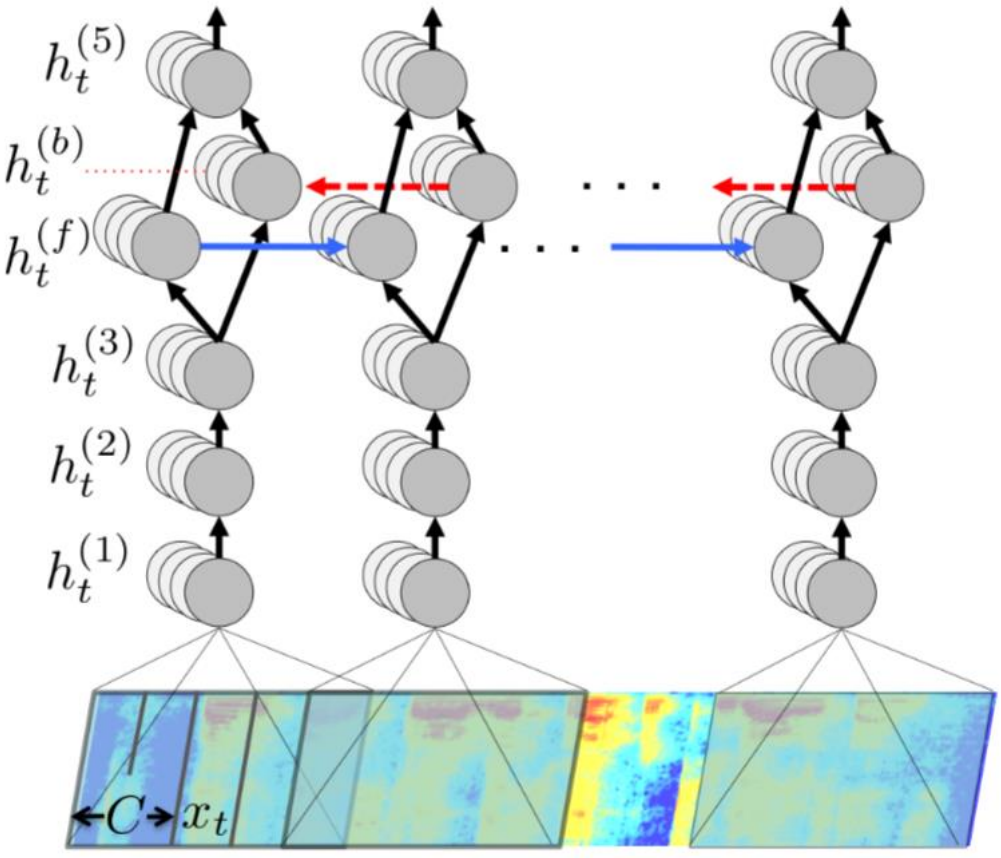
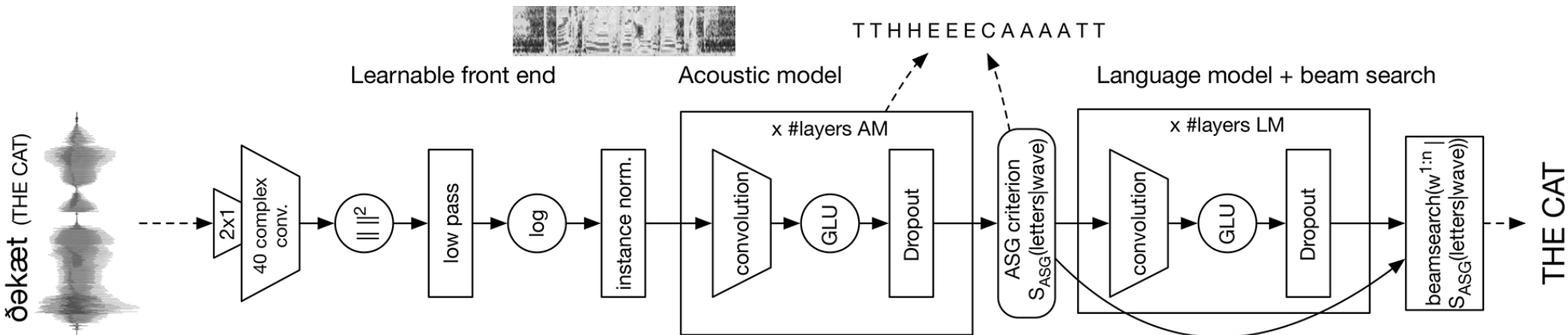


Figure 1: Structure of our RNN model and notation.

Wav2Letter

- Facebook (2016, 2019)
- Необходимо ~1000 часов речи для обучения



Функциональные требования

- Русский язык
- Online-декодирование

Преймущества Kaldi

- Хорошо документирована
- Есть модели для русского языка
- 400 часов за 5 часов

Выполнено

- Высокое качество распознавания в идеальных условиях
- Online-декодирование

План

- Исследовать возможности повышения надежности распознавания
- 4000 часов