**Введение**

Распознавание речи человека является актуальной и практически значимой задачей в области обработки сигналов. К современным устройствам предъявляются требования к низкому энергопотреблению, высокой мобильности и производительности. Для удовлетворения этих противоречивых требований требуется разработать специализированное устройство, которое будет решать самую трудоемкую задачу – преобразование оцифрованного акустического сигнала в цифровую последовательность дискретных признаков речевого сигнала. Таким этапом распознавания речи является выделение фонем.

Речь – это последовательность смыслоразличимых фонем. Фонема – это минимальная звуковая единица [языка](https://ru.wikipedia.org/wiki/%D0%AF%D0%B7%D1%8B%D0%BA). Фонема не имеет самостоятельного [лексического](https://ru.wikipedia.org/wiki/%D0%9B%D0%B5%D0%BA%D1%81%D0%B8%D0%BA%D0%B0) или [грамматического](https://ru.wikipedia.org/wiki/%D0%93%D1%80%D0%B0%D0%BC%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D0%BA%D0%B0%D1%82%D0%B5%D0%B3%D0%BE%D1%80%D0%B8%D1%8F) значения, но служит для различения и отождествления значимых единиц языка ([морфем](https://ru.wikipedia.org/wiki/%D0%9C%D0%BE%D1%80%D1%84%D0%B5%D0%BC%D0%B0) и слов).

Микрофон преобразует речь человека из акустического сигнала в электрический сигнал. Электрический сигнал поступает на аналого-цифровой преобразователь, отсчеты которого являются входным сигналом для устройства выделения фонем.

Для устойчивого, помехозащищенного выделения фонем, отсчеты оцифрованного сигнала подвергаются специальному преобразованию для уменьшения объема обрабатываемых данных. Одним из таких преобразований является вычисление MFCC коэффициентов, которые позволяют сократить объем обрабатываемых данных и увеличить надежность верного распознавания.

**Формирование имитационных фонем**

Существуют различные математические модели для описания звуков и речевого аппарата человека [7,8]. Для простоты изложения метода выделения фонем из непрерывной речи человека ограничимся моделированием имитационных фонемы.

Сначала формируется дискретный сигнал, состоящий из суммы синусоид

.

В таблице 1 каждому коду фонемы соответствует набор частот.

Таблица 1. Соответствие кода фонемы и наличия частот в сигнале

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Код фонемы |  |  |  |  |  |
| 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 0 | 1 | 0 |
| 3 | 0 | 0 | 0 | 1 | 1 |
| 4 | 0 | 0 | 1 | 0 | 0 |
| 5 | 0 | 0 | 1 | 0 | 1 |
| 6 | 0 | 0 | 1 | 1 | 0 |
| 7 | 0 | 0 | 1 | 1 | 1 |
| 8 | 0 | 1 | 0 | 0 | 0 |
| 9 | 0 | 1 | 0 | 0 | 1 |
| 10 | 0 | 1 | 0 | 1 | 0 |
| 11 | 0 | 1 | 0 | 1 | 1 |
| 12 | 0 | 1 | 1 | 0 | 0 |
| 13 | 0 | 1 | 1 | 0 | 1 |
| 14 | 0 | 1 | 1 | 1 | 0 |
| 15 | 0 | 1 | 1 | 1 | 1 |
| 16 | 1 | 0 | 0 | 0 | 0 |
| 17 | 1 | 0 | 0 | 0 | 1 |
| 18 | 1 | 0 | 0 | 1 | 0 |
| 19 | 1 | 0 | 0 | 1 | 1 |
| 20 | 1 | 0 | 1 | 0 | 0 |
| 21 | 1 | 0 | 1 | 0 | 1 |
| 22 | 1 | 0 | 1 | 1 | 0 |
| 23 | 1 | 0 | 1 | 1 | 1 |
| 24 | 1 | 1 | 0 | 0 | 0 |
| 25 | 1 | 1 | 0 | 0 | 1 |
| 26 | 1 | 1 | 0 | 1 | 0 |
| 27 | 1 | 1 | 0 | 1 | 1 |
| 28 | 1 | 1 | 1 | 0 | 0 |
| 29 | 1 | 1 | 1 | 0 | 1 |
| 30 | 1 | 1 | 1 | 1 | 0 |
| 31 | 1 | 1 | 1 | 1 | 1 |

Речь человека лежит в диапазоне частот от 100 Гц до 4000 Гц, поэтому, на основании теоремы Котельникова[1], достаточно взять частоту дискретизации: 

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Обозначение | F5 | F4 | F3 | F2 | F1 |
| Частота, Гц | 2093 | 1396 | 659 | 293 | 196 |

Затем сигнал будет проходить через нелинейный элемент вида

,

где коэффициенты  задаются случайно в диапазоне значений от 1 до 50, далее через фильтр имитирующий АЧХ тракта человека (фильтр тракта).

