ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ - MFCC ДЛЯ ОБРАБОТКИ РЕЧИ

До А.Т., магистрант гр.467311

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

Зельманский О.Б. – канд. тех. наук, доцент

Аннотация. Извлечение признаков играет ключевую роль в системах обработки речи, и коэффициенты мел-частотного цепстра (MFCC) становятся одним из самых эффективных представлений, которые близко приближают человеческое слуховое восприятие. MFCC эффективно имитируют нелинейное восприятие человеческой слуховой системы, преобразуя звуковые сигналы в компактный набор дискриминативных признаков. Процесс извлечения включает несколько ключевых этапов: предварительная усиление для повышения высокочастотных компонент, фреймирование и окно для захвата краткосрочных признаков, преобразование Фурье для получения спектральной информации, обработка фильтров Мел для приближения чувствительности человеческого слуха, логарифмическое сжатие для акцентирования перцептивно важных компонент и, наконец, дискретное косинусное преобразование (DCT) для декорреляции признаков и получения цепстральных коэффициентов.

Ключевые слова. Речевые сигналы, распознавание речи, MFCC, извлечение признаков, преобразование Фурье, DCT.

Чтобы преобразовать человеческую речь в информацию, которую компьютер может понять и обработать, система распознавания речи должна пройти через процесс цифровой обработки сигналов, состоящий из нескольких важных этапов. Каждый этап играет ключевую роль в обеспечении точности и эффективности системы. Данная статья сосредоточена на методе извлечения признаков MFCC (мел-частотные кепстральные коэффициенты) – популярной технике в обработке речи, особенно в задачах распознавания речи (ASR) и обработки звуковых сигналов. Процесс начинается с деления входного звукового сигнала на короткие фреймы по 25 мс с перекрытием в 10 мс для обеспечения непрерывности. Затем каждый фрейм преобразуется в вектор размерности 39 с помощью обработки, включающей преобразование Фурье, применение фильтра шкалы Мел и вычисление кепстральных коэффициентов. Это создает признаки с низким уровнем шума, высокой независимостью и эффективной акустической репрезентацией, отвечающими требованиям алгоритмов машинного обучения, при этом обеспечивая разумные вычислительные затраты. Весь процесс, начиная от сырого звукового сигнала до окончательных признаков MFCC, подробно иллюстрирован на рисунке ниже.

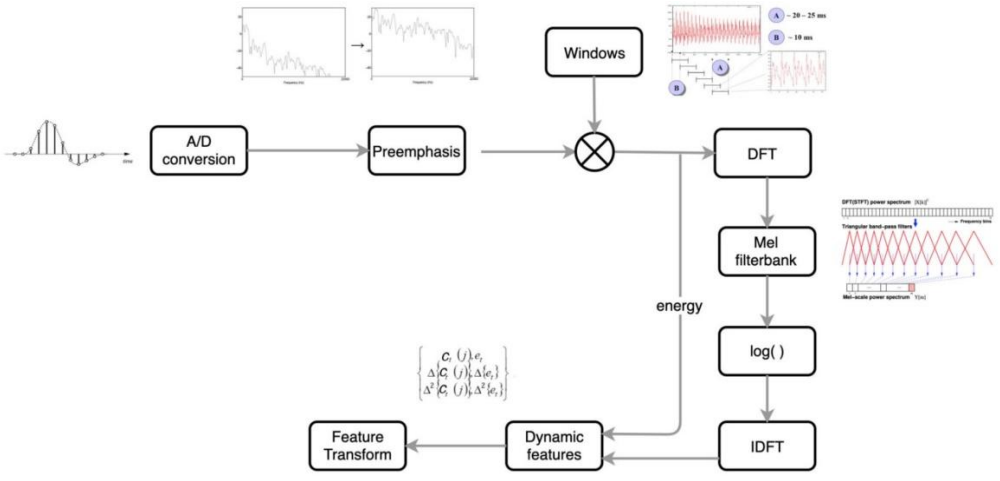

Рисунок 1 – Процесс извлечения признаков MFCC из речевого сигнала

Поскольку звук является непрерывным (аналоговым) сигналом, в то время как компьютеры обрабатывают данные в дискретной (цифровой) форме, процесс преобразования требует выборки

сигнала в равномерно распределенные моменты времени с определенной частотой, называемой частотой дискретизации (sample rate). Например, при sample_rate = 8000 Hz система будет получать 8000 значений амплитуды звука каждую секунду, создавая точное цифровое представление оригинального сигнала.

