

# Sprawozdanie z Analizy Sygnału Audio

Urszula Szczęsna

27 marca 2025

## Spis treści

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Opis Aplikacji</b>                                      | <b>2</b>  |
| <b>2</b> | <b>Opis Metod</b>  | <b>2</b>  |
| 2.1      | Wczytywanie Sygnału Audio . . . . .                        | 2         |
| 2.2      | Wizualizacja Przebiegu Czasowego . . . . .                 | 2         |
| 2.3      | Detekcja Ciszy . . . . .                                   | 2         |
| 2.4      | Parametry na poziomie ramki . . . . .                      | 2         |
| 2.4.1    | Głośność . . . . .   | 2         |
| 2.4.2    | STE Short Time Energy . . . . .                            | 2         |
| 2.4.3    | ZCR Zero Crossing Rate . . . . .                           | 3         |
| 2.5      | Detekcja Częstotliwości Podstawowej (F0) . . . . .         | 3         |
| 2.5.1    | Autokorelacja . . . . .                                    | 3         |
| 2.5.2    | AMDF . . . . .   | 3         |
| 2.6      | Segmentacja Dźwięczne/Bezdźwięczne . . . . .               | 4         |
| 2.7      | Klasyfikacja Mowa/Muzyka . . . . .                         | 4         |
| 2.8      | Cechy sygnału audio na poziomie klipu . . . . .            | 4         |
| 2.8.1    | Bazujące na głośności . . . . .                            | 4         |
| 2.8.2    | Bazujące na energii . . . . .                              | 5         |
| 2.8.3    | Bazujące na ZCR . . . . .                                  | 5         |
| <b>3</b> | <b>Prezentacja Wyników Działania</b>                       | <b>6</b>  |
| 3.1      | Zdanie - kobiecy głos . . . . .                            | 6         |
| 3.2      | Porównanie głos ludzki (radio) i muzyka (gitara) . . . . . | 8         |
| 3.3      | Porównanie głosu męskiego i żeńskiego . . . . .            | 9         |
| <b>4</b> | <b>Wnioski</b>   | <b>10</b> |

# 1 Opis Aplikacji

Aplikacja jest interaktywnym narzędziem do odczytywania oraz przedstawiania cech sygnału audio w dziedzinie czasu. Główne biblioteki użyte w projekcie to:

- **Numpy** - operacje matematyczne na wektorach.
- **Librosa** - wczytywanie sygnału audio
- **Plotly** - interaktywne wizualizacje
- **Pandas** - operacje tabelaryczne
- **Streamlit** - budowa interfejsu użytkownika

## 2 Opis Metod

### 2.1 Wczytywanie Sygnału Audio

Sygnał audio w formacie WAV jest wczytywany za pomocą biblioteki Librosa. Funkcja `librosa.load(file, sr=None)` zwraca częstotliwość próbkowania oraz sygnał w postaci wektora.

### 2.2 Wizualizacja Przebiegu Czasowego

Przebieg czasowy sygnału jest wizualizowany za pomocą biblioteki Plotly. Wykres przedstawia amplitudę sygnału w funkcji czasu.

### 2.3 Detekcja Ciszy

Detekcja ciszy jest realizowana poprzez analizę parametru Zero Crossing Rate, który określa liczbę przejść przez zero w ramce zadanej przez użytkownika (domyślnie 20 ms), oraz średniej energii. Ramki, w których parametr ZCR jest mniejszy od zadanego progu (`threshold`), a energia jest mniejsza od określonego procentu średniej energii w całym nagraniu audio, są klasyfikowane jako cisza..

### 2.4 Parametry na poziomie ramki

Aby analizować poniższe parametry, w pierwszej kolejności należy podzielić sygnał audio na ramki o zadanej przez użytkownika długości. Operację tę wykonuje się za pomocą funkcji `split_into_frames(data, rate, frame_ms)`. Dzieli ona sygnał na ramki o określonej długości w milisekundach i generuje tablicę ramek o długości `frame_size`.

