



Deskryptory dźwięku w standardzie MPEG-7



Standard MPEG7

- Multimedia Content Description Language
- Główne cele:
 - opis zawartości multimedialnej
 - elastyczność w zarządzaniu danymi
 - globalizacja i wewnętrzna kompatybilność zasobów danych

Standard MPEG7

- Części:
 - Part I: System
 - Part II: Description Definition Language (DDL)
 - Part III: Visual
 - Part IV: Audio
 - Part V: Multimedia Description Schemes (MDS)
 - Part VI: Reference Software
 - Part VII: Conformance Testing
 - Part VIII: Extraction And Use Of MPEG7 Descriptions
 - Part IX: Profiles
 - Part X: Schema Definition

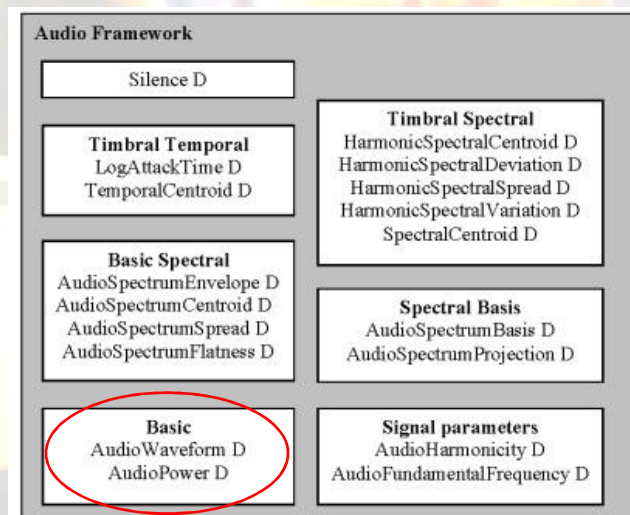
MPEG7 Audio

- Standard MPEG-7 zawiera szereg rozwiązań przeznaczonych opisu danych multimedialnych.
- Dla danych dźwiękowych są to m. in.:
 - format danych (rodzaj kodowania, częstotliwość próbkowania itp.)
 - informacje takie jak autor bądź nazwa instrumentu
 - deskryptory wyznaczone z sygnału dźwiękowego

Low Level Audio Descriptors

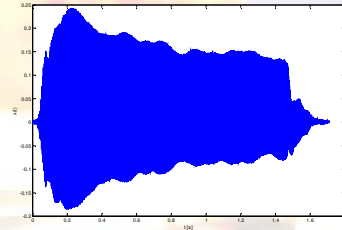
- Reprezentacja danych dźwiękowych w standardzie MPEG7
 - obliczane na podstawie reprezentacji czasowej i widmowej sygnału,
 - deskryptory widmowe są wyznaczane na podstawie analizy widma sygnału obliczonego w kolejnych ramkach czasowych,
 - możliwość opisu ciągu danych za pomocą parametrów statystycznych (wartość minimalna, maksymalna, średnia, wariancja).

Low Level Audio Descriptors

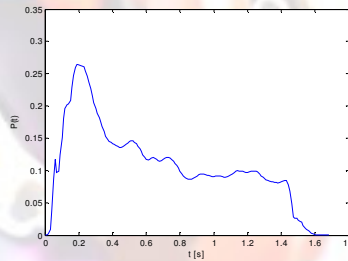


Basic

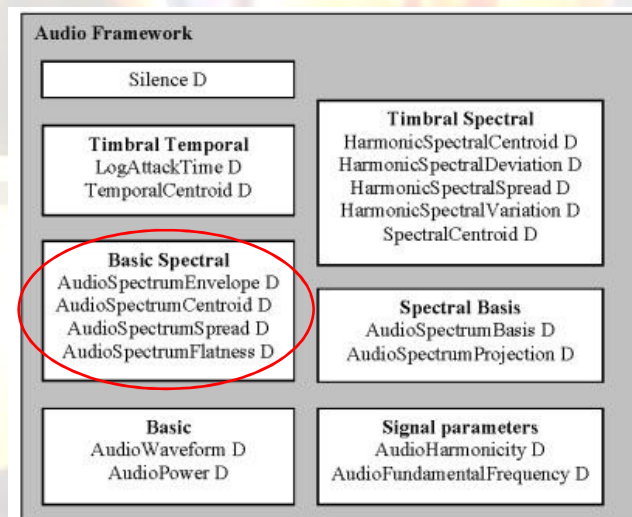
- AudioWaveform
 - Przebieg czasowy sygnału $s(t)$



