Dokumentacja

Analiza i przetwarzanie dźwięku

Projekt 1 - cechy sygnału audio w dziedzinie czasu

Bogumiła Okrojek

27 marca 2025

Spis treści

1 Opis aplikacji

Aplikacja została opracowana jako narzędzie do analizy i przetwarzania sygnałów audio. Jej głównym celem jest umożliwienie użytkownikom analizy charakterystyk sygnałów w dziedzinie czasu oraz przekształcanie ich za pomocą różnych funkcji.

1.1 Wykorzystane technologie i struktura

Aplikacja została zrealizowana w języku Python z wykorzystaniem następujących bibliotek i modułów:

- `Tkinter` – jako główny interfejs graficzny użytkownika (GUI),
- `matplotlib` – do wizualizacji danych i wyników analizy na wykresach,
- `sounddevice` – do odtwarzania próbek audio,
- `scipy.io.wavfile` – do wczytywania plików w formacie `.wav`.

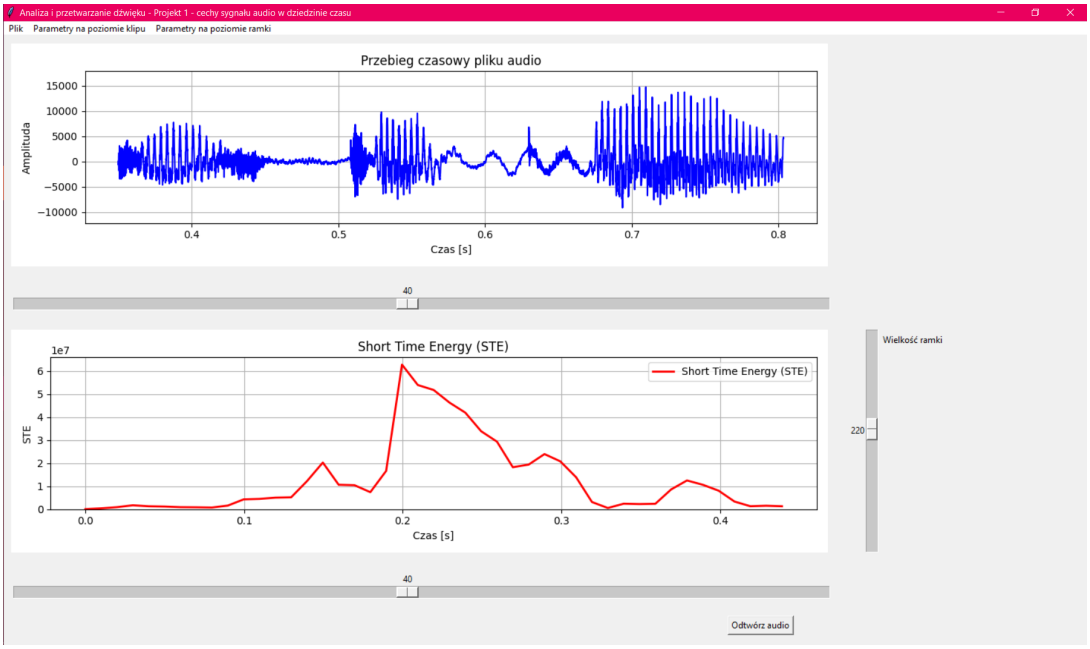
Główne elementy interfejsu widoczne na Rysunek 1 to:

- **Główne okno interfejsu** – zbudowane na bazie klasy `Tk()` z elementami, takimi jak paski menu, suwaki i przyciski do nawigacji,
- **Moduł funkcji analizy sygnału** – zawiera funkcje do analizy sygnałów, takie jak `Short Time Energy (STE)`, `Zero Crossing Rate (ZCR)`, czy wykrywanie tonu podstawowego za pomocą algorytmów autokorelacji,
- **Moduł wizualizacji** – umożliwia wyświetlanie wykresów za pomocą biblioteki `matplotlib`.

1.2 Interfejs użytkownika

Interfejs graficzny został zaprojektowany w sposób intuicyjny i funkcjonalny, oferując:

- Menu podzielone na sekcje, takie jak `Plik`, `Parametry na poziomie klipu` oraz `Parametry na poziomie ramki`, umożliwiające użytkownikom wykonywanie operacji związanych z analizą sygnałów,
- Suwaki umożliwiające interaktywną regulację parametrów, takich jak poziom RMS czy wielkość ramki analizy,
- Panel wizualizacji, który prezentuje wyniki analizy w formie wykresów.



Rysunek 1: GUI aplikacji

1.3 Uzasadnienie wyboru technologii

Python został wybrany ze względu na dostępność bogatych bibliotek do przetwarzania i analizy danych audio oraz możliwości integracji z interfejsem graficznym. Zastosowane technologie, takie jak `matplotlib` i `sounddevice`, zapewniają zarówno efektywność, jak i wszechstronność w realizacji zadań związanych z analizą sygnałów dźwiękowych.