#### 2.4.1 Głośność

Funkcja `loudness(data, rate, frame_ms)` oblicza głośność dla każdej ramki na podstawie poniższego wzoru:

$$p(n) = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} s_n^2(i)}$$

gdzie:

- $N$  to długość ramki
- $s_n(i)$  to amplituda  $i$ -tej próbki w danej ramce

#### 2.4.2 STE Short Time Energy

Funkcja `short_time_energy(data, rate, frame_ms)` pozwala na obliczenie parametru `ste` dla każdej ramki w całym audio o zadanej przez użytkownika długości. Parametr STE to głośność podniesiona do kwadratu, czyli:

$$STE(n) = p(n)^2$$

### 2.4.3 ZCR Zero Crossing Rate

Funkcja `zero_crossing_rate(data, rate, frame_ms)` pozwala na obliczenie parametru zcr dla każdej ramki audio. Parametr ten oznacza liczbę przejść amplitudy przez zero w jednej ramce czasu sygnału audio.

$$ZCR(n) = \frac{1}{2N} \sum_{i=1}^{N-1} |sign(s_n(i)) - sign(s_n(i+1))|$$

gdzie:

- $N$  to długość ramki
- $s_n(i)$  amplituda w  $i$ -tej próbki w danej ramce
- $sign()$  to funkcja znaku.

## 2.5 Detekcja Częstotliwości Podstawowej (F0)

Częstotliwość podstawowa (F0) to najniższa częstotliwość sinusoidalna obecna w sygnale dźwiękowym. Jest ona związana z wysokością dźwięku i odpowiada częstotliwości drgań źródła dźwięku -np. strun głosowych w mowie lub struny instrumentu muzycznego. W aplikacji zostały zaimplementowane 2 sposoby wyznaczania jej.

### 2.5.1 Autokorelacja

Funkcja `autocorrelation(frame)` oblicza korelację sygnału z jego wersją opóźnioną, co pozwala wykryć okresowość w sygnale. Funkcja `np.correlate` w trybie `full` oblicza korelację całkowitą, a następnie odrzucamy część wyników przed środkowym punktem, ponieważ interesują nas tylko dodatnie opóźnienia.

Funkcja: `compute_f0_autocorrelation(data, rate, frame_ms, min_f0=50, max_f0=400)`

1. **Ramki:** Sygnał jest dzielony na małe ramki o zadanej długości w milisekundach. Długość ramki jest określona przez `frame_ms` oraz częstotliwość próbkowania `rate`.
2. **Autokorelacja:** Dla każdej ramki sygnału obliczana jest autokorelacja.
3. **Wyszukiwanie szczytów:** Po obliczeniu autokorelacji, wykrywane są szczyty, które wskazują na okresowość sygnału. Szczyt w autokorelacji odpowiada okresowi podstawowemu (częstotliwości).
4. **Wyznaczanie F0:** Częstotliwość F0 jest obliczana jako odwrotność opóźnienia, które odpowiada pierwszemu szczytowi w autokorelacji (poza zerowym opóźnieniem).
5. **Filtrowanie:** Jeżeli wyznaczone F0 leży poza zakresem zdefiniowanym przez `min_f0` i `max_f0`, jest ustawiane na 0, aby odrzucić błędne wyniki.

Wartości F0 są zbierane dla wszystkich ramek sygnału i zwracane jako lista wyników.

### 2.5.2 AMDF

AMDF jest alternatywną metodą detekcji F0, bazującą na analizie różnic między próbkami w sygnale z opóźnieniem.

Funkcja: `estimate_f0_amdf(data, rate, frame_ms=30, min_freq=50, max_freq=500)`

1. **Ramki:** Sygnał dzielony jest na ramki o długości zależnej od `frame_ms`, a jej rozmiar obliczany na podstawie częstotliwości próbkowania `rate`.
2. **AMDF:** Dla każdej ramki oblicza się różnice między próbkami w zadanym zakresie opóźnień (`min_lag` do `max_lag`). AMDF oblicza średnią różnicę magnitudy między próbkami z opóźnieniem.
3. **Minimalna różnica:** Wartość F0 jest określana przez opóźnienie, które daje najmniejszą wartość AMDF, ponieważ najmniejsza różnica wskazuje na największe podobieństwo sygnału, co odpowiada okresowi F0.
4. **Wyznaczanie F0:** Po znalezieniu najlepszego opóźnienia, F0 obliczane jest jako odwrotność tego opóźnienia.