- AudioPower
 - moc sygnału $P(t) = |s(t)|^2$



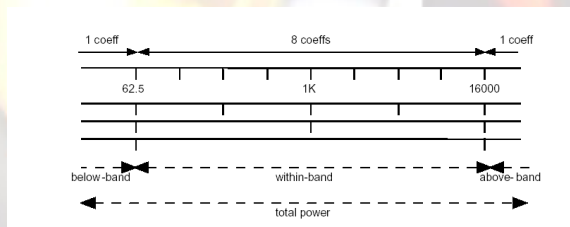
Low Level Audio Descriptors



Audio Spectrum Envelope

- *Audio Spectrum Envelope* (ASE) – jest zdefiniowany jako krótkookresowe widmo gęstości mocy P_x wyznaczone dla częstotliwości w odstępach logarytmicznych (pasma o szerokości 1/16, 1/8, 1/4, 1/2, 1, 2, 4 lub 8 oktaw)

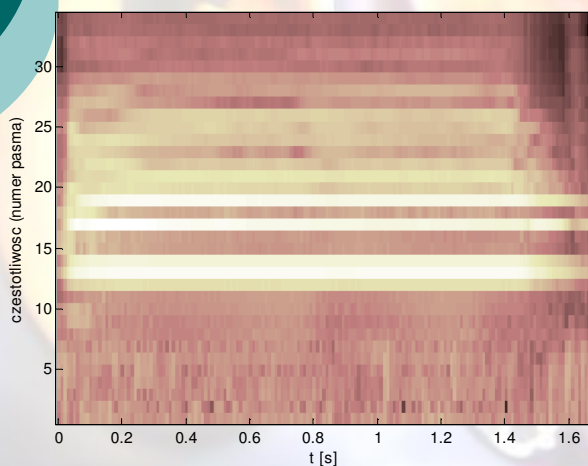
$$P_x(n) = \frac{1}{lw \cdot NFFT} |X(n)|^2$$



Przykład podziału widma na pasma o szerokości jednej oktawy

Audio Spectrum Envelope

Rozdzielczość 1/4 oktawy



Pasma:

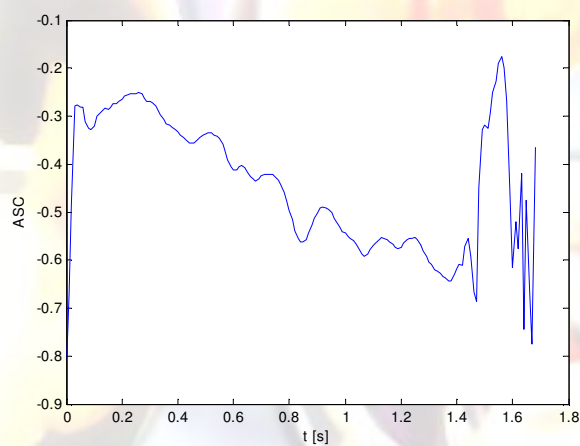
- 1: do 62,5 Hz
- 2-5: 62,5 – 125 Hz
- 6-9: 125 – 250 Hz
- 10-13: 250 – 500 Hz
- 14-17: 500 – 1000 Hz
- 18-21: 1 – 2 kHz
- 22-25: 2-4 kHz
- 26-29: 4-8 kHz
- 30-33: 8 – 16 kHz
- 34: powyżej 16 kHz

Audio Spectrum Centroid

- *Audio Spectrum Centroid* (ASC) – jest zdefiniowany jako środek ciężkości widma gęstości mocy, wyskalowany w oktawach w stosunku do 1 kHz

$$ASC = \frac{\sum_n \log_2(f(n)/1000)P_x(n)}{\sum_n P_x(n)}$$

Audio Spectrum Centroid



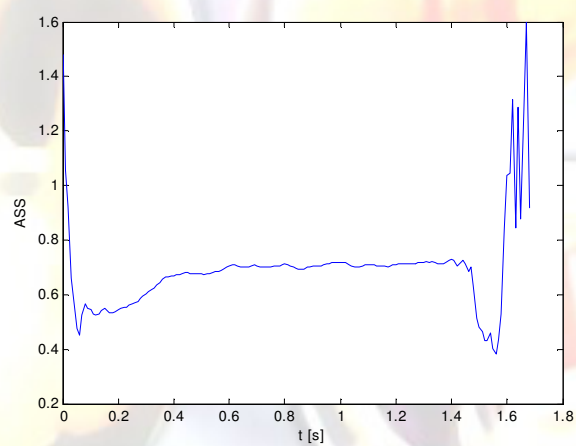
ASC = -0.44555

Audio Spectrum Spread

- *Audio Spectrum Spread (ASS)* – jest zdefiniowany jako odchylenie średniokwadratowe (RMS) widma gęstości mocy w skali oktaawowej, względem Audio Spectrum Centroid

$$ASS = \sqrt{\frac{\sum_n ((\log_2(f(n)/1000) - ASC)^2 P_x(n))}{\sum_n P_x(n)}}$$