2 Opis metod

2.1 Cechy sygnału audio w dziedzinie czasu na poziomie ramki (Frame-Level)

Analiza cech sygnału audio w dziedzinie czasu na poziomie ramki (*Frame-Level*) polega na podzieleniu sygnału na krótkie, nie nakładające się odcinki czasowe, zwane ramkami, oraz wyznaczeniu dla każdej z nich parametrów opisujących lokalne właściwości sygnału. Takie podejście umożliwia dokładniejszą analizę zmian w sygnale w czasie oraz identyfikację kluczowych cech, które mogą być wykorzystane w takich zastosowaniach jak przetwarzanie mowy, identyfikacja emocji, analiza muzyczna czy detekcja ciszy.

2.1.1 Volume – głośność

Głośność (ang. Volume) jest obliczana jako wartość RMS (Root Mean Square) amplitudy sygnału audio w określonych ramkach czasowych. RMS pozwala na reprezentację poziomu energetycznego

sygnału, uwzględniając zarówno dodatnie, jak i ujemne wartości amplitudy.

Idea i metoda

Operacja głośności jest realizowana na poziomie ramek sygnału o określonym rozmiarze, co umożliwia analizę przebiegu głośności w czasie. Każda ramka jest przetwarzana osobno poprzez:

- Podział danych sygnałowych na ramki o zdefiniowanym rozmiarze,
- Obliczenie RMS dla każdej ramki przy użyciu wzoru:

$$RMS = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{1}$$

  gdzie N oznacza liczbę próbek w ramce, a x_i to amplituda kolejnych próbek.

- Wizualizację wyników RMS na wykresie obok przebiegu czasowego sygnału audio.

W aplikacji dostępna jest możliwość dynamicznej regulacji rozmiaru ramki za pomocą suwaka znajdującego się obok wykresu głośności. Zmiana rozmiaru ramki pozwala użytkownikowi dostosować analizę do charakterystyki sygnału audio.

Wady i ograniczenia

Metoda RMS ma swoje ograniczenia:

- Wyniki zależą od rozmiaru ramki – zbyt małe ramki mogą powodować wysoką zmienność, a zbyt duże mogą utrudnić wychwycenie lokalnych zmian głośności.
- Nie uwzględnia percepcji ludzkiego ucha. Aby to wziąć pod uwagę możnaz astosować inne mierniki głośności, takie jak A-weighted RMS.

2.1.2 STE (Short Time Energy)

Short Time Energy (STE) to miara energii sygnału w krótkich przedziałach czasowych (tzw. ramkach). Jest stosowana do analizy zmienności sygnału w czasie, szczególnie przydatna w identyfikacji charakterystyk dźwięku, takich jak występowanie ciszy, dźwięków mowy czy hałasu.

Idea i metoda

Operacja STE przebiega następująco:

- Dzielimy sygnał audio na ramki o ustalonej wielkości,
- Dla każdej ramki obliczana jest wartość energii STE za pomocą wzoru:

$$STE = \sum_{i=1}^{N} x_i^2 \tag{2}$$

gdzie N oznacza liczbę próbek w ramce, a x_i to wartości amplitudy kolejnych próbek.

- Wyniki są wizualizowane w czasie na wykresie, gdzie STE jest przedstawiana jako funkcja czasu.

W aplikacji użytkownik ma możliwość regulacji rozmiaru ramki za pomocą suwaka, co pozwala na dopasowanie analizy do charakterystyki sygnału audio.

Wady i ograniczenia

Metoda STE ma swoje ograniczenia:

- Wyniki są zależne od wybranego rozmiaru ramki – zbyt małe ramki mogą prowadzić do nadmiernej zmienności wyników, natomiast zbyt duże utrudniają wychwycenie lokalnych zmian w energii sygnału.
- Na wynik może wpływać szum tła, aby go usunąć można wprowadzić filtracji sygnału przed obliczeniem STE

2.1.3 ZCR – Zero Crossing Rate

Zero Crossing Rate (ZCR) to miara opisująca liczbę przejść sygnału przez wartość zero w jednostce czasu lub w określonej ramce sygnałowej. Jest szeroko stosowana w analizie sygnałów audio, szczególnie w celu odróżnienia dźwięków tonalnych (np. mowa) od szumów.

