Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Факультет: «Информатика и системы управления»

Кафедра: «Программное обеспечение ЭВМ и информационные технологии»

Расчетно-пояснительная записка по НИР на тему: «Метод распознавания звуков в звучащей речи на Естественном Языке»

Студент: Левушкин И. К.
Группа: ИУ7-72Б
Научный руководитель: Градов В.М.

Москва, 2020 г.

Содержание

Введение **3**

1 Аналитический раздел **4**

1.1 Классический подход к распознаванию речи 4

1.2 Входные данные 4

1.3 Извлечение признаков (Feature Extraction) 5

1.4 MFCC 6

1.4.1 Разложение в ряд Фурье 7

1.4.2 Расчет mel-фильтров 7

1.5 Применение фильтров, логарифмирование энергии спектра 8

1.6 Косинусное преобразование 9

1.7 Акустическая модель 9

2 Конструкторский раздел **10**

3 Технологический раздел **11**

Заключение **12**

Список используемой литературы **13**

Введение

Распознавание речи — одна из самых интересных и сложных задач искусственного интеллекта. Здесь задействованы достижения весьма различных областей: от компьютерной лингвистики до цифровой обработки сигналов.

1 Аналитический раздел

1.1 Классический подход к распознаванию речи

Классический подход к распознаванию речи представляет собой последовательность следующих действий (этапов):

- Извлечение признаков (Feature Extraction)
- Построение и обучение акустической модели (Acoustic model)
- Декодер, выбирающий наиболее вероятный путь перехода по HCLG-графу:
  - H модуль на базе HMM
  - C модуль контекстной зависимости
  - L модуль произношения
  - G модуль языковой модели
- Rescoring - перевзвешивание гипотез и выдача окончательного результата

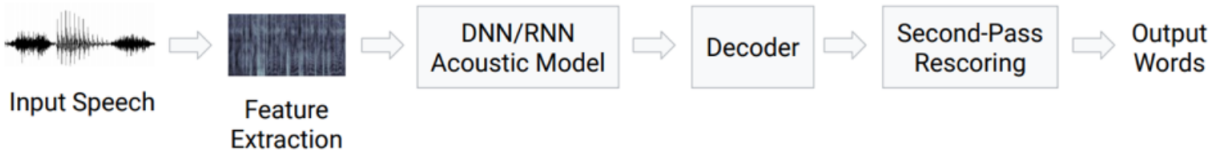

Рисунок 1: Классический подход.

1.2 Входные данные

Для начала нужно понимать, что наша речь - это последовательность звуков. Звук в свою очередь — это суперпозиция (наложение) звуковых колебаний (волн) различных частот. Волна характеризуются двумя атрибутами — амплитудой и частотой.