Audio Spectrum Spread



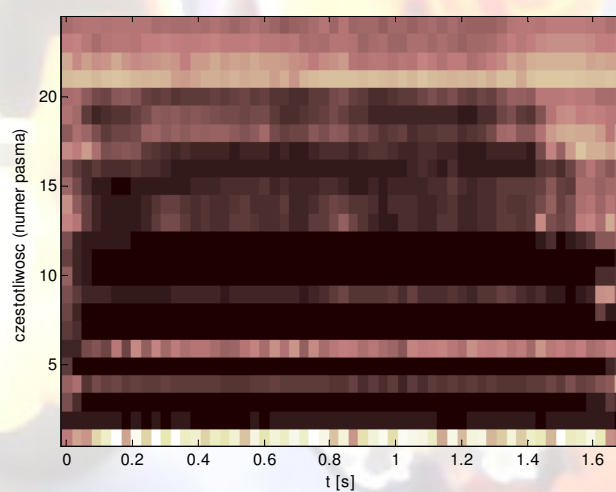
ASS = 0.69174

Audio Spectrum Flatness

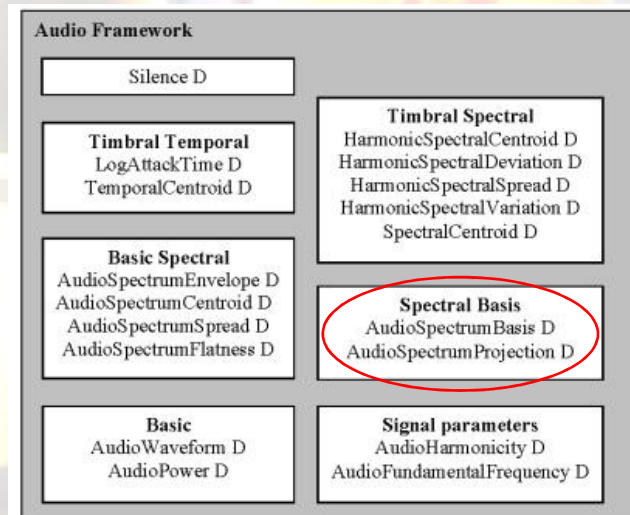
- *Spectral Flatness Measure* (SFM) – jest zdefiniowany jako stosunek średniej geometrycznej i średniej arytmetycznej współczynników widma gęstości mocy w pasmach (b) o szerokości $\frac{1}{4}$ oktawy

$$SFM_b = \frac{\sqrt[ih(b)-il(b)+1]{\prod_{i=il(b)}^{ih(b)} c(i)}}{\frac{1}{(ih(b)-il(b)+1) \sum_{i=il(b)}^{ih(b)} c(i)}}$$

Audio Spectrum Flatness



Low Level Audio Descriptors



Audio Spectrum Basis Audio Spectrum Projection

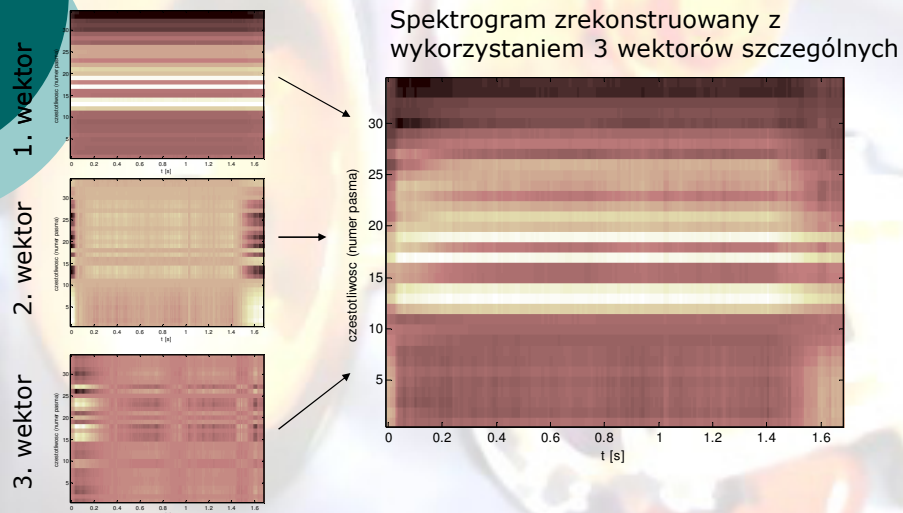
- Deskryptory te zawierają informacje o widmie mocy w postaci zredukowanej liczby danych, uzyskane za pomocą rozkładu macierzy względem wartości szczególnych (SVD).
- Dekompozycji poddawana jest macierz \mathbf{X} , której wiersze zawierają widmo dla kolejnych ramek czasowych (*AudioSpectrumEnvelope*). W wyniku otrzymuje się macierze wektorów \mathbf{U} i \mathbf{V} oraz diagonalną macierz wartości szczególnych \mathbf{S} .

