Особенности алгоритмов распознавания речи и их влияние на опыт пользователей

Шмырёв Н.В.

АЦ Технологии

АЦ Технологии

Воск

https://github.com/alphacep/vosk-api

```
pip3 install vosk
```

Vosk

- Работает без доступа к сети даже на мобильных устройствах - Raspberry Pi, Android, iOS
- Поддерживает 18 языков и диалектов - русский, английский, индийский английский, немецкий, французский, португальский, испанский...
- Работает на серверах и интегрируется в системы телефонии (Asterisk, Freeswitch, Unimrcp, Jitsi)
- Поддерживает несколько популярных языков программирования - Java, C#, Javascript, Swift, Python

Vosk

- Сделана для потоковой обработки звука, что позволяет реализовать мгновенную реакцию на команды
- Модели для каждого языка занимают всего 50Мб, но есть и гораздо более точные большие модели для более точного распознавания
- Позволяет быстро настраивать словарь распознавания для улучшения точности распознавания
- Позволяет идентифицировать говорящего

Не всё так замечательно

- Тяжело добавлять слова в словарь
- Не поддерживается активация по ключевому слову
- Точность зависит от задачи
- Не поддерживает GPU

Речевые технологии 2011-2021

Десятилетие нейросетей

- Распознавание речи (4x более more accurate, about 80% accurate for realistic call-center and video)
- Синтез речи (3/5 MOS $\rightarrow$ 4.5/5 MOS)
- Сжатие речи (40% MUSRA quality at 2.4kbps $\rightarrow$ 80% quality at 1.6kbps for LPCnet codec)
- Проверка голоса (0.5% EER for 10k speakers on 10 second sample)
- Преобразование голоса (5 seconds to good clone)
- Очистка шумов, обнаружение искусственной речи, коррекция заиканий, эмоции...

Быстрое развитие технологий

Распознавание речи: DNN(2011) $\rightarrow$ LSTM(2014) $\rightarrow$ TDNN(2015) $\rightarrow$ Deepspeech(2016) $\rightarrow$ Transformers(2017) $\rightarrow$ Conformer(2019) $\rightarrow$ Wav2Vec2.0(2020) $\rightarrow$ Conformer + LAS(2021)

Синтез речи: Wavenet(2016) $\rightarrow$ WaveRNN(2017) $\rightarrow$ Waveglow(2018) $\rightarrow$ Melgan(2019) $\rightarrow$ MBMelgan(2020) $\rightarrow$ HiFi-Gan(2021) $\rightarrow$ Diffusion networks (2021)...

Библиотеки для распознавания речи

- Kaldi/K2/Lhotse/Snowfall (распознавание)
- Espnet (распознавание, синтез, перевод речи, проверка голоса)
- Nvidia Nemo (распознавание, синтез, обработка языка)
- Facebook Flashlight(formerly wav2letter) (распознавание)
- Mozilla Deepspeech/TTS/Coqui(распознавание, синтез)
- Speechbrain(распознавание, идентификация, разделение)
- Wenet(распознавание)
- Много много других - Tensorflow/lingvo, Didi/delta ...

Общепринятые методы оценки

- Точность распознавания
- Задержка результатов
- Пропускная способность

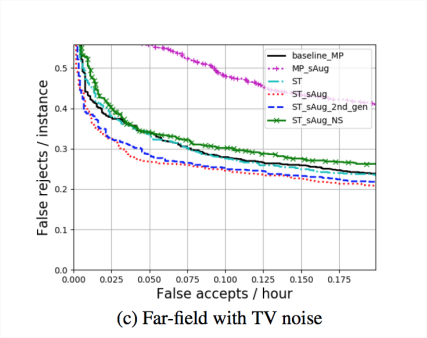
Поиск ключевого слова/фразы

Noisy student-teacher training for robust keyword spotting

Hyun-Jin Park, Pai Zhu, Niranjan Subrahmanya, Ignacio Lopez Moreno

https://arxiv.org/pdf/2106.01604v1.pdf

- Точность распознавания
- Количество ложных срабатываний



(c) Far-field with TV noise

Умный поиск ключевой фразы

Domain-Specific Utterance End-Point Detection for Speech Recognition Roland Maas, Ariya Rastrow, Kyle Goehner, Gautam Tiwari, Shaun Joseph, Björn Hoffmeister