Idea i metoda

ZCR jest obliczane w następujących krokach:

- Podział sygnału audio na ramki o ustalonym rozmiarze (*frame size*),
- Dla każdej ramki obliczana jest liczba przejść sygnału przez zero ze wzoru:

$$ZCR = \frac{1}{2N} \sum_{i=1}^{N-1} |\text{sgn}(x[i]) - \text{sgn}(x[i+1])| \tag{3}$$

gdzie N to liczba próbek w ramce, $x[i]$ to wartość amplitudy i-tej próbki, a $\text{sgn}(x)$ to funkcja znaku zdefiniowana jako:

$$\text{sgn}(x) = \begin{cases} 1, & \text{jeśli } x > 0, \\ 0, & \text{jeśli } x = 0, \\ -1, & \text{jeśli } x < 0. \end{cases}$$

- Wartości ZCR są wizualizowane w czasie na wykresie, umożliwiając analizę dynamiki sygnału.

W aplikacji użytkownik ma możliwość regulacji rozmiaru ramki za pomocą suwaka, co pozwala na dostosowanie analizy do charakterystyki sygnału.

Wady i ograniczenia

Metoda ZCR ma swoje ograniczenia:

- Nie uwzględnia amplitudy sygnału – sygnały o niskiej amplitudzie mogą generować podobne wartości ZCR jak sygnały o dużej amplitudzie.
- ZCR jest mniej efektywny dla sygnałów o niskiej częstotliwości lub małej zmienności znaku.
- Wyniki zależą od rozmiaru ramki – zbyt duże ramki mogą wygładzić zmienność ZCR, a zbyt małe mogą wprowadzić nadmierną fluktuację.

2.1.4 SR – Silent Ratio

Silent Ratio (SR) to miara określająca proporcję czasu, w którym sygnał audio jest klasyfikowany jako „cichy" w danej ramce. Cisza jest definiowana na podstawie dwóch parametrów: głośności (vol_level) oraz współczynnika Zero Crossing Rate (ZCR). Ramka jest uznawana za „cichą", jeśli jej głośność (wyrażona jako RMS) jest mniejsza niż ustalony próg głośności oraz jeśli ZCR jest niższe niż zadany poziom.

Idea i metoda

Operacja Silent Ratio przebiega następująco:

- Podział sygnału audio na ramki o ustalonym rozmiarze (*frame size*),
- Obliczenie wartości RMS i ZCR dla każdej ramki,
- Klasyfikacja ramki jako „cicha" lub „głośna" na podstawie zadanych progów głośności i ZCR według wzoru:

$$SR = \begin{cases} 1, & \text{jeśli RMS} < \text{vol\_level oraz ZCR} < \text{zcr\_level}, \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

- Wizualizacja wartości SR w czasie na wykresie, gdzie 1 oznacza ciszę, a 0 brak ciszy.

Użytkownik może regulować wartości progów głośności, współczynnika ZCR oraz rozmiaru ramki za pomocą suwaków znajdujących się w aplikacji.

Wady i ograniczenia

Metoda SR ma swoje ograniczenia:

- Skuteczność klasyfikacji ciszy zależy od wartości progów – niedopasowane parametry mogą prowadzić do błędnej klasyfikacji. Można zastosować automatyczne dostosowanie wartości progów na podstawie charakterystyki sygnału.
- Cisza jest definiowana wyłącznie na podstawie parametrów głośności i ZCR, bez uwzględnienia spektralnych cech sygnału.
- Zbyt duże ramki mogą utrudnić identyfikację krótkich okresów ciszy, natomiast zbyt małe ramki mogą wprowadzić nadmierne fluktuacje wyników.

2.1.5 Częstotliwość tonu podstawowego F0 (Fundamental Frequency) bazująca na funkcji autokorelacji

Częstotliwość tonu podstawowego (*Fundamental Frequency*, F_0) jest podstawową częstotliwością, która odpowiada za wysokość dźwięku. Jedną z metod jej wyznaczania jest analiza autokorelacji, która pozwala na znalezienie dominującej okresowości w sygnale.

Idea i metoda

Operacja wyznaczania F_0 przebiega według następujących kroków:

- Podział sygnału audio na ramki o ustalonej wielkości (*frame size*),
- Obliczenie funkcji autokorelacji dla każdej ramki za pomocą wzoru:

$$R(lag) = \sum_{i=0}^{N-lag} x[i] \cdot x[i + lag], \tag{4}$$

  gdzie N to liczba próbek w ramce, $x[i]$ to wartość amplitudy i-tej próbki, a lag to opóźnienie próbek.

- Wyszukanie indeksu lag maksymalizującego funkcję autokorelacji (z pominięciem $lag = 0$),
- Wyznaczenie F_0 na podstawie wzoru:

$$F_0 = \frac{fs}{lag}, \tag{5}$$

  gdzie fs to częstotliwość próbkowania sygnału.

- Wizualizacja wartości F_0 w czasie na wykresie.

