академия больших данных

mail.ru group

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

ЦРТ ГРУППА КОМПАНИЙ

Обработка речевых сигналов

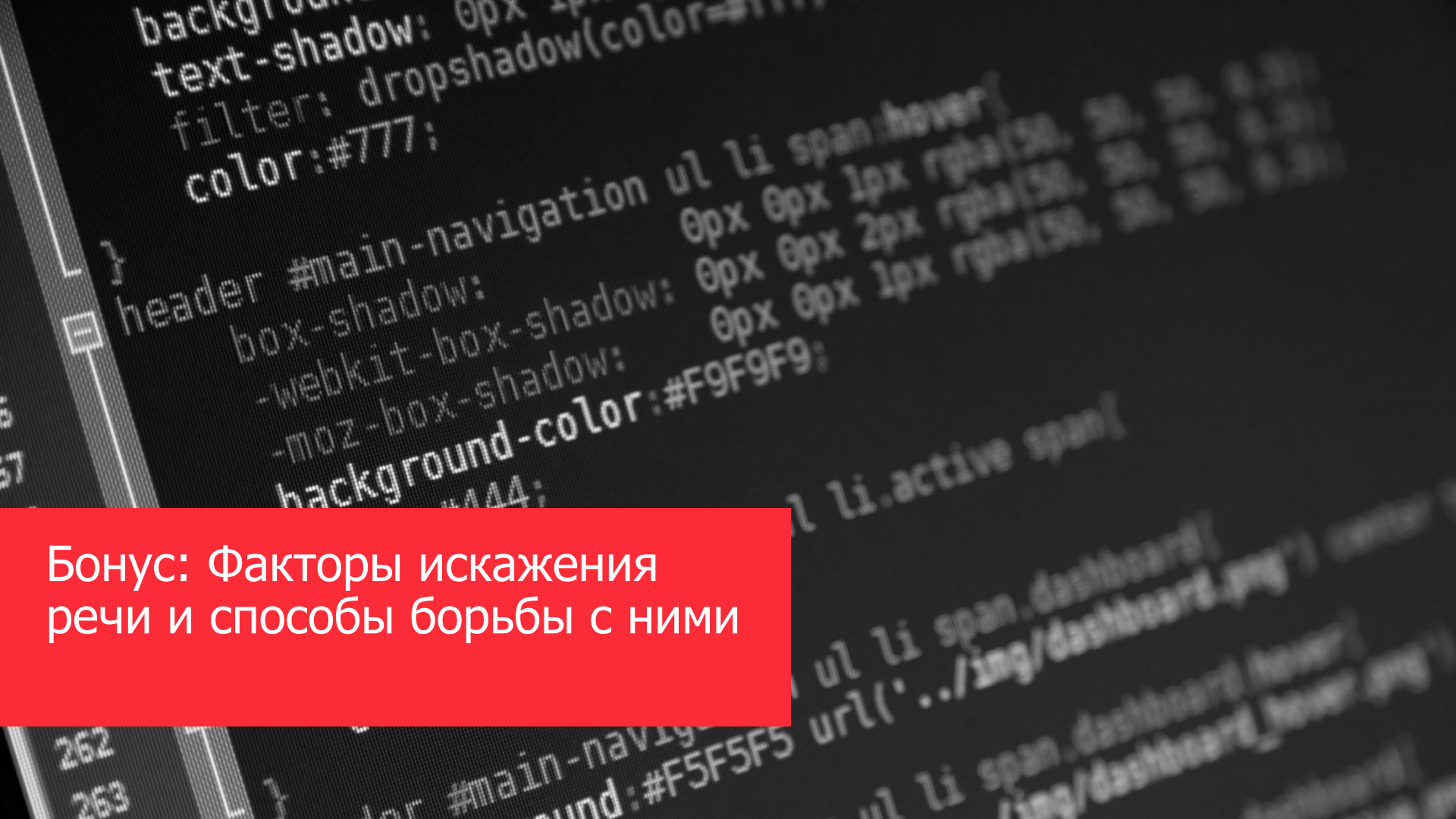
Блок 2. Автоматическое распознавание речи

Максим Кореневский

Старший научный сотрудник

ООО «ЦРТ-инновации», к.ф.-м.н.