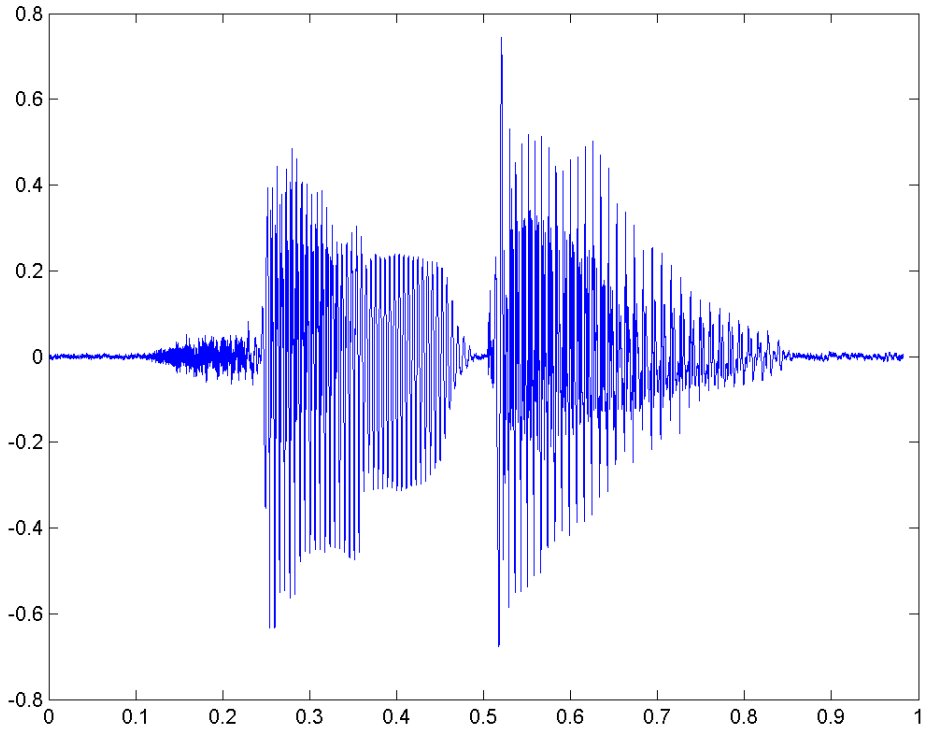

Рисунок 2: Пример звуковой волны.

Другими словами, наши данные - это записанные дорожки аудиофайлов, которые необходимо далее преобразовать в вектор признаков для обучения модели.

1.3 Извлечение признаков (Feature Extraction)

Для извлечения признаков из звуковой дорожки необходимо разбить ее на мелкие единицы размером около 10 мс - *фреймы*. Причём фреймы должны идти не строго друг за другом, а «внахлёст». То есть конец одного фрейма должен пересекаться с началом другого.

Фреймы являются более подходящей единицей анализа данных, чем конкретные значения сигнала, так как анализировать волны намного удобней на некотором промежутке, чем в конкретных точках. Расположение же фреймов «внахлёст» позволяет сгладить результаты анализа фреймов, превращая идею фреймов в некоторое «окно», движущееся вдоль исходной функции (значений сигна-

ла).

Разбив дорожку на фреймы, необходимо преобразовать ее в вектор признаков.

Для такой задачи существует множество способов, но наиболее зарекомендовавший себя среди других - это Мел-частотные кепстральные коэффициенты (Mel-frequency cepstral coefficients).

1.4 MFCC

Mel-frequency cepstral coefficients — это представление энергии спектра сигнала, где mel - единица высоты звука, основанная на восприятии этого звука органами слуха человека.

Плюсы его использования:

- Используется спектр сигнала (то есть разложение по базису ортогональных [ко]синусоидальных функций), что позволяет учитывать волновую “природу” сигнала при дальнейшем анализе;
- Спектр проецируется на специальную mel-шкалу, позволяя выделить наиболее значимые для восприятия человеком частоты;
- Количество вычисляемых коэффициентов может быть ограничено любым значением, что позволяет «сжать» фрейм и, как следствие, количество обрабатываемой информации;

Рассмотрим процесс вычисления MFCC коэффициентов для некоторого фрейма.

Представим наш фрейм в виде вектора $x[k], 0 <= k < N$, где N - размер фрейма.

1.4.1 Разложение в ряд Фурье

Рассчитываем спектр сигнала с помощью дискретного преобразования Фурье.

$$X[k] = \sum_{n=0}^{N-1} x[n] * e^{-2\pi i k \frac{n}{N}}, 0 <= k < N$$

К полученным значениям применяется оконная функция Хэмминга, чтобы «сгладить» значения на границах фреймов.

$$H[k] = 0.54 - 0.46cos(2\pi k/(N-1))$$

В результате будет вектор следующего вида:

$$X[k] = X[k]H[k], 0 <= k < N$$

В результате проведенного преобразования по оси X мы имеем частоту (hz) сигнала, а по оси Y — магнитуду (как способ уйти от комплексных значений):

