MFCC

입력된 신호에서 노이즈 및 배경 소리로부터 실제 유효한 소리의 특징을 추출하는 방법으로, 일정 **구간**(Short time)식 나누어, 이 구간에 대한 **스펙트럼을 분석**하여 특징을 추출하는 기법

6개의 단계로 나눠서 사용

**1.입력 시간 도메인의 소리 신호를 작은 크기 프레임으로 분리**

MFCC에서는 필터 프레임의 길이를 20~40ms 정도로 정하고 있다. 각각의 Sampling Rate에 따라 샘플의 수가 달라지게 된다.

예제에서는 프레임의 길이가 23ms인 샘플링을 거친 후 12 개의 독립적인 MFCC Coefficient 가 만들어진다.

**2. 각 프레임에 대하여 Power Spectrum의 Periodogram estimate(Periodogram Spectral Estimate)을 계산**

인간은 소리를 들을 때 인간의 귀는 입력된 소리의 주파수에 따라 달팽이관의 진동하는 부위가 다른 걸 이용해서 달팽이관의 어떤 지점이 진동하느냐에 따라, 뇌에서 어떤 주파수가 들어왔는지를 알 수 있게 된다. Periodogram Estimation (주기도 평가)는 기계가 인간이 소리를 판단하는 방식을 이용하여 소리를 구별하게 된다.

-단일 프레임에 대해 DFT를 취하는 식

https://dthumb-phinf.pstatic.net/?src=%22https%3A%2F%2Fssl.pstatic.net%2Fimages.se2%2Fsmedit%2F2017%2F4%2F21%2Fj1rbor2zcb8oum.jpg%22&type=w2

s(n)  : 시간 영역에서의 신호

si(n) : n 은 1 ~ 2048 의 샘플, i 는 프레임의 번호

h(n) : N 개의 샘플을 가지는 윈도우

K : Length of DFT

Periodogram-based power spectral Estimate 구하는 식

https://dthumb-phinf.pstatic.net/?src=%22https%3A%2F%2Fssl.pstatic.net%2Fimages.se2%2Fsmedit%2F2017%2F4%2F21%2Fj1rboxnb21m8hg.jpg%22&type=w2

1번의 결과 에 따라서 2048 포인트 DFT를 하게 되면, Coefficent는 1025개를 구할 수 있다.

**3. 2번에서 구한 Power Spectrum 에 Mel Filter bank를 적용하고, 각 필터에 에너지를 함**

**Mel Filter Bank**

여기서 사용자는 구간에서 얼마만큼의 에너지가 발생하는지 대략적으로 아는 것이 중요하다 **Mel Scale은** Filter Bank를 나눌 때 어떤 간격으로 나누어야 하는지 알려주며, 간격을 나누는 방법은 다음과 같다.

Mel Scale

Convert Frequency to Mel Scale :    (1) https://dthumb-phinf.pstatic.net/?src=%22https%3A%2F%2Fssl.pstatic.net%2Fimages.se2%2Fsmedit%2F2017%2F4%2F21%2Fj1r8vi71njeet6.jpg%22&type=w2

Convert Mel Scale to Frequency :  (2)

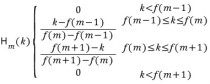
https://dthumb-phinf.pstatic.net/?src=%22https%3A%2F%2Fssl.pstatic.net%2Fimages.se2%2Fsmedit%2F2017%2F4%2F21%2Fj1ranzdgzgknsx.jpg%22&type=w2

FFT 결과 생성된 1025개 Coefficient가  26개의 삼각파형 벡터에 저장된다. 이렇게 만들어진 필터는 Periodogram-Based Power Spectral Estimate (1025개 Coefficient) 의 결과에 적용할 시 자기 스팩트럼에 해당하는 일부 구간을 제외한 다른 주파수 영역에 대해서는 모두 0의 값을 가지게 된다.

Filter Bank Energy = Filter Bank와 Power Spectrum 사이의 행렬 곱

그 결과 26개의 수만 남게 되고, Filterbank 에서 얼마만큼의 에너지를 가지는 보여주게 된다.

Filter Bank의 일반식



M : Filter의 수

m : Filter의 인덱스

f() : M+2Mel-Spaced Frequency 의 리스트

**4. 3번에서 구한 모든 필터 뱅크 에너지의 Log를 취함.**

 압축 연산은 인간이 실제로 듣는 것과 유사하게 소리의 특징을 만듦,

로그를 취하는 이유는 이후에 채널 Noramlization기법인 Cepstral Mean Subtraction을 사용 할 수 있다. 인간의 귀는 소리의 크기를 Linear -Scale 로 감지하는 것이 아니기 때문에 소리의 전체 크기가 크다면, 에너지 변화가 좀 크더라 하더라도 실제 큰 차이가 없을 수 있다.

**5. 4번 값에 DCT를 계산함.**

DCT를 사용하는 이유는 Filter Bank는 모두 Overlapping 되어 있기 때문에 Filter Bank 에너지들 사이에 상관관계가 존재하기 때문에 에너지들 사이에 이러한 상관관계를 분리 해주는 역할을 함. HMM Classifier 와 유사한 Diagonal Covariance Matrice 사용이 가능해진다.

**6. DCT를 취한 값에 Coefficients 2~13 만 남기고 나머지는 버림**

26개 DCT Coefficient 들 중 12개만 남긴다. DCT Coefficient 가 많으면, Filter Bank 에너지의 빠른 변화를 나타내게 되고, 이는 음성인식의 성능을 낮추게 된다.