W aplikacji użytkownik ma możliwość regulacji rozmiaru ramki (*frame size*) za pomocą suwaka, co pozwala na dopasowanie analizy do charakterystyki sygnału audio.

Wady i ograniczenia

Metoda autokorelacji ma swoje ograniczenia:

- Wyniki mogą być nieprecyzyjne w przypadku sygnałów o niskim poziomie energii lub dużej zawartości szumów. Dlatego można najpierw zastosować filtr odszumiający.
- Przy sygnałach wielotonowych (np. muzyka) mogą pojawić się trudności w jednoznacznym określeniu F_0.
- Wartość F_0 jest zależna od rozmiaru ramki – zbyt małe ramki mogą być niewystarczające do analizy okresowości, a zbyt duże mogą rozmywać lokalne zmiany.

2.1.6 Częstotliwość tonu podstawowego F0 (Fundamental Frequency) bazująca na funkcji AMDF (Average Magnitude Difference Function)

Częstotliwość tonu podstawowego (*Fundamental Frequency*, F_0) reprezentuje podstawową częstotliwość drgań sygnału i jest kluczowym parametrem w analizie dźwięku. Metoda AMDF (*Average Magnitude Difference Function*) pozwala na określenie F_0 poprzez analizę różnic między wartościami amplitudy sygnału przesuniętego o różne opóźnienia (*lags*).

Idea i metoda

Obliczanie F_0 z wykorzystaniem AMDF przebiega następująco:

- Sygnał audio jest dzielony na ramki o ustalonej wielkości (*frame size*),
- Dla każdej ramki obliczana jest funkcja AMDF dla różnych wartości *lag* za pomocą wzoru:

$$AMDF(lag) = \frac{1}{N - lag} \sum_{i=1}^{N-lag} |x[i] - x[i + lag]|, \tag{6}$$

  gdzie N to liczba próbek w ramce, $x[i]$ to wartość amplitudy i-tej próbki, a *lag* to opóźnienie.

- Na podstawie funkcji AMDF wyszukiwane jest pierwsze lokalne maksimum, które ogranicza zakres poszukiwania,
- Po pierwszym maksimum określane jest minimum $lag_{\min}$, które służy do wyliczenia F_0 ze wzoru:

$$F_0 = \frac{fs}{lag_{\min}}, \tag{7}$$

gdzie fs to częstotliwość próbkowania sygnału.

- Wyniki F_0 są wizualizowane w czasie na wykresie.

Użytkownik może dostosować rozmiar ramki (*frame size*) za pomocą suwaka w interfejsie aplikacji, aby lepiej dostosować analizę do charakterystyki sygnału.

Wady i ograniczenia

Metoda AMDF ma swoje ograniczenia:

- Wyniki są wrażliwe na szum w sygnale – obecność szumu może wprowadzić fałszywe minima. Można zastosować filtry wstępne.
- Rozdzielczość metody zależy od rozmiaru ramki – zbyt duże ramki mogą zniekształcić wyniki, a zbyt małe mogą być niewystarczające do analizy okresowości,
- Działa najlepiej na sygnałach jednotonowych (np. mowa), ale może mieć trudności z analizą sygnałów wielotonowych (np. muzyka).

2.2 Cechy sygnału audio w dziedzinie czasu na poziomie klipu (Clip-Level)

Analiza sygnału audio w dziedzinie czasu na poziomie klipu (*Clip-Level*) polega na ocenie globalnych właściwości sygnału dla całego klipu lub jego istotnych fragmentów. Zamiast skupiać się na krótkich odcinkach (ramkach), jak w analizie *Frame-Level*, podejście *Clip-Level* umożliwia uzyskanie uśrednionych lub skumulowanych informacji o całym sygnale. Tego rodzaju analiza jest szczególnie przydatna w zadaniach takich jak klasyfikacja dźwięków, rozpoznawanie gatunków muzycznych, ocena jakości nagrań czy detekcja mowy.

2.2.1 Bazujące na głośności: VSTD, VDR, VU

Cechy bazujące na głośności pozwalają na ocenę zmienności, zakresu i równomierności głośności sygnału audio w całym klipie.

Idea i metoda

W analizie wykorzystywane są trzy parametry:

- **VSTD (Volume Standardized Deviation)** – opisuje znormalizowane odchylenie standardowe głośności. Wartość $VSTD$ jest obliczana jako stosunek odchylenia standardowego głośności σ_{vol} do maksymalnej wartości głośności $\max_{vol}$:

$$VSTD = \frac{\sigma_{vol}}{\max_{vol}}. \tag{8}$$

- **VDR (Volume Dynamic Range)** – określa zakres dynamiki głośności w klipie i jest zdefiniowany jako stosunek różnicy maksymalnej i minimalnej wartości głośności do maksymalnej głośności:
$$VDR = \frac{\max_{vol} - \min_{vol}}{\max_{vol}}. \tag{9}$$

- **VU (Volume Uniformity)** – ocenia różnice między lokalnymi pikami i dolinami głośności w sygnale. Dla każdego okna analizy VU obliczane jest jako suma wartości bezwzględnych różnic między pikami p_i a dolinami v_i:
$$VU = \sum_{i=1}^{P} |p_i - v_i|, \tag{10}$$
gdzie P to liczba par pików i dolin w analizowanym oknie.

