Математические модели обработки сигналов

Тема 11: MFCC

Лектор: Кривошеин А.В.

MFCC

Мел-частотные кепстральные коэффициенты (англ. MFCC, mel-frequency cepstrum coefficients) разработаны для обработки речевых сигналов.

MFCC позволяет выделить признаки, которые удобно использовать для распознавания речевых сигналов. Звуки, произносимые человеком зависят от формы голосового тракта в момент произнесения звуков. MFCC до некоторой степени позволяют представить эту форму через набор коэффициентов.

Мел-шкала (сокр. от англ. слова melody) — это психофизическая единица измерения высоты слышимого тона.

Мел-шкала основана на статистической обработке значительного объёма эмпирических данных о субъективном восприятии высоты слышимых звуков. Шкала основана на том, что человек лучше различает частоты 500 и 600 Гц, чем частоты 1000 и 1100 Гц.

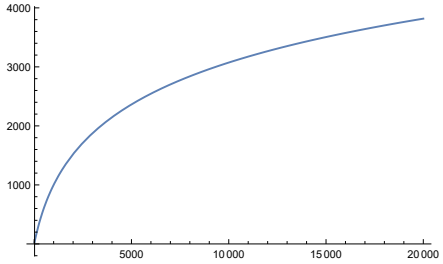
Мел-шкала

Есть различные формулы для перевода частоты в Гц в высоту звука в мелах.

Один из распространённых вариантов имеет вид

$$m = 2595 \log_{10}\left(1 + \frac{\omega}{700}\right) = 1127 \ln\left(1 + \frac{\omega}{700}\right),$$ где ω — это частота в Гц.

Out[ ]=



Кепстр

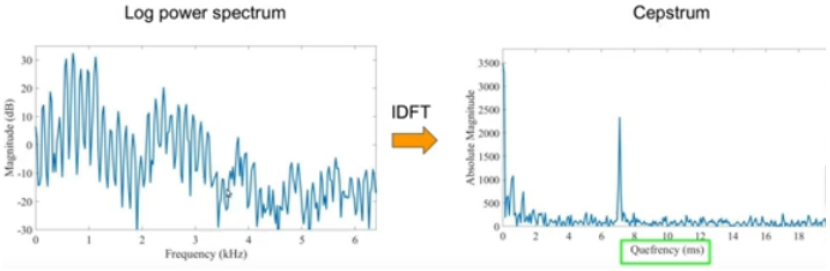
Кепстр — это спектр от спектра (англ. cepstrum).

Кепстр изначально использовался для изучения эха в сейсмических сигналах (1960-е). Позже было замечено, что кепстр позволяет получить хороший набор признаков для распознавания и идентификации речи (1970-е).

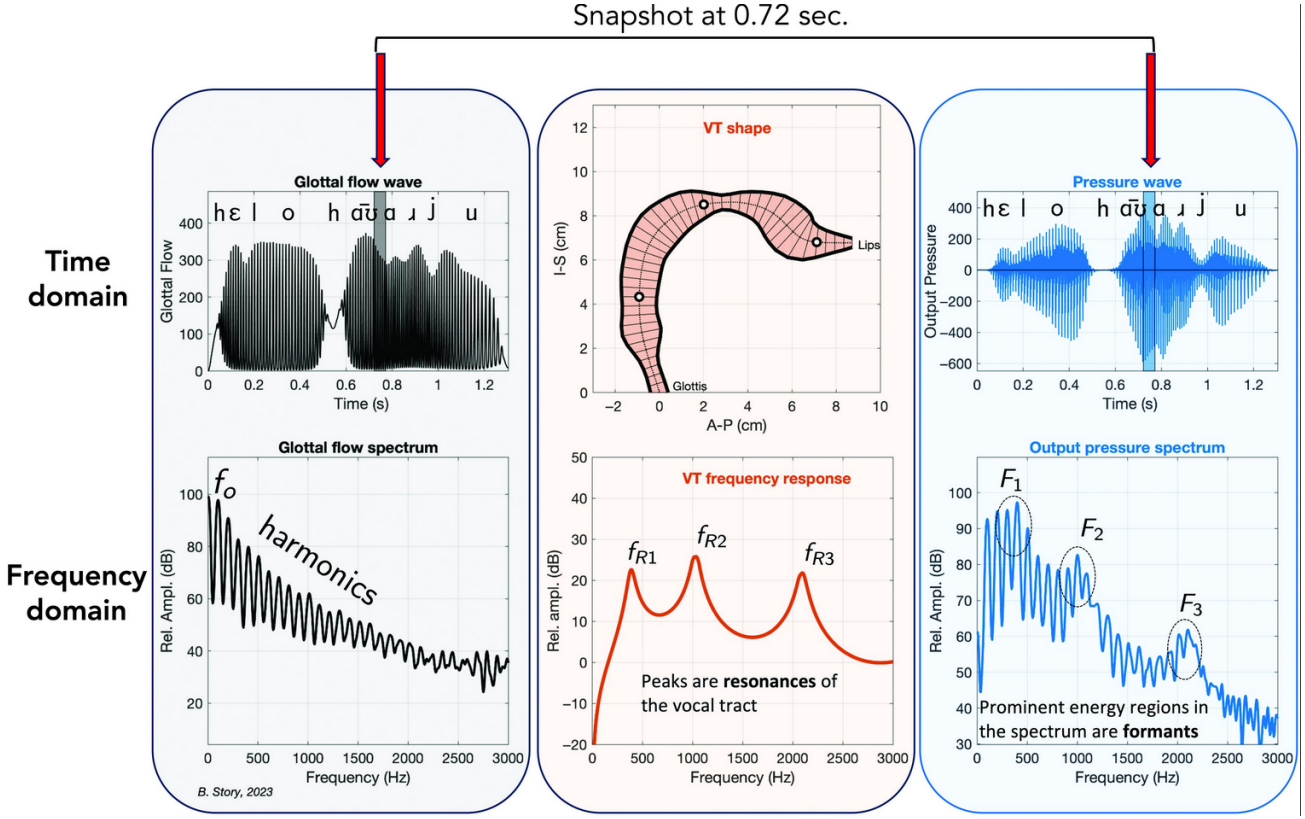
Пусть $x \in \mathbb{R}^d$ некоторый сигнал. Его кепстр имеет вид

$$\text{Cepstrum}(x) = \text{DFT}^{-1}\left\{\log\left(|\text{DFT}(x)|^2\right)\right\}$$