$$\mathbf{X} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

Audio Spectrum Basis Audio Spectrum Projection

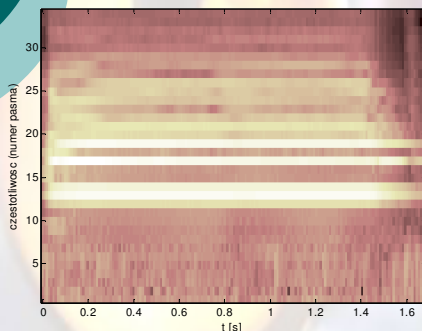
- Do wyznaczenia deskryptorów *AudioSpectrumBasis* i *AudioSpectrumProjection* wykorzystuje się kilka pierwszych kolumn macierzy \mathbf{V} . Pozwala to dokonać aproksymacji macierzy \mathbf{X} , przez ograniczenie się jedynie do kilku pierwszych wektorów szczególnych.
- *AudioSpectrumBasis* zawiera funkcje bazowe widma (pierwszych kilka, zwykle 3-10, kolumn macierzy \mathbf{V}).
- *AudioSpectrumProjection* zawiera funkcje przekształcające wyznaczone w oparciu o analizowaną macierz \mathbf{X} oraz kilka pierwszych wektorów szczególnych macierzy \mathbf{V} .
- Rekonstrukcję spektrogramu uzyskuje się poprzez odpowiednie wymnożenie zawartości deskryptorów *AudioSpectrumBasis* i *AudioSpectrumProjection*