Metoda wyznaczania parametrów VSTD, VDR i VU obejmuje następujące kroki:

- Normalizacja sygnału audio w celu zapewnienia spójności danych,
- Podział sygnału na ramki o ustalonym rozmiarze (*frame size*),
- Obliczanie wartości RMS (Root Mean Square) dla każdej ramki w celu wyznaczenia głośności,
- Wyznaczanie wartości $VSTD$, VDR i VU na podstawie wartości RMS w całym klipie oraz w kolejnych oknach 1-sekundowych.,
- Prezentacja wyników zarówno dla całego klipu, jak i dla poszczególnych okien czasowych.

Wady i ograniczenia

Metoda bazująca na głośności posiada następujące ograniczenia:

- Obliczenia są podatne na szum – obecność zakłóceń w sygnale może wpływać na wartości $VSTD, VDR$ i VU,
- Wyniki zależą od rozmiaru ramki i okien analizy – nieodpowiedni dobór tych parametrów może prowadzić do zniekształcenia wyników,
- Wysoka dynamika sygnału (np. w muzyce) może generować złożone wzorce pików i dolin, co utrudnia analizę VU.

2.2.2 Bazujące na energii: LSTER, Energy Entropy

Parametry bazujące na energii dostarczają informacji o globalnych oraz lokalnych właściwościach energetycznych sygnału audio, które są kluczowe w rozpoznawaniu mowy, identyfikacji ciszy czy analizy emocji w dźwięku.

Idea i metoda

W tej metodzie zastosowano dwa podstawowe parametry:

- **LSTER (Low Short-Time Energy Ratio)** – określa stosunek liczby ramek o niskiej energii do całkowitej liczby ramek w sygnale, w danym oknie czasowym. Ramki o niskiej energii definiowane są jako te, które mają wartość energii mniejszą niż 50% średniej energii w danym oknie czasowym:

$$LSTER = \frac{\text{Liczba ramek o } STE < 0.5 \cdot \overline{STE}}{\text{Liczba wszystkich ramek w oknie}}. \tag{11}$$

- **Energy Entropy** – mierzy nierównomierność dystrybucji energii w podziałach ramki na segmenty. Wyższa entropia wskazuje na bardziej równomierne rozłożenie energii w ramce, natomiast niższa entropia sugeruje koncentrację energii w określonych fragmentach. Entropia energii jest obliczana jako:

$$EE = -\sum_{i=1}^{N} p_i \cdot \log_2(p_i), \tag{12}$$

  gdzie p_i to znormalizowana energia i-tego segmentu, a N to liczba segmentów w ramce.

Parametry LSTER i Energy Entropy są obliczane według następującego procesu:

- Normalizacja sygnału audio w celu ujednolicenia wartości amplitudy,
- Podział sygnału na ramki o ustalonym rozmiarze (*frame size*),
- Obliczenie energii (Short-Time Energy, STE) dla każdej ramki,
- Wyznaczanie $LSTER$ jako stosunku liczby ramek o niskiej energii do całkowitej liczby ramek w poszczególnych oknach czasowych,
- Podział ramek na mniejsze segmenty, obliczenie energii w każdym segmencie oraz wyznaczenie Energy Entropy na podstawie znormalizowanego rozkładu energii w segmentach,
- Prezentacja wyników dla okien czasowych (np. 1-sekundowych) oraz dla całego klipu audio.

Wady i ograniczenia

Metoda posiada kilka ograniczeń:

- Obliczenia są podatne na obecność szumów w sygnale, co może zafałszować wartości $LSTER$ i entropii energii,
- Wyniki zależą od rozmiaru ramek i segmentów – niewłaściwy dobór tych parametrów może prowadzić do zniekształcenia wyników,
- Entropia energii zakłada równy podział segmentów w ramach, co może nie być idealne w sygnałach o nieregularnej strukturze.

2.2.3 Bazujące na energii: LSTER, Energy Entropy

Parametry bazujące na energii umożliwiają ocenę globalnych i lokalnych właściwości energetycznych sygnału audio, co jest kluczowe w takich zastosowaniach jak analiza mowy, detekcja ciszy czy klasyfikacja dźwięków.