## 2.6 Segmentacja Dźwięczne/Bezdźwięczne

Segmentacja dźwięczne/bezdźwięczne jest realizowana przy użyciu parametrów STE oraz ZCR. Proces klasyfikacji:

1. **Obliczenie STE i ZCR** funkcja `short_time_energy`) oblicza energię sygnału w krótkich ramkach a funkcja `zero_crossing_rate`) mierzy liczbę przekroczeń zera w ramach.
2. **Klasyfikacja** - Na podstawie tych cech, ramka jest klasyfikowana jako:
  - **Dźwięczna** (1) jeśli  $STE > ste\_threshold$  i  $ZCR < zcr\_threshold$ .
  - **Bezdźwięczna** (0) w przeciwnym przypadku.

Funkcja zwraca klasyfikację ramek oraz rozmiar ramki co pozwala na stworzenie wykresu z zaznaczonymi fragmentami.

## 2.7 Klasyfikacja Mowa/Muzyka

Funkcja `plot_speech_music` klasyfikuje ramki sygnału jako mowa lub muzyka na podstawie współczynnika przekroczeń zera (ZCR).

1. **Obliczanie ZCR:** Funkcja `zero_crossing_rate` oblicza ZCR dla każdej ramki.
2. **Klasyfikacja:**
  - **Mowa:** Ramki z  $ZCR > 0.08$  są klasyfikowane jako mowa.
  - **Muzyka:** Ramki z ZCR w przedziale (0.01, 0.08) są klasyfikowane jako muzyka.

Dla każdej ramki określany jest czas początkowy i końcowy, a ramki przypisane do mowy lub muzyki są przechowywane w osobnych listach.

## 2.8 Cechy sygnału audio na poziomie klipu

Kolejna grupa cech jest wyliczana na poziomie całego klipu lub ramek o długości sekundy. Dlatego poniższe funkcje są jedynie dostępne dla nagrań dłuższych niż 2 sekundy.

### 2.8.1 Bazujące na głośności

**VSTD** czyli odchylenie standardowe głośności podzielone przez maksymalną wartość głośności w całym klipie.

**VDR** określa się wzorem

$$VDR = \frac{\max(v) - \min(v)}{\max(v)}$$

gdzie  $v$  to głośność.

**VU** określa falistość głośności. Oblicza się ją poprzez zliczanie sąsiednich pików i dolin wykresu głośności w całym klipie. Opis algorytmu:

1. **Podział na ramki** – Sygnał audio jest dzielony na ramki o długości określonej przez parametr `frame_ms`, a następnie dla każdej ramki obliczana jest jej głośność.
2. **Obliczenie różnic głośności** – Na podstawie wartości głośności kolejnych ramek obliczana jest różnica między sąsiednimi próbkami (`np.diff(frame_loudness)`).
3. **Wykrywanie zmian trendu** – Zmiany kierunku różnic (czyli przejścia między wzrostem a spadkiem głośności) są wykrywane poprzez sprawdzenie iloczynu sąsiednich wartości różnic (`dif[1:] * dif[:-1] < 0`).
4. **Sumowanie wartości zmian** – Ostatecznie, wartość VU jest określana jako suma bezwzględnych wartości różnic dla wykrytych zmian trendu.

### 2.8.2 Bazujące na energii

**LSTER** Jest to miara zdefiniowana jako odsetek liczby ramek, dla których wartość parametru STE jest mniejsza niż 50% średniej wartości STE w oknie 1 sekundowym. Funkcja `lster(data, rate, frame_ms)` zwraca listę ramek o długości 1 sekundy z obliczonym parametrem `lster` dla każdej z nich.

