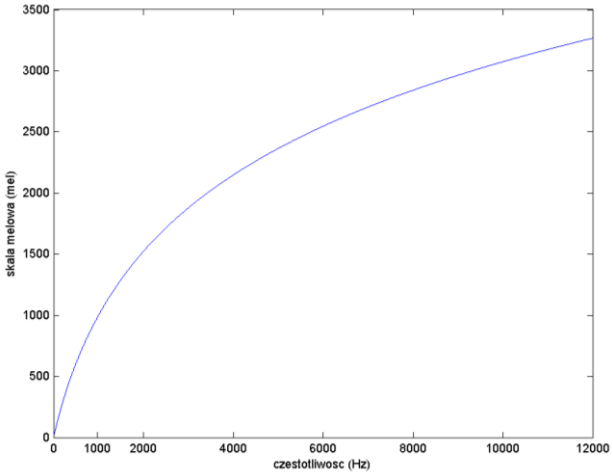
Instrukcja do laboratorium nr 8
Przetwarzanie Sygnałów Diagnostycznych

Cechy Mel-Cepstralne Sygnału Mowy.

1. Wprowadzenie.

Ucho człowieka reaguje nieliniowo na częstotliwości sygnału dźwięku – różnice w zakresie niskich częstotliwości (poniżej 1 kHz) są łatwiej wykrywane niż podobne różnice w zakresie wysokich częstotliwości słyszalnego spectrum sygnału. Czyli im wyższa częstotliwość, tym gorsza dokładność – tym większe odstępy między kolejnymi pasmami są potrzebne dla zrekompensowania nieliniowości. Omawiana zależność jest opisana poprzez skale Mel przeliczoną dla pasma częstotliwościowego mowy i przedstawioną na rysunku 1.



Rys. 1. Związek między skala częstotliwości a skala Mel.

Jako punkt odniesienia przyjmuje się ton 1 kHz, dla którego krzywa znajduje się 40 dB ponad progiem słyszenia człowieka i oznacza się go jako 1000 meli. Powszechnie używa się zależności (1.1 lub 1.2) na przybliżenie skali Mel ze skali częstotliwościowej i odwrotnie.

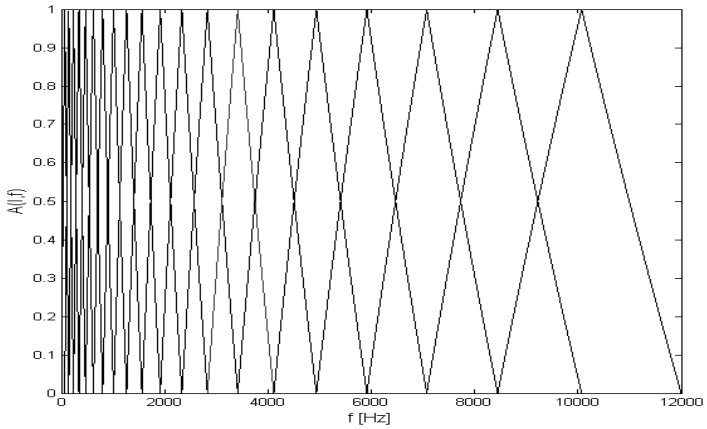
$$mel(f) = 2595 \cdot \log(1 + \frac{f}{700}), \quad f(mel) = 700 \cdot \left(10^{m/2595} - 1\right) \tag{1.1}$$

$$mel(f) = 1127{,}01048 \cdot \ln(1 + \frac{f}{700}), \quad f(mel) = 700 \cdot \left(e^{m/1127{,}01048} - 1\right) \tag{1.2}$$

2. Filtry pasmowe.

Tworzony jest zbiór filtrów dla kolejnych pasm częstotliwości, rozmieszczonych w nieliniowy sposób w dziedzinie częstotliwości a liniowo w skali Mel. Filtrami są najczęściej symetryczne trójkąty o zadanej podstawie (np. 200 lub 300 melów), które są przesuwane z 50% nakładkowaniem się (np. co 100 lub 150 melów) od dolnej (np. 100 lub 150 melów) do górnej (np. 2840 melów ≅ 8000 Hz, 3266 melów ≅ 12000 Hz) częstotliwości badanej[1].
Na rysunku 2. przedstawiono charakterystykę banku 20 filtrów trójkątnych rozmieszczonych zgodnie ze skala Mel w zakresie częstotliwości 1 Hz do 12 kHz.

[1] Górna granica częstotliwości badanej dla tzw. mowy prawidłowej wynosi 8 kHz, dla mowy zdeformowanej 12 kHz.



Rys. 2. Charakterystyka banku 20 filtrów trójkątnych rozmieszczonych zgodnie ze skalą Mel

3. Cechy Mel-Spektralne

Wykorzystując zbiór trójkątnych filtrów D(l,k) dla obliczenia (np. L = 20) tzw. współczynników mel-spektralnych MFC(l,τ) dla każdej ramki sygnału (τ - okna czasowego) zgodnie z (1.3):

$$MFC(l,\tau) = \sum_{k=0}^{M-1}[D(l,k)\cdot G(k,\tau)], \quad l = 1,...,L \tag{1.3}$$

gdzie:

$G(k,\tau) = |STFT(k,\tau)|^2$ spektrum energii dla każdej ramki sygnału (τ - okna czasowego).