Audio Spectrum Basis Audio Spectrum Projection



Audio Spectrum Basis Audio Spectrum Projection

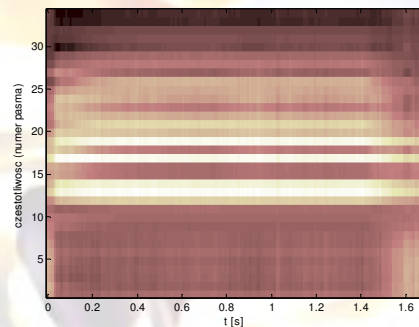
Spektrogram oryginalny



$M \times N$

Dla $M=100$ i $N=34$:
3400 wartości

Spektrogram zrekonstruowany z wykorzystaniem 3 wektorów szczególnych



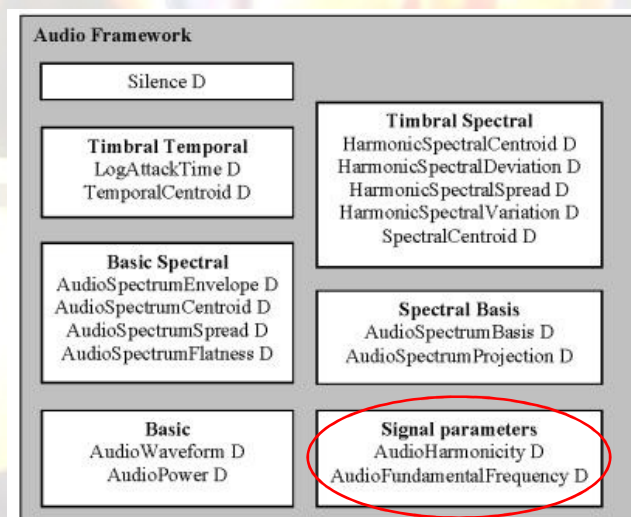
$3(M+N)$

Dla $M=100$ i $N=34$:
402 wartości

Zajętość pamięci

M – ilość pasm częstotliwości
 N – ilość ramek czasowych

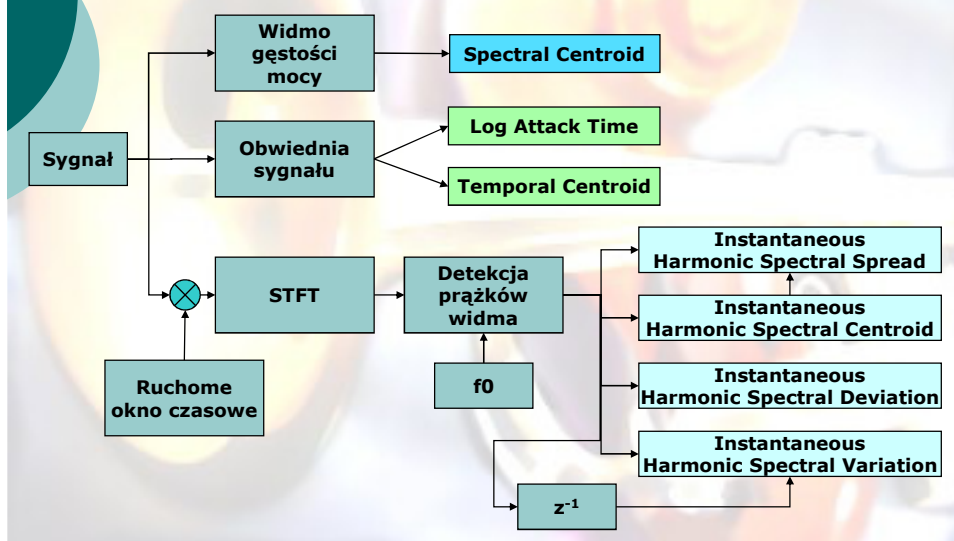
Low Level Audio Descriptors



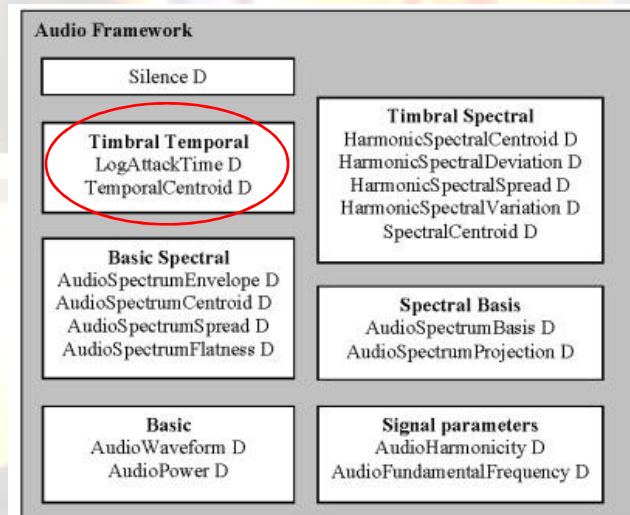
Audio Fundamental Frequency Audio Harmonicity

- **Audio Fundamental Frequency** – częstotliwość podstawowa dźwięku [Hz]
 - Sposób detekcji częstotliwości podstawowej nie jest ustalony w standardzie MPEG7
- **AudioHarmonicity** – zawiera informacje o stopniu harmoniczności (okresowości) sygnału
 - Harmonic Ratio – zawartość składowych harmonicznych w widmie sygnału (0 – biały szum, 1 – sygnał okresowy)
 - Upper Limit Of Harmonicity – częstotliwość (skala oktawa w odniesieniu do 1 kHz), powyżej której widmo nie wykazuje cech harmoniczności

Timbre Descriptors



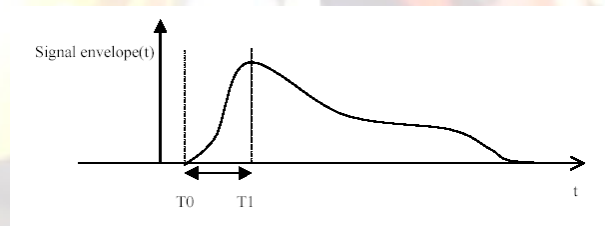
Low Level Audio Descriptors



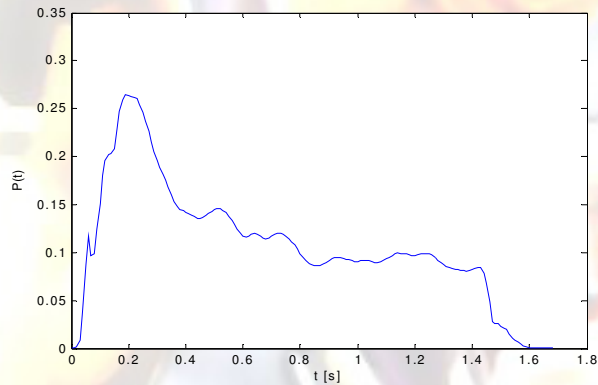
Log Attack Time

- Jednostka: [log s]
- Zakres: [log 1/SR, określone przez czas trwania sygnału]
- *Log Attack Time (LAT)* – jest zdefiniowany jako logarytm dziesiętny czasu od chwili, gdy sygnał się rozpoczyna (T_0) do chwili, gdy osiąga stan ustalony (T_1).