**Entropia** Entropia informuje o tym, jak energia jest rozłożona w krótkoterminowych segmenach sygnału. Wysoka entropia oznacza, że energia jest równomiernie rozproszona, natomiast niska entropia wskazuje na dominację kilku segmentów o wysokiej energii. Funkcja `energy_entropy(data, rate, frame_ms)` oblicza entropie sygnału audio w następujących krokach:

1. Obliczenie krótkoterminowej energii (`ste`) na ramkach o długości 1 sekundy oraz długości ramki (liczba próbek w pojedynczej ramce)
2. Podział ramki na segmenty o długości  $K = 441$  próbek i wyliczenie energii dla każdego segmentu
3. Normalizacja energii segmentów, czyli podzielenie przez całkowitą energię ramki.
4. Obliczanie entropii ze wzoru :  $H = -\sum_{i=1}^J p_i \log_2 p_i$ , gdzie  $p_i$  to znormalizowana energia i-tego segmentu klipu.

### 2.8.3 Bazujące na ZCR

**ZSTD** Odchylenie standardowe ZCR.

**HZCRR** Wyraża się wzorem :

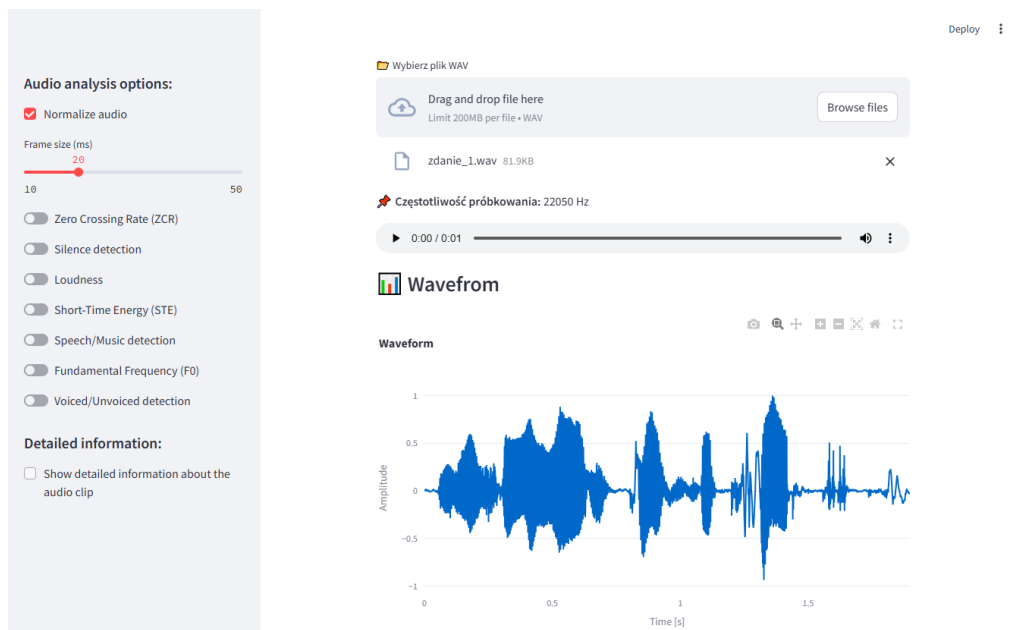
$$HZCRR = \frac{1}{2N} \sum_{n=0}^{N-1} \text{sign}(ZCR(n) - 1.5 \times \text{avZCR}) + 1$$

gdzie:

- $N$  to całkowita liczba ramek
- $ZCR(n)$  to  $zcr$  w  $n$ -tej ramce
- $\text{avZCR}$  to średnia wartość ZCR w 1 sekundowym oknie