Wartość pojedynczego współczynnika MFC odpowiada ważonej sumie wartości FC należnych do zakresu trójkątnego filtra pasmowego odpowiadającego danemu MFC.

4. Współczynniki Mel-Cepstralne

Ostatecznie wyznacza się pewna liczbę (np. K=12) współczynników mel-cepstralnych (MFCC[2]) według wzorów (1.4):

$$MFCC(k,\tau) = \sqrt{\frac{2}{L}}\sum_{l=1}^{L}\log[MFC(l,\tau)]\cdot\cos\left[\frac{\pi k(2l-1)}{2L}\right] \quad k = 1,2,...,K \tag{1.4}$$

gdzie:
L – liczba zastosowanych funkcji wagowych (filtrów)
K – liczba wyznaczonych współczynników mel-cepstralnych

Ponieważ sygnał mowy ma charakter ciągły, zatem poziomy energii w sąsiednich pasmach są skorelowane. Dlatego też niezbędna do tego transformata odwrotna Fouriera (IDFT) zamienia zbiór logarytmów energii na nieskorelowane ze sobą współczynniki cepstralne. Aby zmniejszyć ilość obliczeń ograniczając je wyłącznie do liczb rzeczywistych zamiast IDFT stosuje się odwrotną transformację cosinusową (IDCT), ponieważ widmo amplitudowe jest parzyste, gdyż sygnał mowy przyjmuje wyłącznie wartości rzeczywiste.

Nadal jednak podstawowej częstotliwości drgań krtaniowych i jej harmoniczne mogą nakładać się na amplitudy mierzonych częstotliwości. Dlatego z punktu widzenia rozpoznawania głosek stosuje się przetwarzanie końcowe zwane ***liftrowaniem*** (*ang*. *fil* | *ter -> lif* | *ter*). Celem kroku liftrowania jest usunięcie szkodliwego wpływu podstawowych drgań krtaniowych na zestaw cech poprzez ograniczenie ilości cech MFCC.

[2] Mel Frequency Cepstral Coefficient

5. Przykład
Niech częstotliwość sygnału wynosi 16 000 Hz, a szerokość okna czasowego odpowiada 256 próbkom, czyli czasowi 16 ms. Wtedy 128 rozpatrywanych współczynników spektrum obejmuje zakres częstotliwości 0÷8000 Hz. Niech liczba cech MFC wynosi L=32 (odpowiada to liczbie filtrów trójkątnych). W takich warunkach, po powrocie do dziedziny czasu, jednostka „czasu cepstralnego" jest 4-krotnie (128/32) większa i wynosi 16 ms/(4*16 ms) = 0.25 ms. To oznacza, że częstotliwość próbkowania w dziedzinie cepstralnej wynosi 4000 Hz. Pierwsza cecha MFCC odpowiada wtedy za pasmo częstotliwości oryginalnego sygnału wokół 4000 Hz (4000 Hz / 1), a 12 cecha – za pasmo wokół 333 Hz (4000 Hz / 12). Załóżmy, że częstotliwość podstawowa mówcy (tzw. F0) wynosi ok. 220 Hz. Odpowiada temu cecha cepstralna o indeksie k= 18, gdyż 222 Hz ≈ 4000 Hz / 18. Bardzo często praktykuje się odwrócenie numeracji cech MFCC, niskie współczynniki odpowiadają za niskie pasmo częstotliwościowe, a wysokie współczynniki odpowiadają za wysokie pasmo częstotliwościowe.
Przykład ten ilustruje, że cechy MFCC o indeksie równym lub wyższym niż indeks cechy, który odpowiada częstotliwości podstawowej mówcy, nie są reprezentatywne dla wymawianej treści, a jedynie dla samego mówcy. Dlatego nie powinny być one brane pod uwagę przez system rozpoznawania mowy.

3. Zadania do wykonania
a) przy wykorzystaniu funkcji mfcc (plik mfcc.p) wyznaczyć współczynniki Mel-Cepstralne dla typowych głosek i słów języka polskiego.

gdzie:
```
[MFCC]=mfcc(fname,nfft,K,w)
%wyznaczenie mel-cepstralnych wspolczynnikow sygnalu
%fname - nazwa pliku typu wave
%nfft - szerokość okna czasowego dla STFFT
%K - liczba wyznaczonych współczynników
%w – rodzaj okna czasowego
if nargin < 2 nfft = 256; end
if nargin < 3 K = 12; end
if nargin < 4 w = 'hamming'; end
```

b) wyznaczyć uśrednione w czasie współczynniki Mel-Cepstralne, wyniki porównać dla różnych mówców, ale tych samych wypowiedzi. Wyniki przedstawić graficznie.

c)* opracować własną funkcję realizującą wyznaczanie współczynników Mel-Cepstralnych.

4. Sprawozdanie
Sprawozdanie z laboratorium obejmuje wykresy wykonanych zadań z punktu 3a) i 3b).
Zadanie 3c)* na ocenę 5! dla chętnych osób (proszę dołączyć kod źródłowy i wyniki w formie graficznej dla przykładowego sygnału mowy). Sprawozdanie (jedno na osobę) wyłącznie w wersji PDF przesłanej przez stronę kursu Platformy e-Learningowej AGH.