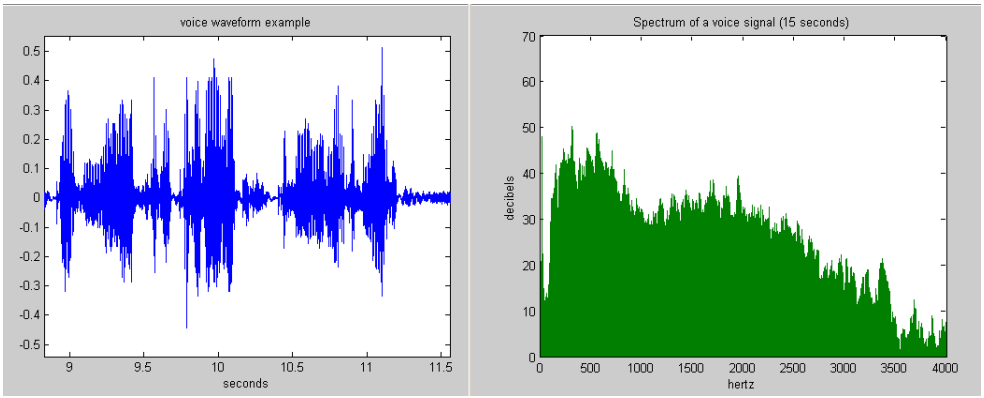

Рисунок 3

1.4.2 Расчет mel-фильтров

Для того, чтобы рассчитать mel-фильтры, необходимо определить, что такое mel.

Mel — это «психофизическая единица высоты звука», основанная на субъективном восприятии среднестатистическими людьми. Зависит в первую очередь от частоты звука (а так же от громкости и тембра). Другими словами, эта величина, показывающая, на сколько звук определённой частоты «значим» для нас.

Преобразовать частоту в мел можно по следующей формуле

$$M = 1127log(1 + \frac{F}{700})$$

Обратное преобразование выглядит следующим образом

$$F = 700(e^{\frac{M}{1127}-1})$$

Ниже приведен график зависимости mel от частоты

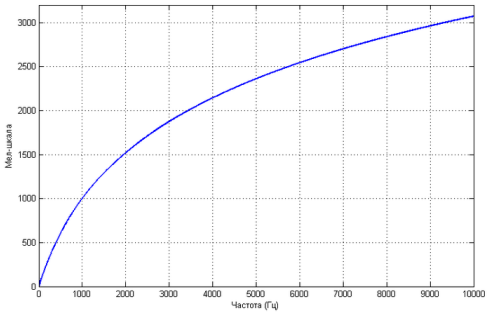

Рисунок 4

1.5 Применение фильтров, логарифмирование энергии спектра

Применение фильтра заключается в попарном перемножении его значений со значениями спектра. Результатом этой операции является mel-коэффициент.

$$S[m] = log(\sum_{k=0}^{N-1} |X[k]|^2 H_m[k]), 0 <= m < M$$

Необходимо применить mel-фильтры не к значениям спектра, а к его энергии. После чего прологарифмировать полученные результаты. Считается, что таким образом понижается чувствительность коэффициентов к шумам.

1.6 Косинусное преобразование

Дискретное косинусное преобразование (DCT) используется для того, чтобы получить кепстральные коэффициенты.

Делается это, чтобы пронормировать полученные результаты, повысив значимость первых коэффициентов и уменьшив значимость последних.

Ниже используется DCTII без домножений на scale factor ($\sqrt{\frac{2}{N}}$).

$$C[l] = \sum_{m=0}^{M-1} S[m] cos(\pi l(m + \frac{1}{2})/M), 0 <= l < M$$

На выходе имеем для каждого фрейма набор из M mfcc-коэффициент
которые могут быть использованы для дальнейшего анализа.

1.7 Акустическая модель

В акустической модели самыми распространенными подходами являются:

- скрытые марковские модели (CMM);
- глубокие нейронные сети (DNN).

2 Конструкторский раздел

Выводы по конструкторскому разделу

3 Технологический раздел

Выводы по технологическому разделу

Заключение

Список литературы

[1] Распознавание речи от Яндекса. Под капотом у Yandex.SpeechKit [Электронный ресурс]. – Режим доступа: https://m.habr.com/ru/company/yandex/blog/198556/, свободный – (20.12.2020)

[2] Понижаем барьеры на вход в распознавание речи [Электронный ресурс]. – Режим доступа: https://m.habr.com/ru/post/494006/, свободный – (20.12.2020)

[3] A Fast, Extensible Toolkit for Sequence Modeling [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1904.01038.pdf, свободный – (20.12.2020)

[4] Мел-кепстральные коэффициенты (MFCC) и распознавание речи [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/140828/, свободный – (20.12.2020)

[5] Распознавание речи для чайников [Электронный ресурс]. – Режим доступа: https://habr.com/ru/post/226143/, свободный – (20.12.2020)