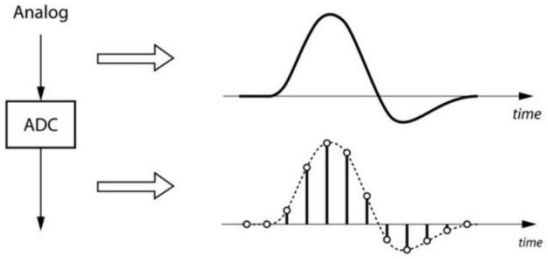

Рисунок 2 – Процесс преобразования аналогового сигнала в цифровой (АЦП)

Человеческое ухо может воспринимать звук в диапазоне от 20 Гц до 20 000 Гц. Согласно теореме дискретизации Найквиста-Шеннона: для сигнала с компонентами частоты ≤ f_m, чтобы гарантировать, что дискретизация не приведет к потере информации (aliasing), частота дискретизации f_s должна удовлетворять условию $f_s \geq 2\, f_m$.

Таким образом, чтобы обеспечить отсутствие потерь информации при дискретизации, частота дискретизации должна составлять $f_s = 44100$ Гц. Однако во многих случаях достаточно использовать $f_s = 8000$ Гц или $f_s = 16000$ Гц.

Из-за особенностей строения гортани и органов артикуляции наша речь имеет следующие характеристики: звуки низкой частоты обладают высокой энергией, тогда как звуки высоких частот имеют довольно низкий уровень энергии. В то же время высокие частоты содержат много информации о фонемах. Поэтому нам нужен этап предварительного усиления (pre-emphasis), чтобы повысить уровень этих высокочастотных сигналов.

$$x'[t_d] = x'[t_d] - \alpha x[t_d\text{-}1] \qquad 0{,}95 < \alpha < 0{,}99$$

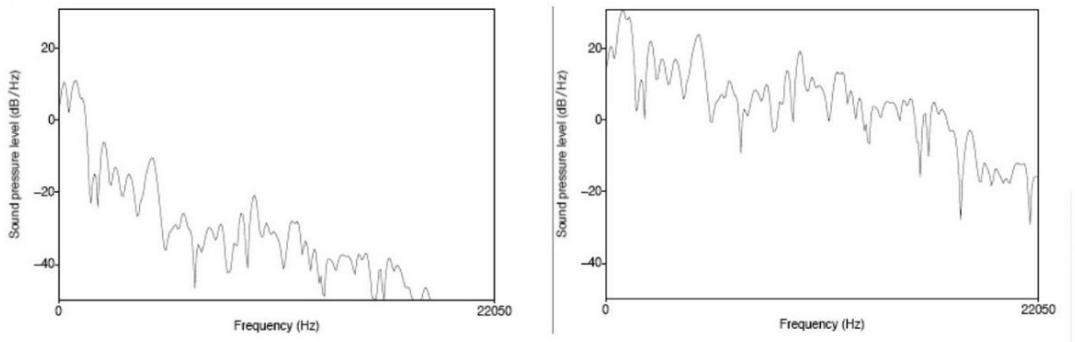

Рисунок 3 – Предварительное усиление высоких частот (Pre-emphasis) в обработке речи

Вместо того чтобы применять преобразование Фурье к длинному участку звука, мы скользим по сигналу с помощью окна, чтобы извлечь фреймы, а затем применяем ДПФ (дискретное преобразование Фурье) к каждому из этих фреймов. Средняя скорость речи человека составляет около 3-4 слов в секунду, каждое слово состоит из 3-4 звуков, а каждый звук делится на 3-4 части. Таким образом, 1 секунда звука делится на 36-40 частей. Мы выбираем ширину каждого фрейма около 20-25 мс, что достаточно для охвата одной части звука. Фреймы перекрываются друг с другом на 10 мс, чтобы можно было захватить изменения контекста.