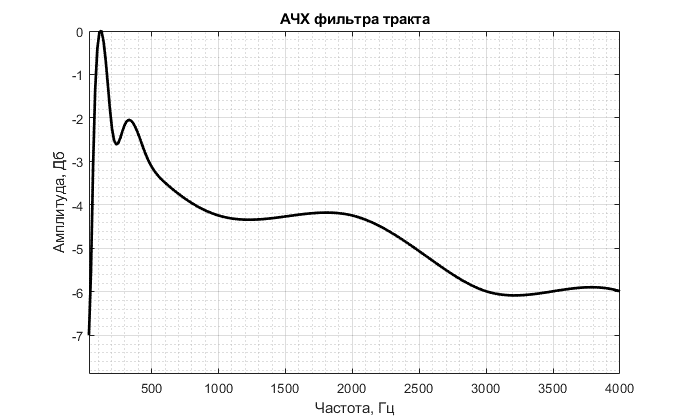


Рисунок 1. АЧХ фильтра тракта



Рисунок 2. Структурная схема тракта человека

Возьмём, в качестве примера, две фонемы (код 9, код 25).

Применим к сигналам дискретное преобразование Фурье[1]



Отобразим амплитудные спектры сигналов в области положительных частот на рисунке 3.

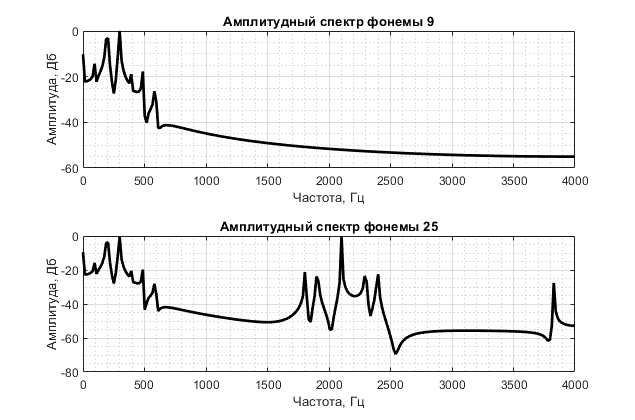


Рисунок 3. Амплитудный спектр

Код фонемы 9 состоит из 2 частот, а код фонемы 25 из 3 частот, на рисунке 2 отчётливо видно основные частоты сигнала.

Что бы получить MFCC коэффициенты нужно понять, что такое MFCC и зачем используется?

# **MFCC**

MFCC (Mel-frequency cepstral coefficients)[2][3][4][5] -это своеобразное представление энергии спектра сигнала. Плюсы его использования заключаются в следующем:

* используется спектр сигнала (то есть разложение по базису ортогональных синусоидальных функций), что позволяет учесть волновую природу сигнала;
* спектр проецируется на специальную [mel-шкалу](http://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%BB_(%D0%B2%D1%8B%D1%81%D0%BE%D1%82%D0%B0_%D0%B7%D0%B2%D1%83%D0%BA%D0%B0), позволяя выделить наиболее значимые для восприятия человеком частоты;
* количество вычисляемых коэффициентов может быть ограничено небольшим значением (например 15), что позволяет уменьшить количество обрабатываемой информации.

MFCC состоит из следующих этапов:

1. Дискретное преобразование Фурье
2. Применение гребёнки треугольных фильтров
3. Вычисление энергии каждого окна
4. Логарифмирование энергии
5. Применение дискретно-косинусного преобразования

Первый этап уже был выполнен ранее. Чтобы увеличить значимость основных частот в спектре, возьмём мощность спектра.



Расчёт mel-фильтров.

Что бы промоделировать логарифмическое ухудшение слуха человека, необходимо запустить в mel-шкале гребёнку треугольных фильтров.

С ростом частоты, разрешающая способность уха ухудшается. То есть человек вполне может отличить звук 300 Гц и 400 Гц, но 3300 Гц и 3400 Гц различить друг от друга, ему уже сложнее.