https://www.isca-speech.org/archive/Interspeech_2017/pdfs/1673.PDF

- Не обязательно искать одну фразу, можно распознавать команды
- Можно обрабатывать интонации
- Снижаются ложные срабатывания
- Повышается время реакции

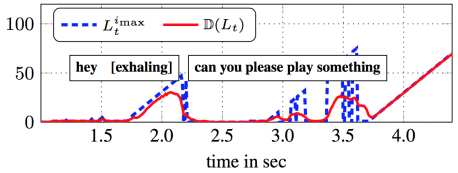


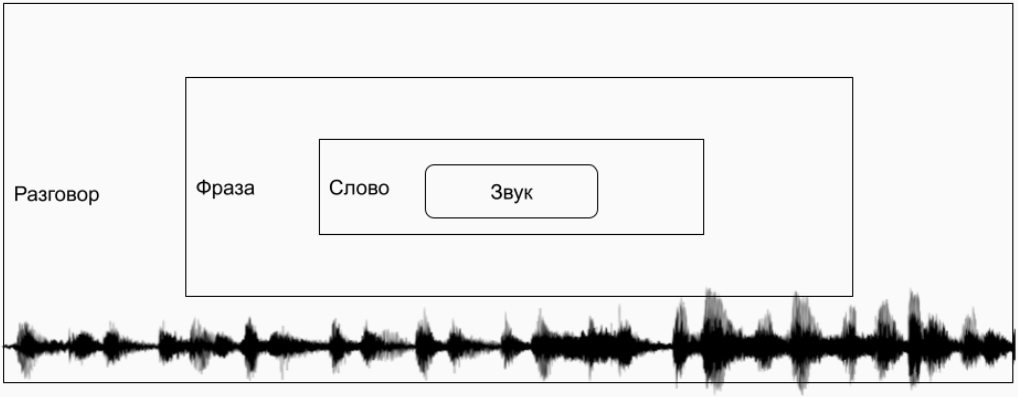
Figure 1: *Feature value comparison of the best path pause duration* $L_t^{i\max}$ *and the expected pause duration* $\mathbb{D}(L_t)$ *for an example utterance.*

Using Complexity-Identical Human- and Machine-Directed Utterances to Investigate Addressee Detection for Spoken Dialogue Systems

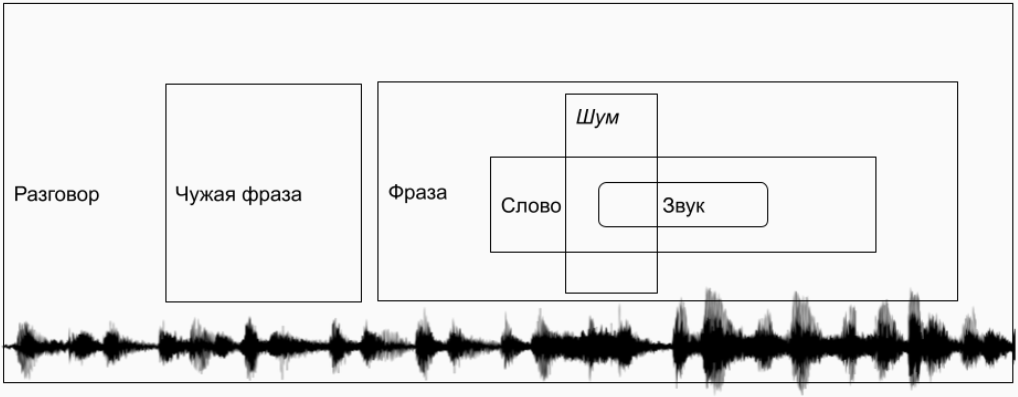
Oleg Akhtiamov, Ingo Siegert, Alexey Karpov, Wolfgang Minker

