МЕТОД МЕЛ-ЧАСТОТНЫХ КЕПСТРАЛЬНЫХ КОЭФФИЦИЕНТОВ В ЗАДАЧЕ РАСПОЗНАВАНИЯ РЕЧИ

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Аксёнов О. Д.

Логин В.М. – ст. преподаватель каф. ПИКС

В данной работе рассматривается один из самых популярных методов представления речевого сигнала в виде вектора признаков – мел-частотные кепстральные коэффициенты. Метод используется для представления слов в виде коэффициентов в задаче распознавания речи, в частности, распознавания команд.

Распознавание речи – это процесс автоматического распознавания слова, произнесенным определенным человеком на основе индивидуальной информации, используя речевые волны. Этот метод позволяет использовать речь для проверки учетных данных, которые обеспечивают контролируемый доступ к службам или базе данных с помощью удаленного доступа [1].

Обработка сигналов и извлечение признаков является важным шагом в любой системе распознавания речи. Речевой сигнал должен быть преобразован в определенный тип параметрического представления для дальнейшего анализа и обработки. Существует достаточно большое количество методов для того, чтобы с помощью вектора признаков представить речевой сигнал. Например, Linear Prediction Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC) и другие. Среди них MFCC – метод мел-частотных кепстральных коэффициентов, является самым известным и популярным. Метод основан на изменении человеческого голоса с критической пропускной способностью с использованием частотных треугольных фильтров. Они расположены с интервалами линейно низких частот и с логарифмическими высокими частотами для получения фонетически важных характеристик речи. Размеры окон рассчитываются при помощи мел-шкалы: первое окно очень узкое и показывает, сколько энергии содержится около нуля герц, по мере возрастания частоты размер окна становится шире, так как с ростом частоты звуки менее различимы человеческим ухом.

MFCC обычно рассчитываются с использованием набора фильтров треугольной формы с фильтром центральной частоты, который расположен с линейными частотными интервалами менее 1000 Гц и логарифмически выше 1000 Гц. Пропускная способность каждого фильтра определяется центральными частотами двух соседних фильтров и зависит от частотного диапазона наборов фильтров и количества фильтров. Для человеческой слуховой системы фильтры имеют собственную полосу пропускания, которая связана с центральной частотой фильтра [2].

Психофизические исследования показали, что восприятие человеком содержания частоты звуков для речевых сигналов не зависит от линейных шкал. Для применения мел-фильтров представления сигнала переносится от частоты Гц к высоте мел звука, используется формула, которая описывает зависимость:

$$H(f) = 1127 \cdot \ln\left(1 + \frac{1}{f}\right)$$

где f – частоты по обычной (линейной) шкале;

$H(f)$ – частоты по мел-шкале.

Энергия сигнала, которая попадает в каждое из окон анализа, получается перемножением векторов энергетического спектра сигнала и оконной функции:

$$x_m = \sum_{k=0}^{N-1} |X_k|^2 H_m(f_k),$$

где x_m – энергетический коэффициент от -ого фильтра;

$m = 1, \dots, M$ – количество фильтров;

X_k – амплитудные коэффициенты спектра сегмента;

$H_m(f)$ – функция -ого фильтра.

В результате вычисления получается набор коэффициентов x_m, содержащих спектральную информацию речевых сегментов. После вычисления энергии, выполняется логарифмирование коэффициентов. Человек может воспринимать громкость в нелинейной шкале – для удвоения воспринимаемой громкости необходимо увеличить энергию в 8 раз. За счет логарифмирования достигается эффективное сжатие пространства признаков. Но логарифм малых значений стремится

к минусу бесконечности [2]. Чтобы обойти этот эффект, можно применить метод маскировки, добавляя к значениям x_m некоторую константу:

$$x'_m = \log(x_m + c), m = 1, \dots, M.$$

Поскольку коэффициенты спектра mel являются действительными числами, мы можем преобразовать их во временной интервал, используя дискретное косинусное преобразование:

$$c_n = \sum_{m=1}^{M} x_m \cos\left(n(m - 0{,}5)\frac{\pi}{M}\right), \quad n = 1, \dots, N.$$

Также необходимо понимать, что был исключен первый компонент c_0 из ДКП, поскольку он представляет собой среднее значение входного сигнала, который несет меньше информации.

Несмотря на то, что данный алгоритм имеет малую вычислительную сложность, прост в реализации, он имеет существенный недостаток. Эксперименты показывают, что этот подход не смог распознать одно и то же слово, которое произносилось по-разному. Поэтому «сырое» применение MFCC использовать в распознавании речевых команд не рекомендуется. Для эффективного применения необходима большая выборка дикторов и вариаций произношения конкретного слова, чтобы потом создать архитектуру нейронной сети для обучения.