$$LAT = \log_{10}(T_1 - T_0)$$



Log Attack Time



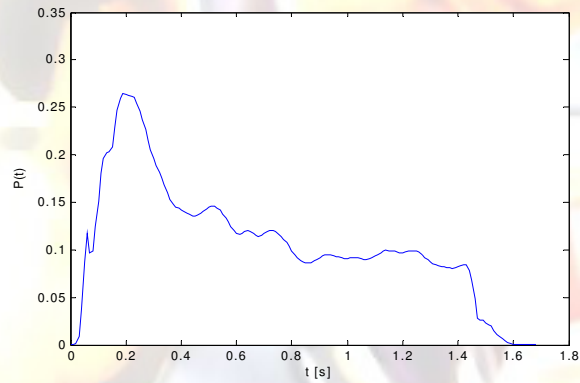
$$\text{LAT} = \log(0,18 \text{ s}) = -0,74 \text{ [log s]}$$

Temporal Centroid

- Jednostka: [s]
- Zakres: [0, określone przez czas trwania sygnału]
- *Temporal Centroid (TC)* – jest zdefiniowany jako środek ciężkości obwiedni mocy sygnału w dziedzinie czasu

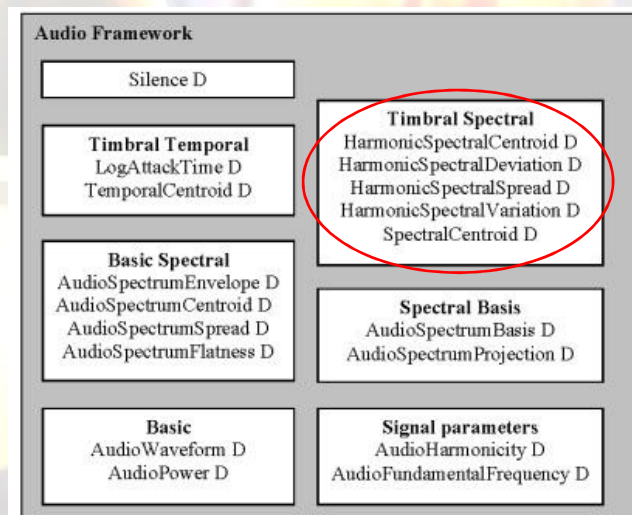
$$TC = \frac{\sum_{n=1}^{\text{length}(SE)} n/SR \cdot SE(n)}{\sum_{n=1}^{\text{length}(SE)} SE(n)}$$

Temporal Centroid



TC = 0,72 [s]

Low Level Audio Descriptors



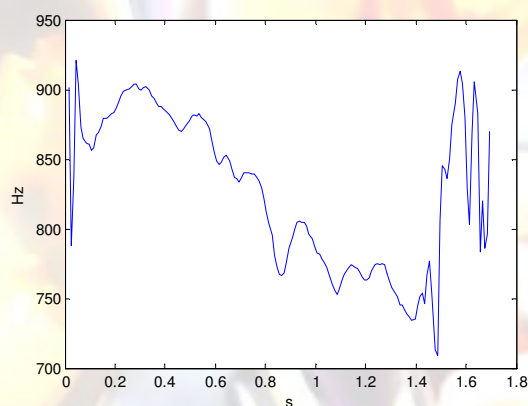
Spectral Centroid

- Jednostka: [Hz]
- Zakres: [0, SR/2]
- *Spectral Centroid (SC)* – jest zdefiniowany jako środek ciężkości widma, czyli średnia ważona częstotliwość współczynników widma gęstości mocy.

$$ISC = \frac{\sum_{k=1}^{\text{length}(S)} f(k) \cdot S(k)}{\sum_{k=1}^{\text{length}(S)} S(k)}$$

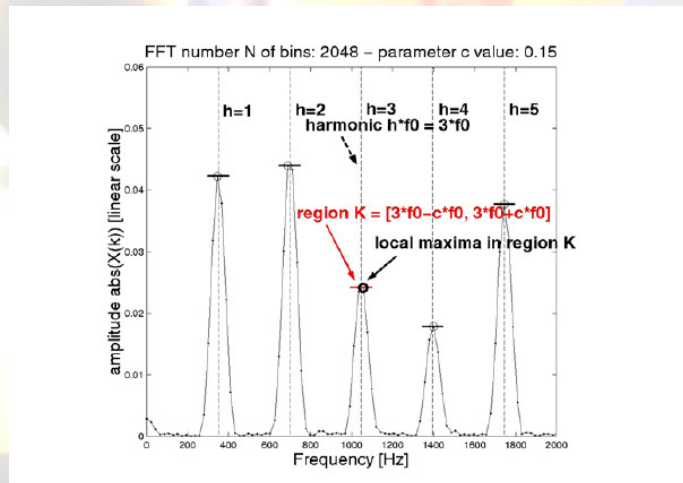
$$SC = \frac{\sum_{i=1}^{nb_f} ISC(i)}{nb_f}$$