## 3 Prezentacja Wyników Działania

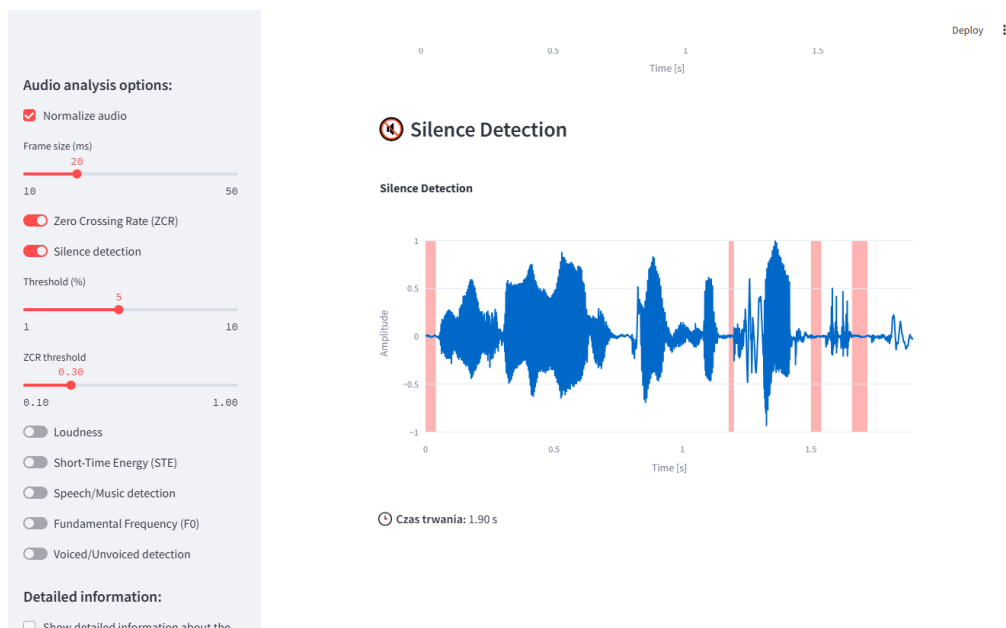
### 3.1 Zdanie - kobiecy głos



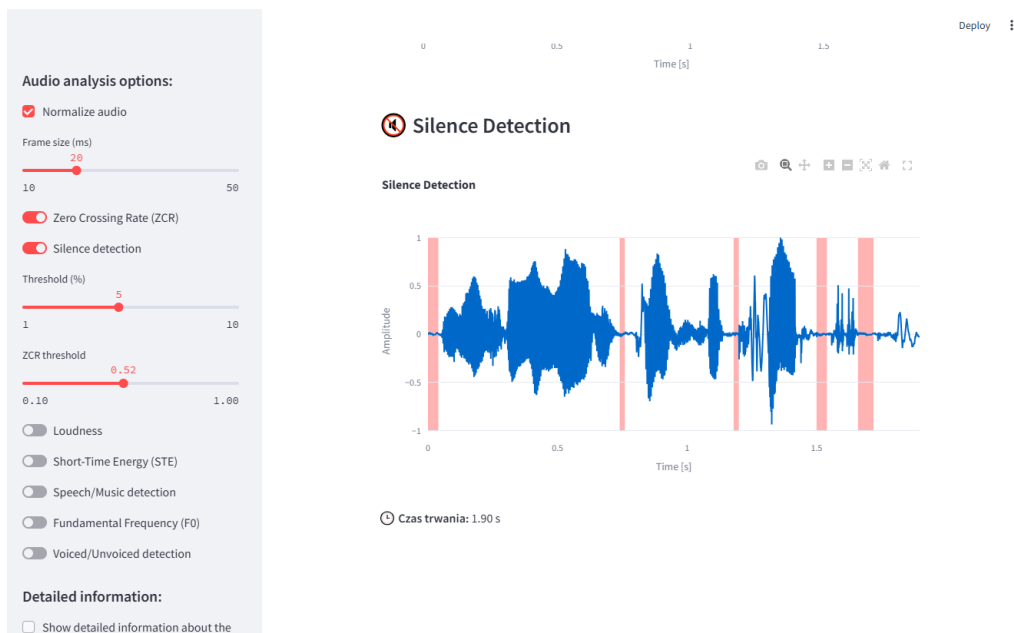
Rysunek 1: Główny panel aplikacji

W głównym panelu aplikacji można załadować i odtworzyć plik audio w formacie .WAV. Po załadowaniu wyświetlany jest wykres amplitudy w funkcji czasu.

Po lewej stronie znajduje się panel, w którym można wybrać wyświetlanie dodatkowych parametrów audio opisanych w raporcie. Dostępny jest również przycisk do normalizacji dźwięku oraz suwak umożliwiający wybór długości ramki w milisekundach.