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


Бонус: Факторы искажения речи и способы борьбы с ними

Факторы искажения речи и способы борьбы с ними

Факторы, влияющие на точность распознавания речи:

- Разнообразие стилей речи
  - Подготовленная vs. спонтанная речь, чтение vs. выступление без бумажки и т.д.
- Междикторская и внутридикторская вариативность
  - Разная высота голоса, акценты, четкость произношения, меняющийся темп речи
- Разнообразие условий записи
  - Канал передачи речи (воздух, характеристика микрофона)
  - Шумы и помехи (кондиционер, телевизор, шум транспорта и т.п.)
  - Реверберация (переотражение от стен и предметов)
  - Расстояние до микрофона (ближний/дальний)
  - Конкурирующая речь (неформальная беседа, совещание и т.д.)

Факторы искажения речи и способы борьбы с ними

Два основных направления

- Снижение вариативности в данных перед обработкой (адаптивное обучение)
  - VTLN
  - Шумоподавление/шумоочистка
  - Дереверберация
  - Снижение зависимости от канала
  - Робастные акустические признаки

Факторы искажения речи и способы борьбы с ними

Два основных направления

- Снижение вариативности в данных перед обработкой (адаптивное обучение)
  - VTLN
  - Шумоподавление/шумоочистка
  - Дереверберация
  - Снижение зависимости от канала
  - Робастные акустические признаки
- Повышение вариативности в обучающих данных
  - Набор/запись различных акустических баз (стиль речи, пол, возраст социальное положение, акценты)
  - Multi-Condition Training
  - Активная data augmentation

Снижение вариативности входных данных

Шумоподавление/шумоочистка

- Подходы на основе методов обработки сигналов (DSP)
  - Оценка параметров шума и фильтрация (Винера, Калмана)
  - Вейвлетная обработка

Снижение вариативности входных данных

Шумоподавление/шумоочистка

- Подходы на основе методов обработки сигналов (DSP)
  - Оценка параметров шума и фильтрация (Винера, Калмана)
  - Вейвлетная обработка
- Подходы на основе использования модели искажения речи
  - Модель искажения речи: $x(t) = s(t) * h + n(t)$. Более простая модель $x(t) = s(t) + n(t)$
  - Оценка шума и «спектральное вычитание»
  - Метод разложения в векторный ряд Тейлора (Vector Tailor Series) для компенсации искажений признаков

Снижение вариативности входных данных

Шумоподавление/шумоочистка

- Подходы на основе методов обработки сигналов (DSP)
  - Оценка параметров шума и фильтрация (Винера, Калмана)
  - Вейвлетная обработка
- Подходы на основе использования модели искажения речи
  - Модель искажения речи: $x(t) = s(t) * h + n(t)$. Более простая модель $x(t) = s(t) + n(t)$
  - Оценка шума и «спектральное вычитание»
  - Метод разложения в векторный ряд Тейлора (Vector Tailor Series) для компенсации искажений признаков
- Подходы на основе неотрицательной матричной факторизации (NMF)
  - Предполагается, что чистую речь можно представить как линейную комбинацию базовых «примитивов»
  - Словарь примитивов обучается по большим объемам чистой речи
  - На зашумленной речи оцениваются коэффициенты разложения, и речь «восстанавливается» из примитивов

Снижение вариативности входных данных

Шумоподавление/шумоочистка/разделение источников

- Подходы на основе машинного обучения
  - Генерация зашумленных данных из чистых и обучение нейросети для удаления шума (DNN, BLSTM, CNN)
  - Шумоподавляющие автоэнкодеры
  - Глубокие генеративные модели (VAE и GAN)
  - Нейронные модели для Speech Separation (например, Deep Clustering)