Spectral Centroid



SC = 827 [Hz]

Detekcja prążków widma



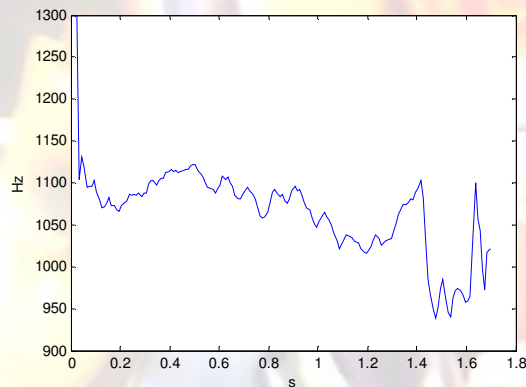
Harmonic Spectral Centroid

- Jednostka: [Hz]
- Zakres: $[0, SR/2]$
- *Harmonic Spectral Centroid (HSC)* – jest zdefiniowany jako ważona amplitudowo średnia częstotliwość prążków widma

$$IHSC = \frac{\sum_{h=1}^{nb_h} f(h) \cdot A(h)}{\sum_{h=1}^{nb_h} A(h)}$$

$$HSC = \frac{\sum_{i=1}^{nb_f} IHSC(i)}{nb_f}$$

Harmonic Spectral Centroid



HSC = 1068 [Hz]

Harmonic Spectral Deviation

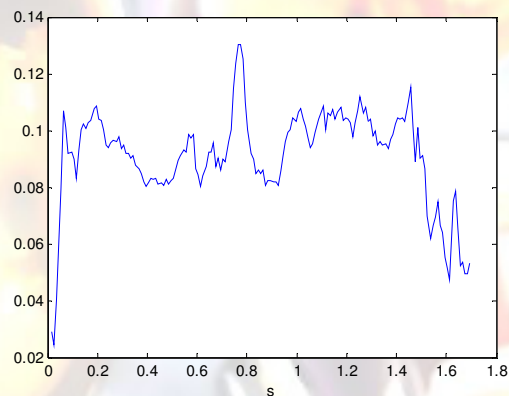
- Jednostka: [-]
- Zakres: [0, 1]
- *Harmonic Spectral Deviation (HSD)* – jest zdefiniowany jako średnie odchylenie logarytmu amplitudy prążków widma od obwiedni widma (SE)

$$SE(h) = \begin{cases} \frac{A(h)+A(h+1)}{2}, & \text{dla } h=1 \\ \frac{\sum_{i=1}^1 A(h+i)}{3}, & \text{dla } h \in [2, nb_h-1] \\ \frac{A(h-1)+A(h)}{2}, & \text{dla } h = nb_h \end{cases}$$

$$IHSD = \frac{\sum_{h=1}^{nb_h} |\log_{10} A(h) - \log_{10} SE(h)|}{\sum_{h=1}^{nb_h} |\log_{10} A(h)|}$$

$$HSD = \frac{\sum_{i=1}^{nb_f} IHSD(i)}{nb_f}$$

Harmonic Spectral Deviation



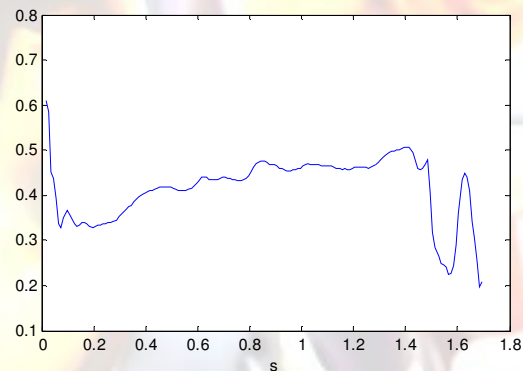
HSD = 0,091

Harmonic Spectral Spread

- Jednostka: [-]
- Zakres: [0, 1]
- *Harmonic Spectral Spread (HSS)* – jest zdefiniowany jako ważone amplitudowo standardowe odchylenie amplitud prążków widma, znormalizowane przez *Harmonic Spectral Centroid (HSC)*

$$IHSS = \frac{1}{IHSC} \sqrt{\frac{\sum_{h=1}^{nb_h} A^2(h) \cdot (f(h) - IHSC)^2}{\sum_{h=1}^{nb_h} A^2(h)}} \quad HSS = \frac{\sum_{i=1}^{nb_f} IHSS(i)}{nb_f}$$