Idea i metoda

W tej metodzie zastosowano dwa podstawowe wskaźniki:

- **LSTER (Low Short-Time Energy Ratio)** – określa stosunek liczby ramek o niskiej energii do całkowitej liczby ramek w sygnale. Ramki o niskiej energii to te, których energia jest mniejsza niż 50% średniej energii w danym oknie czasowym:

$$LSTER = \frac{\text{Liczba ramek o } STE < 0.5 \cdot \overline{STE}}{\text{Liczba wszystkich ramek w oknie}}. \tag{13}$$

- **Energy Entropy** – mierzy nierównomierność dystrybucji energii w segmentach ramki. Wyższa entropia wskazuje na równomierne rozłożenie energii, a niższa entropia na jej koncentrację w określonych obszarach. Entropia energii jest obliczana jako:

$$EE = -\sum_{i=1}^{N} p_i \cdot \log_2(p_i), \tag{14}$$

  gdzie p_i to znormalizowana energia i-tego segmentu, a N liczba segmentów w ramce.

Proces obliczania

Parametry LSTER i Energy Entropy są wyznaczane w następujący sposób:

- Normalizacja sygnału audio - amplituda sygnału jest dzielona przez jego maksymalną wartość.

- Podział sygnału na ramki o ustalonej długości.
- Obliczanie energii (STE) każdej ramki jako sumy kwadratów wartości amplitud znormalizowanych.
- Obliczanie LSTER
- Obliczanie Energy Entropy
- Wyniki LSTER i Energy Entropy są prezentowane zarówno dla całego klipu, jak i dla poszczególnych okien czasowych.

Wady i ograniczenia

Metoda bazująca na energii posiada następujące ograniczenia:

- Wyniki mogą być zniekształcone przez obecność szumów, które wpływają na obliczenia energii.
- Efektywność analizy zależy od rozmiaru ramek i segmentów – niewłaściwy dobór parametrów może prowadzić do błędnych wyników.
- Nierównomierny podział energii w segmentach może prowadzić do błędnej interpretacji entropii.

2.2.4 Bazujące na ZCR: ZSTD, HZCRR

Idea i metoda

Parametry bazujące na współczynniku Zero Crossing Rate (ZCR) pozwalają na szczegółową analizę zmienności sygnału w dziedzinie czasu. Są one szczególnie przydatne w rozróżnianiu typów dźwięków, takich jak szumy czy mowa. W tej metodzie zastosowano dwa kluczowe wskaźniki:

- **ZSTD (Zero Crossing Rate Standard Deviation)** – opisuje zróżnicowanie ZCR w sygnale w określonych oknach czasowych. Odchylenie standardowe ZCR (ZSTD) jest obliczane według wzoru:

$$ZSTD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(zcr_i - \overline{zcr})^2}, \tag{15}$$

gdzie N to liczba ramek w oknie, zcr_i to wartość ZCR dla i-tej ramki, a $\overline{zcr}$ oznacza średnią wartość ZCR w oknie.

- **HZCRR (High Zero Crossing Rate Ratio)** – określa proporcję ramek, których wartości ZCR są równe lub większe niż $1.5 \cdot \overline{\text{ZCR}}$ w danym oknie czasowym. HZCRR jest obliczane według wzoru:

$$HZCRR = \frac{1}{2N} \sum_{n=1}^{N} [\text{sgn}(ZCR_n - 1.5 \cdot \overline{ZCR}) + 1] \tag{16}$$

  gdzie:

  - N – liczba ramek w oknie czasowym,
  - ZCR_n – wartość Zero Crossing Rate dla n-tej ramki,
  - $\overline{ZCR}$ – średnia wartość ZCR w oknie czasowym,
  - $\text{sgn}(x)$ – funkcja signum

Proces obliczania

Parametry ZSTD i HZCRR są wyznaczane w następujący sposób:

- Podział sygnału na ramki o określonym rozmiarze (*frame size*),
- Obliczenie współczynnika ZCR dla każdej ramki jako liczby przejść sygnału przez wartość zero,
- Grupowanie ramek w okna czasowe (np. 1-sekundowe),
- Wyznaczenie średniego ZCR w oknie, a następnie obliczenie odchylenia standardowego ZCR (ZSTD),
- Obliczenie (HZCRR),
- Prezentacja wyników zarówno dla całego klipu, jak i dla poszczególnych okien czasowych.

Wady i ograniczenia

Metoda ta ma pewne ograniczenia:

- Obliczenia są podatne na zakłócenia związane z szumami, które mogą wpłynąć na wartości ZCR.
- Wyniki zależą od rozmiaru ramek i okien czasowych – niewłaściwy dobór tych parametrów może prowadzić do błędnych interpretacji wyników.
- Wysoka dynamika sygnału może generować złożone wzorce ZCR, utrudniając analizę i klasyfikację dźwięku.