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 mel можно по следующей формуле



Обратное преобразование выглядит следующим образом



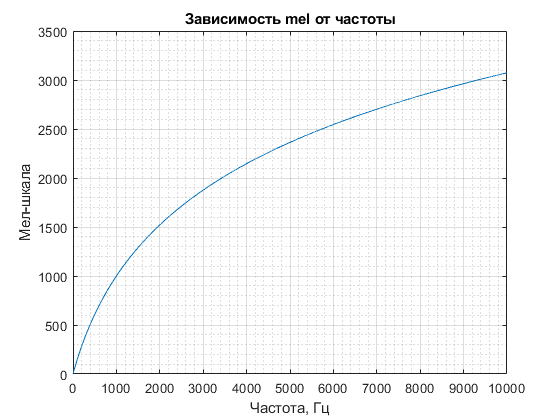


Рисунок 4. Зависимость mel от частоты

У каждого фильтра с ростом частоты основание расширяется, соответственно больший «кусок» сигнала в частотной области он захватывает.



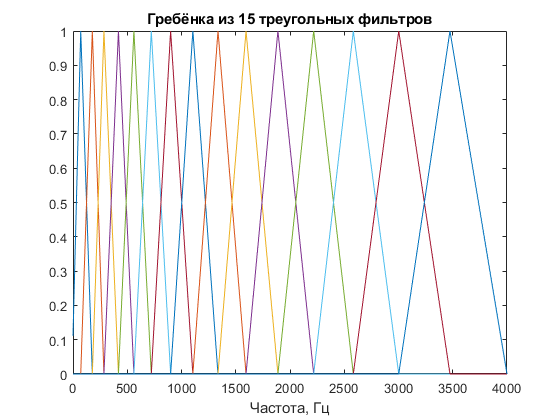


Рисунок 5. Гребёнка треугольных фильтров

Если просуммировать количество энергии приходящуюся на каждый mel-фильтр, то получится сглаженный mel-спектр, избавленный от высокочастотных шумов.



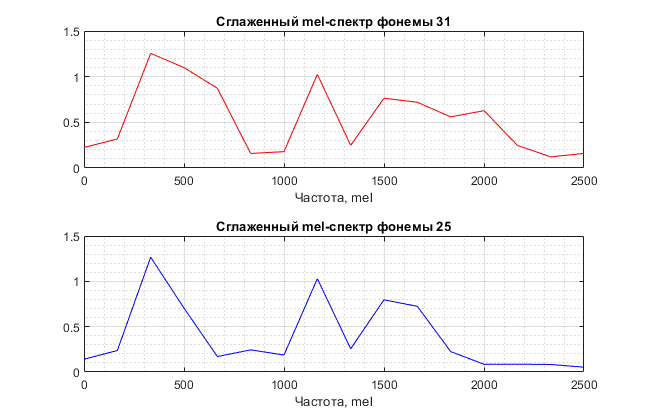


Рисунок 6. Сглаженный mel-спектр

Получились мел-частотные спектральные коэффициенты. На этом этапе некоторые системы распознавания речи останавливаются.

Чтобы получить mel-кепстральные коэффициенты, необходимо ещё проделать несколько этапов.

Для начала нужно пояснить что же значит «кепстральные» коэффициенты.

Гомоморфное преобразование - это преобразование, преобразующее свертку в сумму.

**Кепстр** -это гомоморфное преобразование, которое позволяет отделить источник от фильтра.

Можно найти значения N такие, что кепстр фильтра , и кепстр фильтра .

Благодаря этому, можно приблизительно восстановить  из  с помощью гомоморфной фильтрации.

На рисунке 6 показано, как восстановить h[n] с гомоморфной фильтрацией:  


где D-кепстральный оператор.

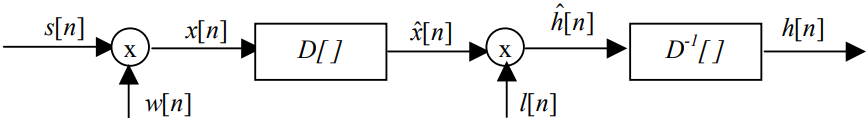


Рисунок 7. Гомоморфная фильтрация для восстановления отклика фильтра от периодического сигнала

После получения мел-частотных спектральных коэффициентов их необходимо прологарифмировать. Это также мотивируется человеческим слухом: люди не слышат громкость в линейном масштабе. Это означает, что большие колебания энергии могут звучать не так уж и по-другому, если звук с самого начала громкий. Эта операция сжатия делает наши функции более близкими к тому, что на самом деле слышат люди.