https://doi.org/10.3390/s20092740

Контекст



Контекст



Контекст

- DNN $\rightarrow$ DNN + i-вектора $\rightarrow$ Трансформер $\rightarrow$ Трансформер с многоголовым вниманием $\rightarrow$ Трансформер с LAS $\rightarrow$ ...
- У Wav2Vec всё плохо с контекстом
- В NLP тоже борьба за контекст: Longformer, Reformer, Linformer $\rightarrow$ Compressive Transformer
- Для тренировки редко моделируется сложный контекст

Контекст

- Чем больше контекста, тем точнее распознавание
- Но: гораздо больше вероятность ошибки
- Но: больше задержка на обработку

Ещё контекст

Vosk TDNN + LM - можно подстраивать языковую модель на лету в зависимости от ситуации

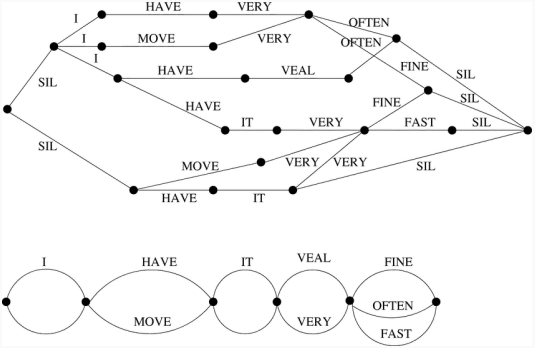
Transformer комбинирует языковую модель в нейросети - подстраивать её гораздо сложнее

Internal Language Model Estimation for Domain-Adaptive End-to-End Speech Recognition Zhong Meng, Sarangarajan Parthasarathy, Eric Sun, Yashesh Gaur, Naoyuki Kanda, Liang Lu, Xie Chen, Rui Zhao, Jinyu Li, Yifan Gong

https://arxiv.org/pdf/2011.01991.pdf

Внутреннее представление результатов декодирования

- наилучший результат (1-best)
- n-лучших результатов (n-best)
- взвешенный граф слов (lattice)
- альтернативы слов (confusion network)
- скрытое состояние
- скрытое распределение



Преимущества графа результатов

- компактное представление
- быстрая переоценка результатов, в том числе и с использованием сложной логики
- возможность отслеживания стабильного промежуточного результата $\rightarrow$ быстрая реакция на речь пользователя

Стабильность результатов в потоковом распознавании

включи с

включи свет

включи свет в

включи свет в кор

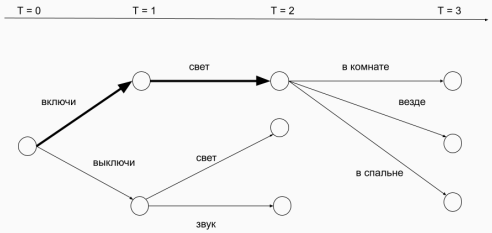
включи свет в корзине

включи свет в коридоре

$\rightarrow$ действие

включи свет в коридоре и

включи свет в коридоре и спальне....



Заключение

- Внедрение систем распознавания речи должно опираться на особенности заложенных в них алгоритмов, существенно влияющих на характеристики систем и определяющих взаимодействие с пользователем
- Многое осталось за кадром – определение голоса, подстройка под пользователя, скорость и меры надёжности распознавания
- В распознавании речи ещё будет много интересного

Заходите на огонёк

Telegram: https://t.me/speech_recognition_ru

Telegram: https://t.me/speechtech

Web: https://alphacephei.com

Github: https://github.com/alphacep

Email: contact@alphacephei.com