2.3 Analiza sygnału

2.3.1 Detekcja ciszy

Detekcja ciszy opiera się na koncepcji Silent Ratio (SR), które klasyfikuje fragmenty sygnału jako „ciche" na podstawie wartości współczynnika Zero Crossing Rate (ZCR) oraz głośności (VOL).

Opis metody

O ile sekcja dotycząca SR skupia się na samych zasadach klasyfikacji, niniejsza część podkreśla różnorodne elementy wizualizacji wyników oraz dodatkowe szczegóły związane z analizą dynamiczną sygnału.

W tej sekcji wyniki detekcji ciszy są przedstawione na wykresie, który łączy kilka kluczowych elementów:

- **Przebieg czasowy sygnału audio (niebieski wykres)**: przedstawia amplitudę sygnału w funkcji czasu, umożliwiając zrozumienie jego ogólnej struktury.
- **Wartości ZCR (zielony wykres) i VOL (żółty wykres)**: pokazują zmienność tych parametrów w czasie, co pozwala ocenić wpływ zmian sygnału na klasyfikację.
- **Zaznaczenie ciszy (szare wypełnienie)**: obszary ciszy są wizualnie wyróżnione na wykresie za pomocą szarego tła, co ułatwia identyfikację ich położenia w czasie.
- **Silent Ratio (czerwony wykres)**: przedstawia wartość SR w danym momencie, gdzie 1 oznacza ciszę, a 0 brak ciszy.

Co nowego wnosi wizualizacja?

Dzięki tej metodzie możliwe jest:

- Wizualne porównanie fragmentów ciszy z przebiegiem czasowym sygnału, co pozwala zidentyfikować nie tylko momenty ciszy, ale także charakterystykę sygnału w czasie.
- Ocena dynamiki ZCR i VOL, które mogą wskazywać, jakie zmiany sygnału wyzwalają klasyfikację ciszy.
- Eksperymentowanie z różnymi progami głośności i ZCR oraz natychmiastowe obserwowanie ich wpływu na klasyfikację ciszy w wizualizacji.

Opis wyników dla różnych parametrów i sygnałów

- **Przypadek 1: Nagranie muzyki z przerwami**

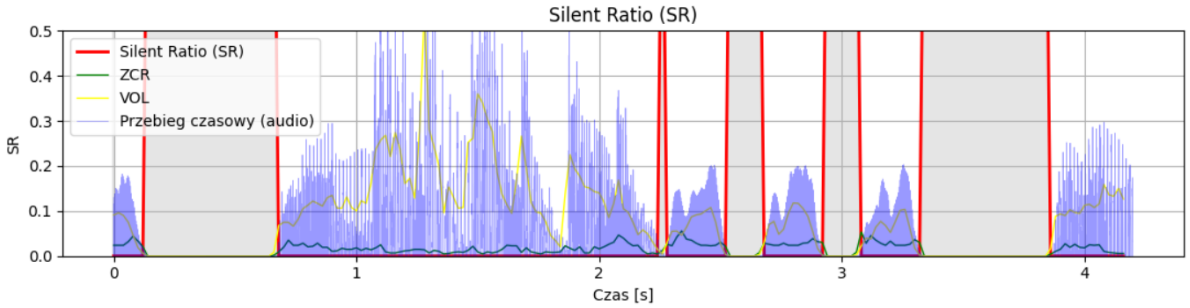
  Dla sygnału muzycznego fragmenty ciszy pojawiają się w pauzach między frazami muzycznymi. Użyte parametry to:

– Próg głośności: vol_level = 0.01,

– Próg ZCR: zcr_level = 0.02,

– Rozmiar ramki: 882 próbek.

Na wykresie przedstawionym na Rysunek 2 widać, że cisza została poprawnie wykryta i zaznaczona w miejscach pomiędzy fragmentami melodii. Algorytm poprawnie identyfikował pauzy przy odpowiednich wartościach progów.