Рисунок 8. Логарифм мощности приходящийся на каждый mel-фильтр

Кепстр частоты mel - это дискретное косинусное преобразование от P[m]:



В итоге получаются MFCC коэффициенты, изображённые на рисунке 9.

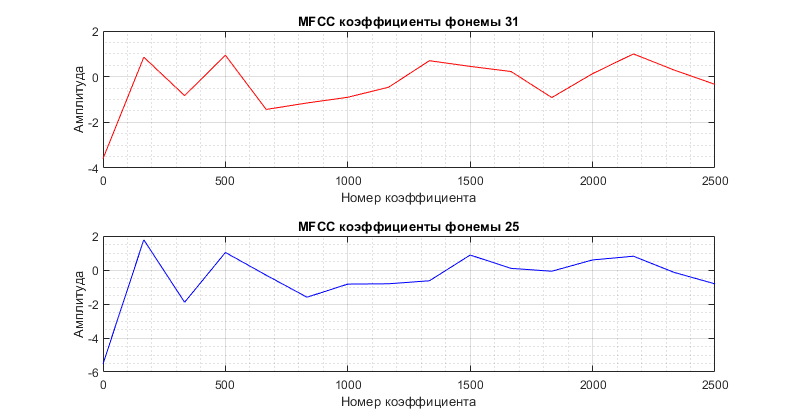


Рисунок 9. MFCC коэффициенты

У каждой фонемы свой набор коэффициентов, по этим коэффициентам происходит выделение из речевого сигнала.

Определяются MFCC коэффициенты для всех 32 двух фонем. Далее оцениваются параметры коэффициентов через GMM[6] модель и кластеризуются. В итоге получается 32 кластера. На этом минимально жизненно необходимая система для распознавания фонем готова.

# **Тестирование системы**

Подадим на такую систему следующий сигнал состоящий из 4 последовательных фонем.



Рисунок 10. Временная диаграмма выхода ГСС

Необходимо сделать следующие операции:

1. Разделить сигнал на кадры.

Ширину кадра выбирают 25.6 мс. Эта ширина достаточна для захвата необходимой информации и внутри такого кадра сигнал остаётся стационарным.

1. Найти коэффициенты MFCC для каждого кадра
2. Отправить эти коэффициенты в обученную GMM модель

На выходе GMM модель даст для каждого кадра номер кластера, связанный с кодом фонемы следующим образом: 



Рисунок 11. Структурная схема определения фонемы

Если взять отношение количества верно определённых кадров к полному количеству кадров, то получим вероятность верного определения последовательности фонем.



# **Заключение**

Таким образом если в качестве критерия сравнения параметров фонемы выбрать MFCC коэффициенты, то получается небольшой набор значений, который при распознавании успешно заменяет тысячи отсчетов речевого сигнала.

# **Список использованных источников**

1. В.Г. Карташев, Е.В. Шалимова, КОНСПЕКТ ЛЕКЦИЙ по курсам «Радиотехнические цепи и сигналы» «Цепи и сигналы в медицинской электронике», Кафедра Основ радиотехники, Москва НИУ «МЭИ», 2017
2. Feature Normalization for Robust Speech Recognition D. S. Pavan Kumar, arXiv:1507.04019v1 [cs.CL] 14 Jul 2015
3. Mel Frequency Cepstral Coefficients Based Similar Albanian Phonemes Recognition Conference Paper  in  Lecture Notes in Computer Science · July 2016
4. Pronunciation recognition of English phonemes /@/, /æ/, /A:/ and /2/ using Formants and Mel Frequency Cepstral Coefficients Keith Y. Patarroyo and Vladimir Vargas-Calder´on, arXiv:1702.07071v1 [cs.CL] 23 Feb 2017
5. https://medium.com/@jonathan\_hui/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9
6. <https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95>
7. II. А. Любимов, Е. В. Захаров, МАТЕМАТИЧЕСКАЯ МОДЕЛЬ АКУСТИЧЕСКОГО РЕЧЕОБРАЗОВАНИЯ С ПОДВИЖНЫМИ СТЕНКАМИ РЕЧЕВОГО ТРАКТА, 2016 г.
8. C.Julian Chen, Elements of human voice, Columbia University, USA, 2016