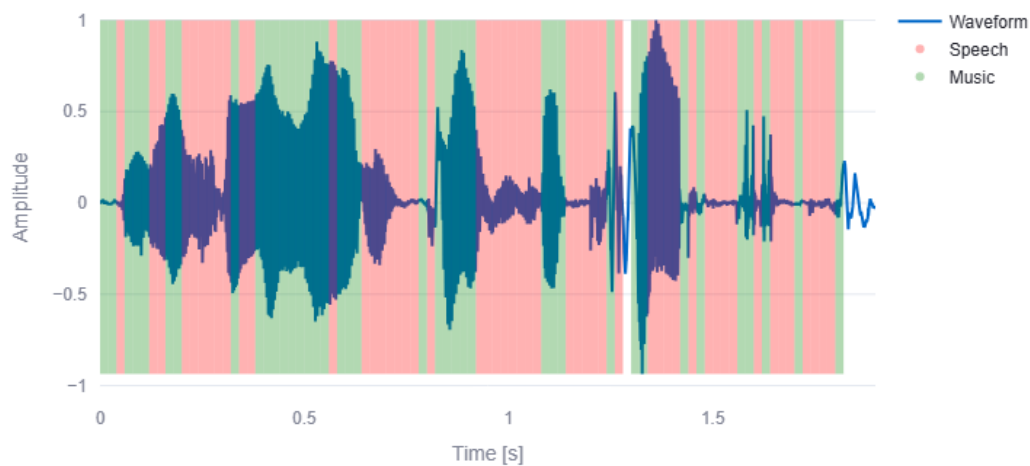
Rysunek 2: Wykrywanie ciszy 1



Rysunek 3: Wykrywanie ciszy 2

Oto wyniki wykrywania fragmentów ciszy zastosowane na dwóch różnych progach parametru ZCR. Widać jak mała zmiana tego parametru wpływa na wychwytywanie trochę innych przedziałów.

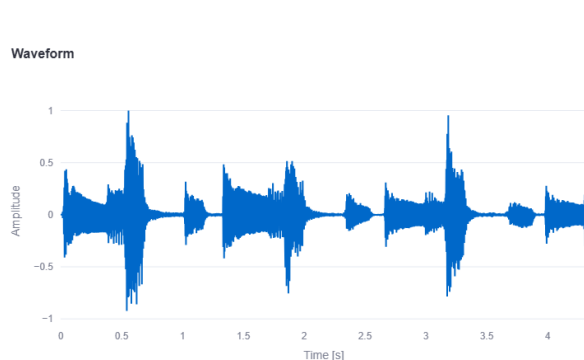
Audio Waveform with Speech/Music Segmentation



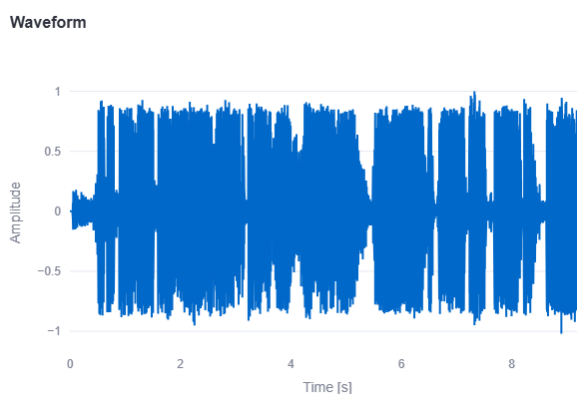
Rysunek 4: Wykrywanie mowy/muzyki

Większość sygnału jest poprawnie zaklasyfikowana jako speech czyli mowa ale widać, że niektóre, krótkie fragmenty błędnie klasyfikujemy jako muzyka.

### 3.2 Porównanie głos ludzki (radio) i muzyka (gitara)



Rysunek 5: Muzyka

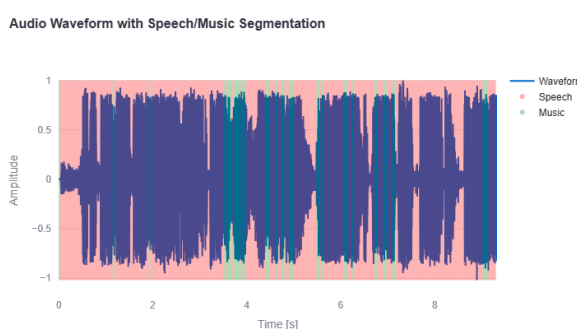


Rysunek 6: Mowa

Rysunek 7: Porównanie wykresów amplitudy





Rysunek 8: Muzyka



Rysunek 9: Mowa


Rysunek 10: Wykrywanie mowy/muzyki



Detailed information about the audio clip 

| Volume Standard Deviation (VSTD) | Volume Dynamic Range (VDR) | Volume Undulation (VU) | Energy Entropy |        |
|----------------------------------|----------------------------|------------------------|----------------|--------|
| 0                                | 0.2247                     | 0.5818                 | 0.0329         | 0.4063 |