Снижение вариативности входных данных

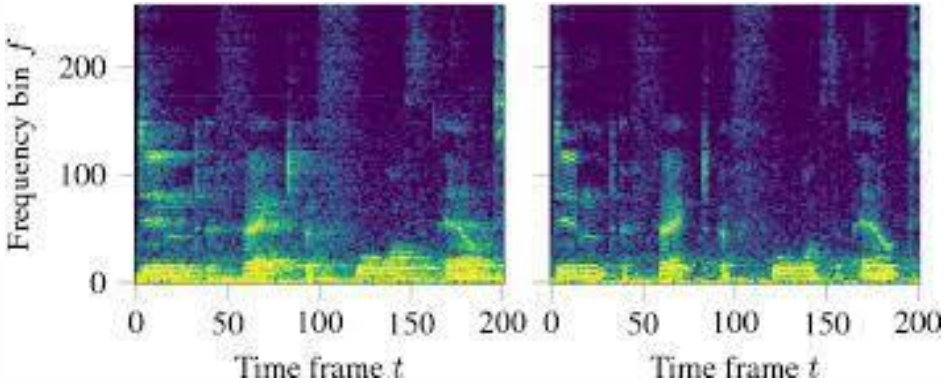
Шумоподавление/шумоочистка/разделение источников

- Подходы на основе машинного обучения
  - Генерация зашумленных данных из чистых и обучение нейросети для удаления шума (DNN, BLSTM, CNN)
  - Шумоподавляющие автоэнкодеры
  - Глубокие генеративные модели (VAE и GAN)
  - Нейронные модели для Speech Separation (например, Deep Clustering)
- Многомикрофонное распознавание речи
  - Массивы микрофонов работают, как антенна
  - Простейший аналог – бинауральный слух
  - С помощью «формирования луча» (beamforming) речь с определенного направления усиливается, а с прочих направлений подавляется
  - Выбор направления на целевой источник (Direction of Arrival, DoA) – отдельная сложная задача, но есть подходы, свободные от этого

Снижение вариативности входных данных

Деверберация

- Сложность в том, что переотраженный звук коррелирует с прямым



- WPE (Weighted prediction error)-деверберация
  - В каждой спектральной полосе делается «линейное предсказание»:

$$x_t = \sum_{k=1}^{L} w_k x_{t-k} + e_t = \sum_{k=1}^{D} w_k x_{t-k} + \sum_{k=D+1}^{L} w_k x_{t-k} + e_t = d_t + r_t$$

  - Лучше работает, когда есть записи с нескольких микрофонов, и спектр предсказывается по всем каналам

Снижение вариативности входных данных

Снижение зависимости от канала и стационарного шума

- Cepstral Mean Vormalization (CMN)
  - В отсутствии шума модель сигнала: $x(t) = s(t) * h$
  - В спектральной (STFT) области: $X(t, \omega) = S(t, \omega)H(\omega)$
  - В лог-спектральной области (и в кепстральной): $\log|X(t, \omega)|^2 = \log|S(t, \omega)|^2 + \log|H(\omega)|^2$
  - Если усреднить по времени, получим: $\overline{\log|X(t, \omega)|^2} = \overline{\log|S(t, \omega)|^2} + \log|H(\omega)|^2$
  - Поэтому, вычитая среднее, мы избавляемся от характеристики канала!

Снижение вариативности входных данных

Снижение зависимости от канала и стационарного шума

- Cepstral Mean Vormalization (CMN)
  - В отсутствии шума модель сигнала: $x(t) = s(t) * h$
  - В спектральной (STFT) области: $X(t, \omega) = S(t, \omega)H(\omega)$
  - В лог-спектральной области (и в кепстральной): $\log|X(t, \omega)|^2 = \log|S(t, \omega)|^2 + \log|H(\omega)|^2$
  - Если усреднить по времени, получим: $\overline{\log|X(t, \omega)|^2} = \overline{\log|S(t, \omega)|^2} + \log|H(\omega)|^2$
  - Поэтому, вычитая среднее, мы избавляемся от характеристики канала!
- Cepstral Mean and Variance Normalization
  - В общем случае присутствует не только канал, но и шум
  - Можно приводить кепстр к нулевому среднему и единичной дисперсии – это снижает зависимость от шума и канала.