Harmonic Spectral Spread



HSS = 0,416

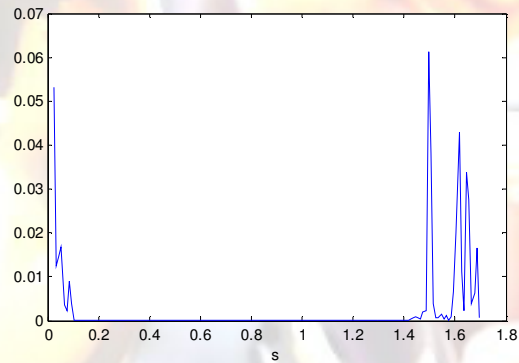
Harmonic Spectral Variation

- Jednostka: [-]
- Zakres: [0, 1]
- *Harmonic Spectral Variation (HSV)* – jest zdefiniowany jako znormalizowana korelacja pomiędzy amplitudami prążków w dwóch sąsiednich ramkach czasowych

$$IHSV = 1 - \frac{\sum_{h=1}^{nb_h} A_{-1}(h) \cdot A(h)}{\sqrt{\sum_{h=1}^{nb_h} A_{-1}^2(h)} \cdot \sqrt{\sum_{h=1}^{nb_h} A^2(h)}}$$

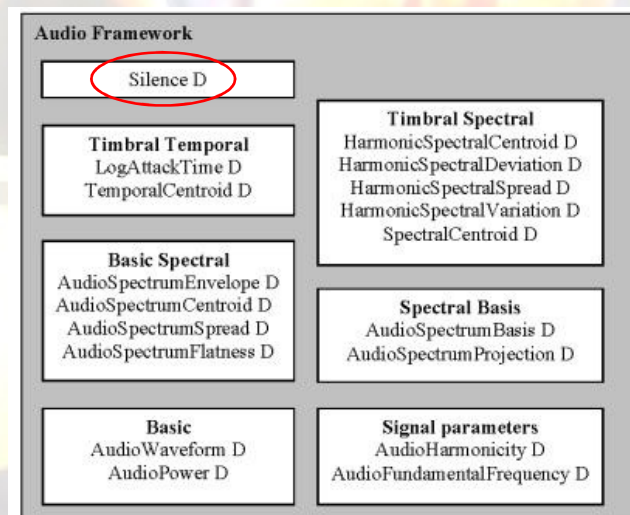
$$HSV = \frac{\sum_{i=2}^{nb_f} IHSV(i)}{nb_f - 1}$$

Harmonic Spectral Variation



HSV = 0,0025

Low Level Audio Descriptors



Silence

- Deskryptor ten pozwala stwierdzić, w którym fragmencie nagrania występują fragmenty ciszy. Podstawowe informacje zawarte w takim deskrytorze to czas rozpoczęcia oraz czas trwania ciszy.
- Deskryptor ten ma charakter semantyczny. Pojęcie ciszy oznacza w tym przypadku brak istotnych dźwięków (np. brak dialogów na ścieżce dźwiękowej filmu, przy obecnych dźwiękach tła).

Silence

- Zastosowaniem tego deskryptora może być automatyczna segmentacja materiału dźwiękowego, np. dzielenie sygnału mowy na zdania lub poszczególne wyrazy, w zależności od przyjętego progu minimalnego trwania ciszy.
- Detekcja ciszy może być implementowana na wiele sposobów. Zwykle uwzględnia się psychofizjologię słuchu i związane z tym pasma krytyczne słyszenia oraz zjawisko maskowania.

Inne deskryptory dźwięku

- *KeyNum* – wyraża wysokość dźwięku zgodnie ze standardem MIDI:

$$KeyNum = 69 + 12 \cdot \log_2 \left(\frac{f_0}{440} \right)$$

- *Ev* – zawartość parzystych harmonicznych w widmie:

$$Ev = \frac{\sqrt{\sum_{h=1}^{\text{int}(nb-h/2)} A^2(2h)}}{\sqrt{\sum_{h=1}^{nb-h} A^2(h)}}$$

Inne deskryptory dźwięku

- Pierwszy, drugi i trzeci zmodyfikowany Tristimulus – określają zawartość grup harmonicznych w widmie sygnału.

$$Tri1 = \frac{A^2(1)}{\sum_{h=1}^{nb-h} A^2(h)} \quad Tri2 = \frac{\sum_{h=2}^4 A^2(h)}{\sum_{h=1}^{nb-h} A^2(h)} \quad Tri3 = \frac{\sum_{h=5}^{nb-h} A^2(h)}{\sum_{h=1}^{nb-h} A^2(h)}$$