Rysunek 11: Muzyka



Detailed information about the audio clip

Volume Standard Deviation (VSTD)

Volume Dynamic Range (VDR)

Volume Undulation (VU)

Energy Entropy

0

0.1004

0.3358

0.0668

0.9001

Rysunek 12: Mowa

Rysunek 13: Porównanie prametrów na poziomie całego klipu

Z analizy wykresów wynika, że w obu przypadkach mowa została poprawnie sklasyfikowana jako mowa, a muzyka jako muzyka. Oznacza to, że model w miarę skutecznie rozróżnia te dwa typy sygnałów dźwiękowych.

Muzykę charakteryzuje znacznie większa dynamika dźwięku (VDR) oraz wyższa wartość odchylenia standardowego energii (VSTD), co wskazuje na większą zmienność amplitudy i intensywności dźwięku w czasie.

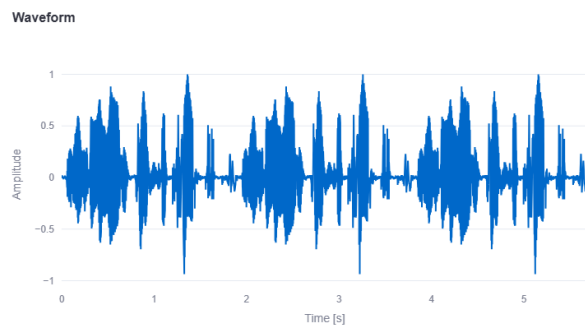
Z kolei głos męski wykazuje wyższą wartość undulacji dźwięku, czyli falistości sygnału. Dodatkowo charakteryzuje się wyższą entropią energii.

Podsumowując, muzyka jest bardziej dynamiczna i zmienna pod względem energii.

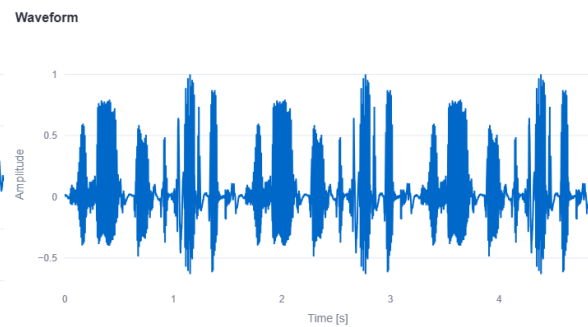


### 3.3 Porównanie głosu męskiego i żeńskiego

Poniżej znajduje się porównanie parametrów dla głosu męskiego i żeńskiego które wypowiadają te same słowa. Oba nagrania zostały znormalizowane.