Снижение вариативности входных данных

Снижение зависимости от канала и стационарного шума

- Cepstral Mean Vormalization (CMN)
  - В отсутствии шума модель сигнала: $x(t) = s(t) * h$
  - В спектральной (STFT) области: $X(t, \omega) = S(t, \omega)H(\omega)$
  - В лог-спектральной области (и в кепстральной): $\log|X(t, \omega)|^2 = \log|S(t, \omega)|^2 + \log|H(\omega)|^2$
  - Если усреднить по времени, получим: $\overline{\log|X(t, \omega)|^2} = \overline{\log|S(t, \omega)|^2} + \log|H(\omega)|^2$
  - Поэтому, вычитая среднее, мы избавляемся от характеристики канала!
- Cepstral Mean and Variance Normalization
  - В общем случае присутствует не только канал, но и шум
  - Можно приводить кепстр к нулевому среднему и единичной дисперсии – это снижает зависимость от шума и канала.
- RASTA (RelAtive SpecTrAl)-обработка: подавление слишком медленных и слишком быстрых изменений в лог-спектре/кепстре.

Снижение вариативности входных данных

Робастные акустические признаки

- Робастность (robustness) = устойчивость к искажениям входных данных
- Основной подход: на основе теории слухового восприятия (auditory features)
  - Процесс обработки звука имитирует многие механизмы слуховой системы
  - Весьма трудозатратный процесс
  - Наиболее известные признаки: Power-Normalized Cepstral Coefficients (PNCC), Gabor & Gammatone filterbanks

Снижение вариативности входных данных

Робастные акустические признаки

- Робастность (robustness) = устойчивость к искажениям входных данных
- Основной подход: на основе теории слухового восприятия (auditory features)
  - Процесс обработки звука имитирует многие механизмы слуховой системы
  - Весьма трудозатратный процесс
  - Наиболее известные признаки: Power-Normalized Cepstral Coefficients (PNCC), Gabor & Gammatone filterbanks
- Альтернативный подход: на основе артикуляции (articulatory features)
  - Для каждой фонемы языка известны ее артикуляционные характеристики
  - Если есть разметка на фонемы, можно обучить классификатор по каждой артикуляционной характеристике

Снижение вариативности входных данных

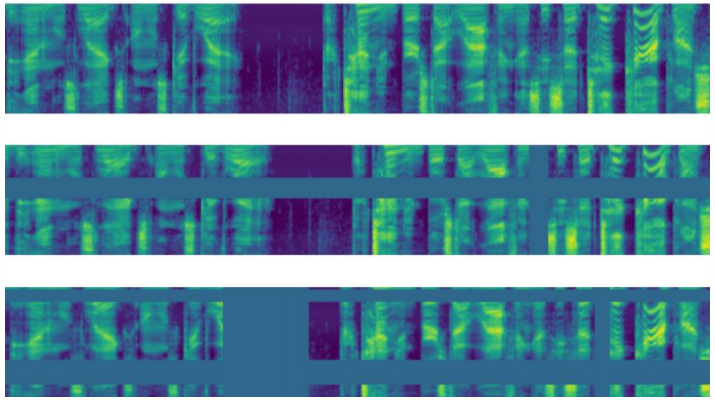
Робастные акустические признаки

- Робастность (robustness) = устойчивость к искажениям входных данных
- Основной подход: на основе теории слухового восприятия (auditory features)
  - Процесс обработки звука имитирует многие механизмы слуховой системы
  - Весьма трудозатратный процесс
  - Наиболее известные признаки: Power-Normalized Cepstral Coefficients (PNCC), Gabor & Gammatone filterbanks
- Альтернативный подход: на основе артикуляции (articulatory features)
  - Для каждой фонемы языка известны ее артикуляционные характеристики
  - Если есть разметка на фонемы, можно обучить классификатор по каждой артикуляционной характеристике
- Еще один подход: модуляционные признаки
  - В спектре присутствуют медленные модуляции, характеризующие слоговую структуру
  - Модуляционный спектр очень устойчив к искажениям