Пример слова «yes» представлен на рисунке 2.

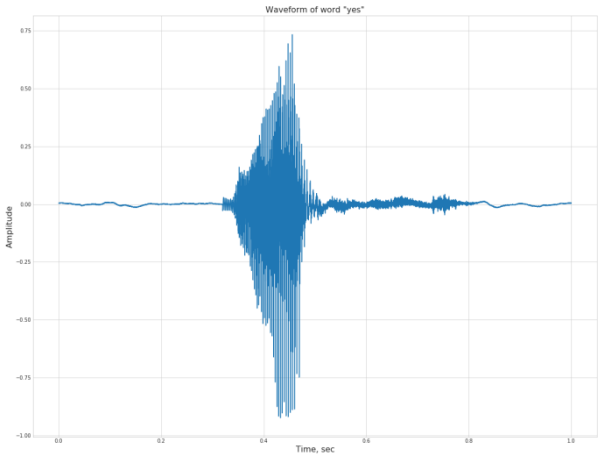

Рисунок 2а – Форма слова «yes»

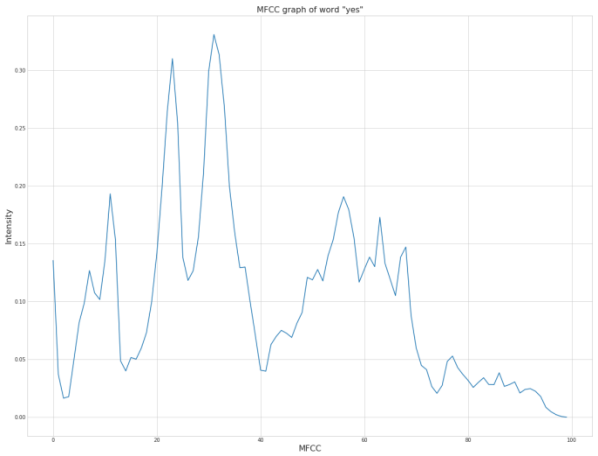

Рисунок 2б – График MFCC коэффициентов слова «yes»

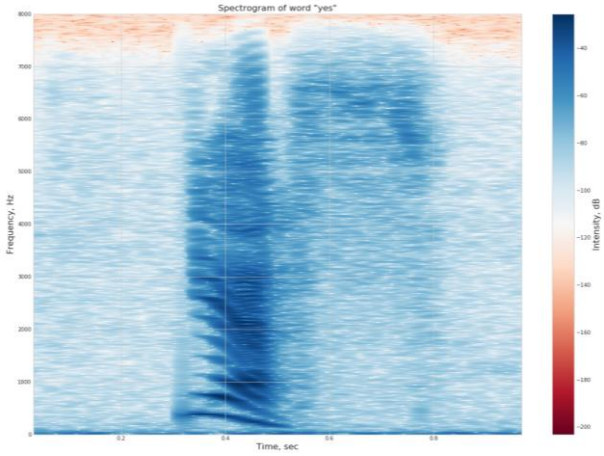

Рисунок 2в – Спектрограмма слова «yes»

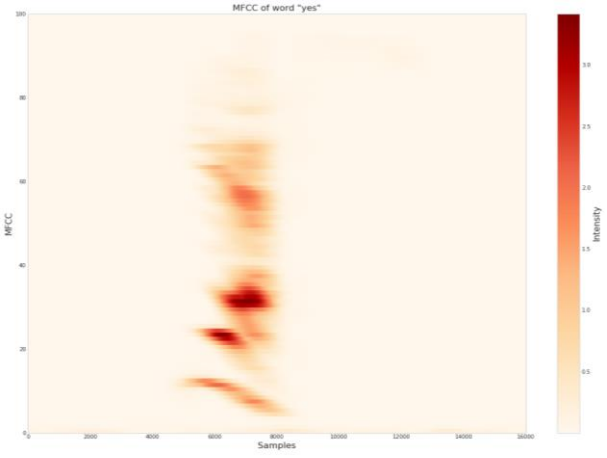

Рисунок 2г – Мел-грамма слова «yes»

Список использованных источников:

1. Мел-кепстральные коэффициенты (MFCC) и распознавание речи [Электронный ресурс] - Режим доступа: https://habr.com/post/140828/. Дата доступа:13.04.2019
2. Рабинер Л., Шафер Р. Цифровая обработка речевых сигналов. — М.: Радио и связь, 1981. — 489 с.
3. Librosa [Электронный ресурс] - Режим доступа: https://librosa.github.io/ Дата доступа: 12.04.2019.