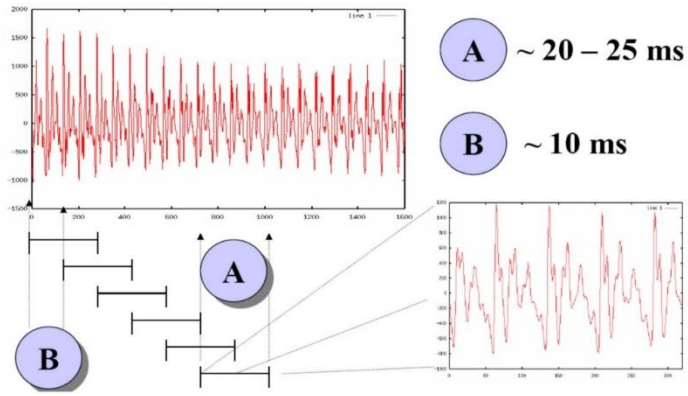

Рисунок 4 – Анализ речи методом скользящего окна (Framing) с ДПФ

Однако обрезка фрейма приводит к резкому снижению значений на обоих краях фрейма (до нуля), что приводит к тому, что при применении ДПФ в частотной области возникает много шумов на высоких частотах. Чтобы исправить это, необходимо сгладить фрейм, умножив его на несколько типов окон. Существуют несколько распространенных окон, таких как окно Хэмминга, окно Ханнинга и другие, которые помогают постепенно снижать значения на краях фрейма.

На каждом фрейме мы применяем ДПФ (дискретное преобразование Фурье) по формуле:

$$X[k] = \sum_{n=0}^{N-1} \mathrm{x[n]} \exp\left(-j\frac{2\pi}{N}kn\right)$$

На каждом фрейме мы получаем список значений амплитуды (magnitude) для каждой частоты от 0 до N. Применяя это ко всем фреймам, мы получаем спектрограмму, как показано ниже. Ось x – это ось времени (соответствующая порядку фреймов), ось y представляет диапазон частот от 0 до 10000 Гц, а значение амплитуды на каждой частоте отображается цветом. Наблюдая за этой спектрограммой, мы замечаем, что на низких частотах обычно высокая амплитуда, а на высоких частотах – низкая амплитуда.

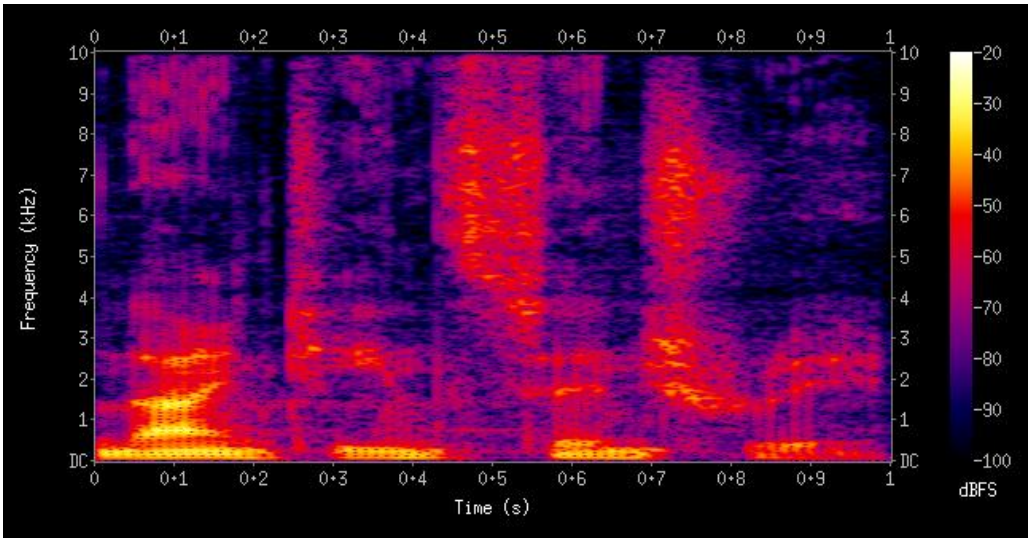

Рисунок 5 – Распределение звуковой энергии по времени и частоте

Человеческое слуховое восприятие имеет особенности нелинейной обработки звука, которые полностью отличаются от обычных измерительных устройств. В частности, человеческое ухо имеет высокую чувствительность в низкочастотном диапазоне (около 20 Гц - 2 кГц), но значительно снижает

чувствительность в высокочастотном диапазоне (выше 5 кГц). Чтобы точно смоделировать, как человек воспринимает звук, нам необходимо создать механизм частотного отображения (frequency mapping), который имитирует эту нелинейную характеристику слуховой системы.