Повышение вариативности обучающих данных

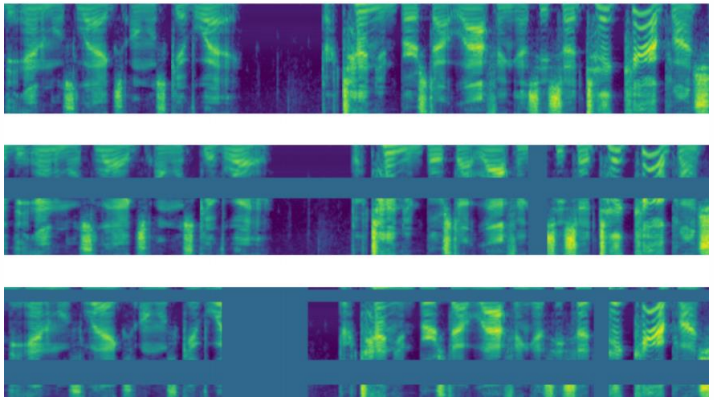
Data augmentation

- Простые подходы:
  - Speed/Tempo/Pitch Perturbation
  - Volume Perturbation
  - SpecAugment (см. картинку)
  - Аугментация кодеками с потерями



Повышение вариативности обучающих данных

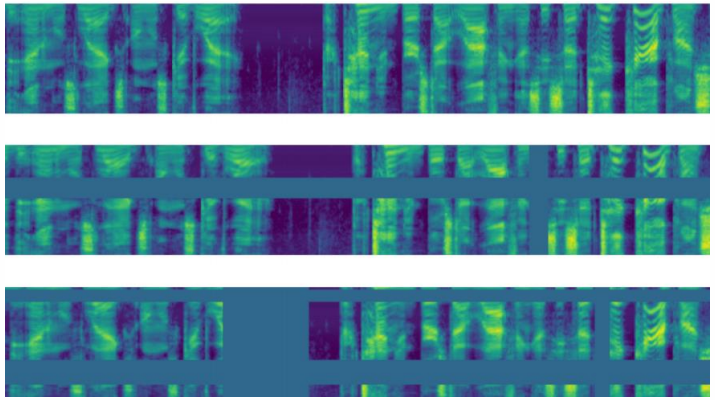
Data augmentation



- Простые подходы:
  - Speed/Tempo/Pitch Perturbation
  - Volume Perturbation
  - SpecAugment (см. картинку)
  - Аугментация кодеками с потерями
- Зашумление и реверберация
  - В интернете есть множество различных шумов, шум добавляется к сигналу с заданным SNR
  - В интернете есть множество импульсных характеристик различных помещений (Room Impulse Response, RIR)
  - RIRы можно генерировать искусственно, задавая размеры помещения и положения источников и микрофона

Повышение вариативности обучающих данных

Data augmentation



- Простые подходы:
  - Speed/Tempo/Pitch Perturbation
  - Volume Perturbation
  - SpecAugment (см. картинку)
  - Аугментация кодеками с потерями
- Зашумление и реверберация
  - В интернете есть множество различных шумов, шум добавляется к сигналу с заданным SNR
  - В интернете есть множество импульсных характеристик различных помещений (Room Impulse Response, RIR)
  - RIRы можно генерировать искусственно, задавая размеры помещения и положения источников и микрофона
- MixUp и подобные подходы
  - Если есть пример x_1 с one-hot меткой y_1 и пример x_2 с one-hot меткой y_2, выберем число $0 \leq \alpha \leq 1$ и добавим к обучающим данным пример $\alpha x_1 + (1 - \alpha) x_2$ с меткой $\alpha y_1 + (1 - \alpha) y_2$

Повышение вариативности обучающих данных

Data augmentation

- Многократно увеличивает объемы обучающих данных
- Возрастает время обучения
- Для новых типов искажений приходится дополнять выборку и переучиваться

Повышение вариативности обучающих данных

Data augmentation

- Многократно увеличивает объемы обучающих данных
- Возрастает время обучения
- Для новых типов искажений приходится дополнять выборку и переучиваться
- Онлайн подход:
  - Задаем набор шумов и RIR-ов
  - На каждой эпохе проводим data augmentation «на лету» выбирая шумы, RIR и SNR-ы случайно из некоторого распределения
  - Также на лету можно проводить SP, VP, SpecAugment и MixUp
  - На каждой эпохе модель видит различные типы искажений, но суммарный объем данных на эпоху НЕ РАСТЕТ!
  - Можно регулировать степень искажений и, например, увеличивать ее в ходе обучения (annealing)

Спасибо за внимание!

Вопросы?