Rysunek 14: Kobieta

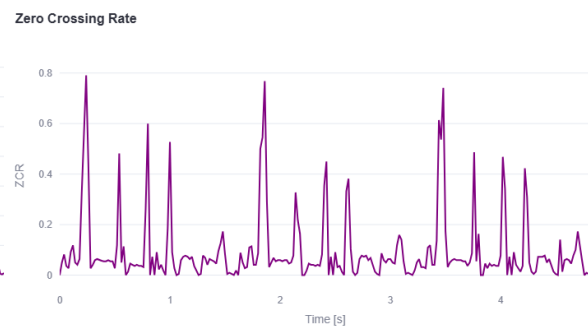


Rysunek 15: Mężczyzna

Rysunek 16: Porównanie wykresów amplitudy

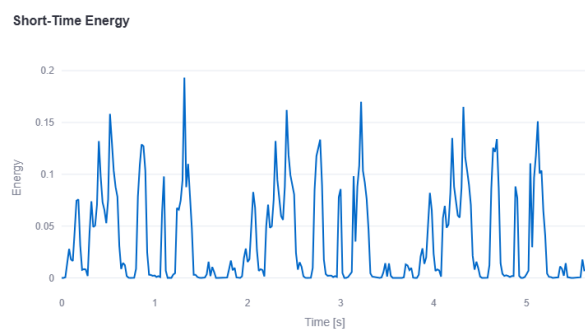


Rysunek 17: Kobieta

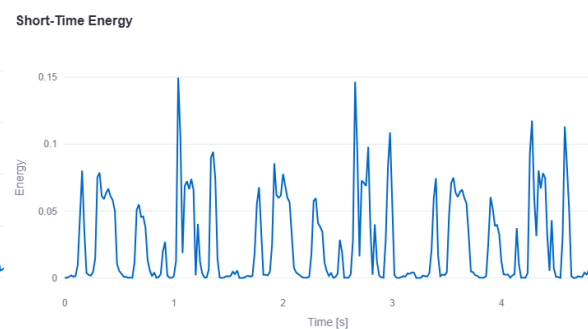


Rysunek 18: Mężczyzna

Rysunek 19: Porównanie wykresów ZCR



Rysunek 20: Kobieta



Rysunek 21: Mężczyzna

Rysunek 22: Porównanie wykresów energii

 Detailed information about the audio clip

| Volume Standard Deviation (VSTD) | Volume Dynamic Range (VDR) | Volume Undulation (VU) | Energy Entropy |
|----------------------------------|----------------------------|------------------------|----------------|
| 0                                | 0.1339                     | 0.3063                 | 0.1735         |
|                                  |                            |                        | 0.7375         |

Rysunek 23: Kobieta

 Detailed information about the audio clip

| Volume Standard Deviation (VSTD) | Volume Dynamic Range (VDR) | Volume Undulation (VU) | Energy Entropy |
|----------------------------------|----------------------------|------------------------|----------------|
| 0                                | 0.0685                     | 0.1491                 | 0.0259         |
|                                  |                            |                        | 0.1422         |

Rysunek 24: Mężczyzna

Rysunek 25: Porównanie parametrów na poziomie całego klipu

Z analizy wykresów wynika, że głos kobiecy charakteryzuje się średnio wyższą energią (STE), co oznacza, że jego amplituda jest większa w porównaniu do głosu męskiego. Dodatkowo zakres energii oraz odchylenie standardowe od średniej (VSTD) są wyższe dla kobiecego głosu, co sugeruje większą zmienność i dynamikę sygnału.

Wartość VDR (różnica między maksymalną a minimalną głośnością) w przypadku głosu kobiecego jest niemal dwukrotnie większa niż w przypadku głosu męskiego. Oznacza to, że głos kobiecy wykazuje większe różnice w natężeniu dźwięku, co może wynikać z większej ekspresyjności oraz szerszego zakresu modulacji.

Dodatkowo, parametr ZCR (Zero-Crossing Rate), czyli liczba przejść sygnału przez oś zerową, również jest wyższy dla kobiecego głosu. Wskazuje to na większą częstotliwość zmian kierunku fali dźwiękowej, co jest zgodne z wyższą częstotliwością podstawową (F0) u kobiet w porównaniu do mężczyzn.

Podsumowując, analiza parametrów akustycznych potwierdza, że głos kobiecy jest zazwyczaj bardziej dynamiczny, charakteryzuje się większą zmiennością energii oraz częstszymi zmianami kierunku sygnału, co przekłada się na jego wyższą tonację i bardziej zróżnicowaną intonację w porównaniu do głosu męskiego.

## 4 Wnioski

- **Szybkość działania** - Dla plików dłuższych niż 3–4 sekundy aplikacja zwalnia, a na obliczenie wszystkich cech audio trzeba poczekać kilka sekund. Z tego powodu najlepiej sprawdza się do analizy bardzo krótkich plików.
- **Częstotliwość F0** - Oba zaimplementowane algorytmy wyznaczania częstotliwości podstawowej nie działają w pełni prawidłowo, dlatego nie zaleca się polegać na ich wynikach.