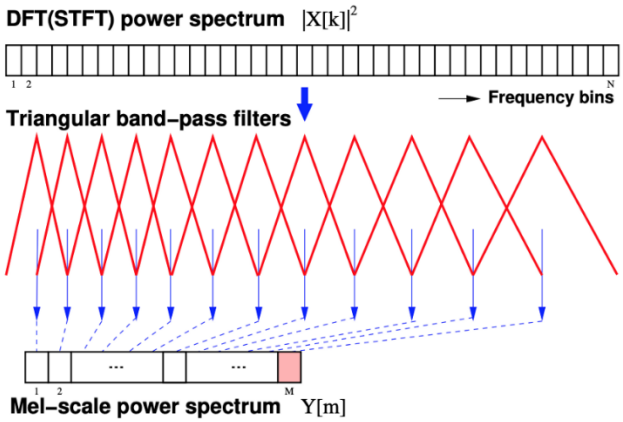

Рисунок 6 – Процесс извлечения мел-спектральных признаков из кратковременного преобразования Фурье (КПФ)

Сначала мы возводим в квадрат значения в спектрограмме, получая спектр мощности ДПФ (DFT power spectrum). Затем мы применяем набор полосовых фильтров Mel-scale на каждом диапазоне частот (каждый фильтр применяется к определенному диапазону частот). Значение на выходе каждого фильтра – это энергия диапазона частот, который этот фильтр покрывает. В результате мы получаем спектр мощности в масштабе Мел (Mel-scale power spectrum). Кроме того, фильтры, используемые для низких частот, обычно уже, чем фильтры для высоких частот.

Этот процесс также можно описать с помощью иллюстрации ниже:

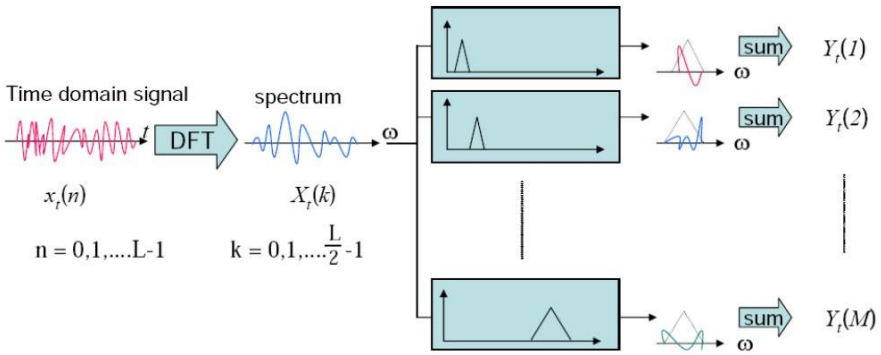

Рисунок 7 – Преобразование временного сигнала в мел-спектр с использованием полосовых фильтров

Фильтровая банка Мел возвращает спектр мощности звука, также известный как спектр энергии. На самом деле, человек менее чувствителен к изменениям энергии на высоких частотах и более чувствителен на низких частотах. Поэтому мы будем вычислять логарифм спектра мощности в масштабе Мел (Mel-scale power spectrum). Это также помогает уменьшить незначительные вариации звука для распознавания речи.

В обработке речи информация о основной частоте F0 (около 125 Гц у мужчин и 210 Гц у женщин) обычно не полезна для распознавания речи. Чтобы исключить F0 и оставить только важные форманты (F1, F2, F3 и так далее), мы используем обратное преобразование Фурье (IDFT) для перехода сигнала в область цепстрюма (Cepstrum) – техника, название которой образовано путем обращения слова "спектр" (spectrum). В области цепстрюма информация о F0 и формантах четко разделяется, что позволяет легко

удалить компонент F0, просто взяв 12 первых цепстральных коэффициентов (соответствующих формантам). Это преобразование также можно выполнить с помощью дискретного косинусного преобразования (DCT), которое создает меньше коррелирующие признаки, что идеально подходит для алгоритмов машинного обучения. В конечном итоге мы получаем 12 цепстральных признаков, сосредоточенных на фонетических характеристиках, не зависимых от высоты голоса каждого человека.