Для речевых сигналов $\log(|\text{DFT}(x)|^2$ выглядит так, как он содержит также частотные компоненты (то есть с изменением частоты периодически меняется значение по оси Oy) и их можно выделить через ДПФ от $\log(|\text{DFT}(x)|^2$.



Мотивация к MFCC



Мотивация к MFCC

Речевой сигнал формируется в голосовом тракте. Звуки из голосовых связок проходят через голосовой тракт. Его форма в момент прохождения звука влияет на произносимый звуки. Фактически звуки из голосовых связок подвергаются фильтрации.

Из спектра произнесённого звука можно форму голосового тракта выделить.

Лог-спектр имеет огибающую. Огибающая фактически является частотным откликом фильтра голосового тракта. Пики огибающей называют формантами — по ним можно идентифицировать звук.

Применение фильтра — это свёртка. В частотной области — это произведение спектров. После логарифмирования — это сумма лог-спектров.

$$x = h * e$$

$$X = H E$$

$$\log |X|^2 = \log |H|^2 + \log |E|^2.$$

Как разделить компоненты речи? По сути сигнал от голосовых связок не важен для распознавания речи (он необходим для решения задачи идентификации спикера). Требуется получить форманты. Подсчитав ДПФ от $\log |X|^2$ можно отсечь быстро меняющиеся частоты от медленно меняющейся огибающей.

Разница между кепстром и мел-частотным кепстром заключается в том, что в MFCC полосы частот равномерно распределены по мел-шкале, что более точно отражает работу слуховой системы человека, чем равномерные полосы частот в обычном спектре.

Шаги для вычисления MFCC

1. Сигнал разделяется на блоки (или фреймы). Размер блока от 20 до 40 мс.

$x[n] \rightarrow x_j[n], \qquad n = 0, \ ..., N-1, \qquad N$ длина блока в отсчётах.

Далее, для каждого блока:

2. Блок сглаживается окном (например, окном Хемминга) для сглаживания на границах. Затем применяется ДПФ.

$$X_j[k] = \sum_{n=0}^{N-1} x_j[n]\, w[n]\, e^{-2\pi i \frac{kn}{N}}, \quad k = 0, \ ..., N-1.$$

3. Вычисляется вектор квадратов амплитуд спектра:

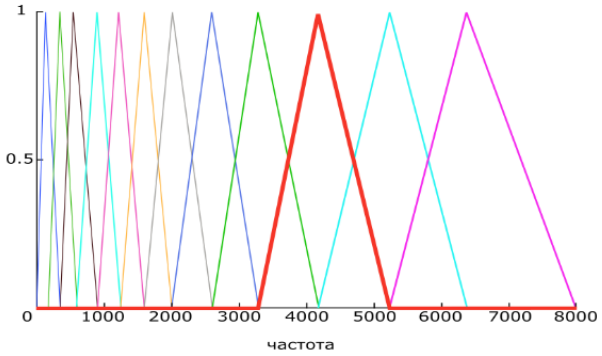
$$P_j[k] = \frac{|X_j[k]|^2}{N}. \quad k = 0, \ ..., N-1$$

Это по сути доля энергии k-го отсчёта спектра в общей энергии сигнала.

Шаги для вычисления MFCC

4. Применяем мел-фильтры. Для этого вектор P_j поточечно умножается на треугольные окна и результат суммируется.

$$H_m(k) = \begin{cases} 0, & k < f(m-1) \\ \frac{k - f(m-1)}{f(m) - f(m-1)}, & f(m-1) \le k < f(m) \\ \frac{f(m+1) - k}{f(m+1) - f(m)}, & f(m) \le k \le f(m+1) \\ 0, & k > f(m+1) \end{cases},$$



Обычно окон несколько десятков. Фильтры собираются в области низких частот, обеспечивая более высокое “разрешение” там, где это необходимо для распознавания.

Полученные значения логарифмируются. Мотивация обоснована нашим способом восприятия звуков:
громкость измеряется в децибелах, то есть в логарифмическом масштабе.

$$S_j[m] = \ln\left(\sum_{k=0}^{N-1} P_j[k]\, H_m[k]\right), \quad m = 0, \ \ldots, M-1.$$

Этот набор отвечает за долю энергии в диапазоне частот по мел-шкале, который был выделен треугольным окном.

Шаги для вычисления MFCC

6. К вектору S_j применяется дискретное косинусное преобразование. В результате получим мел-кепстральные коэффициенты или MFCC:

$$c_j[n] = \sum_{m=0}^{M-1} S_j[m] \cos\left(\pi n \frac{\left(m + \frac{1}{2}\right)}{M}\right), \quad n = 0, \ \ldots, M-1.$$

Как правило, только первые 10-20 коэффициентов сохраняются. Это связано с тем, что более высокие коэффициенты представляют более быстрые изменения энергии по набору фильтров, и оказывается, что эти быстрые изменения фактически ухудшают производительность распознавания речи, поэтому имеет смысл их отбросить.

Коэффициенты по сути ухватывают как быстро изменяется доля энергии сигнала по мел-шкале.