Таким образом, из каждого кадра мы извлекли 12 цепстральных признаков, которые стали первыми 12 признаками MFCC. 13-й признак – это энергия этого кадра, вычисляемая по формуле:

$$Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

В распознавании речи информация о временных изменениях играет важную роль в определении фонем, особенно на переходных позициях между фонемами (таких как начало или конец согласного). Эти изменения очень явные, и фонемы могут быть распознаны на основе этих изменений. Следующие 13 коэффициентов – это первая производная (по времени) первых 13 признаков. Она содержит информацию о изменении от кадра t до кадра t+1. Формула:

$$d(t) = \frac{c(t+1) - c(t-1)}{2}$$

Аналогично, последние 13 значений MFCC представляют собой изменение d(t) во времени – производную d(t), а также вторую производную c(t). Формула:

$$b(t) = \frac{d(t+1) - d(t-1)}{2}$$

Таким образом, из 12 цепстральных признаков и 13-го признака мощности, проведя две производные, мы получаем 39 признаков. Это и есть признаки MFCC.

Метод извлечения признаков MFCC зарекомендовал себя как эффективный в обработке речи благодаря способности точно моделировать характеристики слуха человека через фильтр шкалы Мел, одновременно предоставляя компактный набор признаков благодаря цепстральному преобразованию. Несмотря на некоторые ограничения в чувствительности к шуму и фиксированной структуре фильтра, MFCC сохраняет важную позицию благодаря своим выдающимся преимуществам в вычислительной эффективности и надежности. Будущие направления развития сосредоточены на улучшении адаптивных фильтров, комбинировании с современными методами извлечения признаков и повышении устойчивости к шуму, что продолжает подтверждать ценность MFCC как в традиционных системах, так и в современных архитектурах глубокого обучения.

Список использованных источников:

1. Li, T.LH. Genre classification and the invariance of MFCC features to Key and Tempo / T.LH. Li, A.B. Chan // International Conference on MultiMedia Modeling, Taipei, 2011.

2. Mel-frequency cepstrum. [Электронный ресурс]. – Режим доступа: https://en.wikipedia.org/wiki/Mel-frequency_cepstrum – Дата доступа: 04.03.2025.

3. Распознавание речи. [Электронный ресурс]. – Режим доступа: https://ru.wikipedia.org/wiki/Распознавание_речи – Дата доступа: 04.03.2025.

4. Мел-кепстральные коэффициенты (MFCC) и распознавание речи. [Электронный ресурс]. – Режим доступа: https://habr.com/ru/articles/140828/– Дата доступа: 04.03.2025.

SOFTWARE FOR SPEECH DETECTION IN SIGNAL

Do A. T.

Belarusian State University of Informatics and Radioelectronics[1], Minsk, Republic of Belarus

Zelmansky O.B. – Candidate of Technical Sciences, Associate Professor

Annotation. Automatic detection of speech segments in audio signals is an important task in the field of natural language processing and human-computer interaction. This article proposes software based on the extraction of MFCC features (Mel-Frequency Cepstral Coefficients) combined with deep learning for accurately distinguishing speech from other components (noise, silence, music). The methodology includes signal preprocessing (noise filtering, normalization), extraction of mel-frequency features, and training of a CNN-LSTM model capable of analyzing both frequency characteristics and temporal sequences. Testing on the LibriSpeech and TIMIT datasets demonstrated an accuracy of 93.5% with an F1 score of 95.2%, surpassing traditional methods (SVM, HMM). The developed software also demonstrates the capability to operate in real-time on embedded devices, opening up prospects for applications in virtual assistants, medical dialogue analysis, and intelligent audio monitoring systems.

Keywords. Speech signals, speech recognition, MFCC, Feature extraction, Fourier transform, DCT.