Rysunek 2: Wykres z zaznaczonymi fragmentami ciszy dla nagrania z muzyką

- **Przypadek 2: Nagranie z muzyką, fragmentem ciszy i mową**

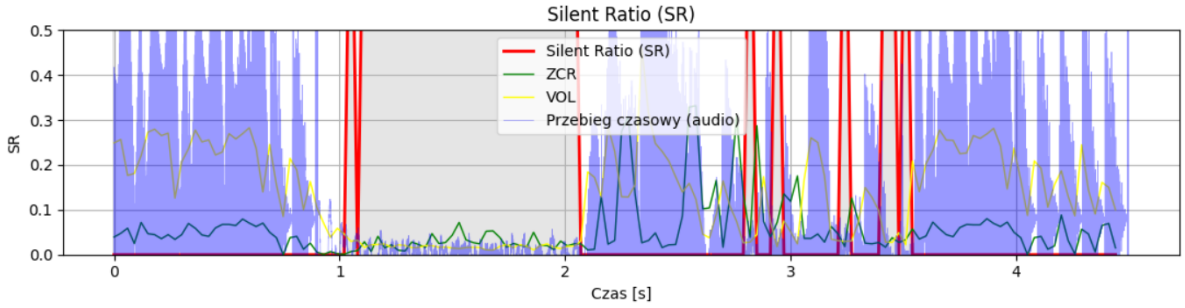
W tym nagraniu występuje fragment ciszy, krótkie przerwy w muzyce oraz przerwy pomiędzy słowami w mowie. Ustawienia parametrów to:

– Próg głośności: vol_level = 0.03,

– Próg ZCR: zcr_level = 0.08,

– Rozmiar ramki: 1323 próbek.

Wykres przedstawiony na Rysunek 3 pokazuje, że cisza została poprawnie oznaczona zarówno w fragmentach ciszy w nagraniu, jak i w krótkich przerwach muzycznych oraz w pauzach pomiędzy wyrazami w mowie.



Rysunek 3: Wykres z zaznaczonymi fragmentami ciszy dla nagrania z muzyką, ciszą i mową

- **Przypadek 3: Nagranie mowy z hałasem tła**

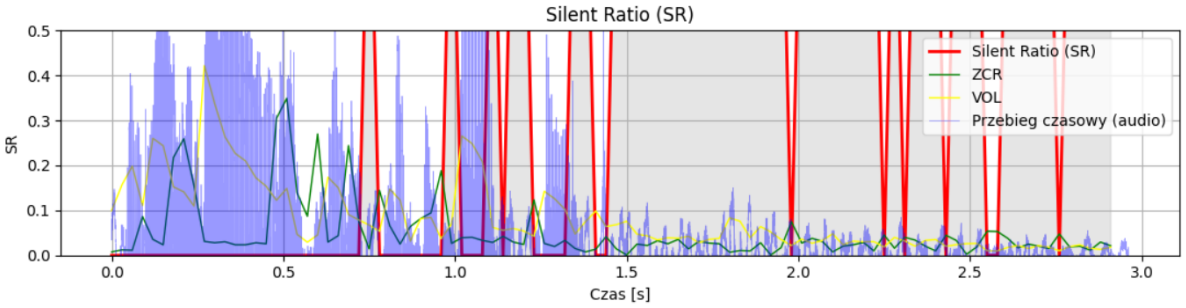
  W ostatnim przykładzie analizowany był sygnał mowy z obecnym hałasem w tle. Parametry analizy to:

  – Próg głośności: vol_level = 0.09,

  – Próg ZCR: zcr_level = 0.04,

  – Rozmiar ramki: 1323 próbek.

  Na wykresie przedstawionym na Rysunek 4 od 1.5 sekundy obecna jest jedynie cisza z szumem, ale nie cały ten zakres został poprawnie oznaczony. Hałas tła wpłynął na obniżenie skuteczności detekcji.



Rysunek 4: Wykres z zaznaczonymi fragmentami ciszy dla nagrania mowy z hałasem tła

Wady i ograniczenia

Detekcja ciszy oparta na SR podlega pewnym ograniczeniom:

- Ograniczona skuteczność w sygnałach zawierających niski poziom szumów tła, które mogą być zaklasyfikowane jako cisza.
- Wpływ rozmiaru ramki (*frame size*) na dokładność – większe ramki mogą maskować krótkie fragmenty ciszy, a mniejsze ramki mogą generować niepotrzebne fluktuacje.
- Wysoka czułość na dobór progów ZCR i VOL, które muszą być dostosowane do charakterystyki analizowanego sygnału.

2.3.2 Określanie dźwięczności / bezdźwięczności

Opis metody

Określanie dźwięczności i bezdźwięczności sygnału opiera się na analizie częstotliwości tonu podstawowego (F_0) wyznaczanego za pomocą funkcji autokorelacji. Podczas gdy sekcja dotycząca

F_0 skupia się na ogólnej charakterystyce tonu podstawowego, niniejsza część przedstawia, w jaki sposób wartości F_0 są wykorzystywane do klasyfikacji sygnału jako dźwięczny lub bezdźwięczny.

Proces klasyfikacji

Operacja przebiega według następujących kroków:

- Wyznaczenie F_0 dla każdej ramki sygnału na podstawie funkcji autokorelacji (zgodnie z metodą opisaną w sekcji F_0).
- Ustawienie wartości progowej (*cut-off*), poniżej której ramka jest klasyfikowana jako bezdźwięczna.
- Klasyfikacja ramki jako „dźwięczna" lub „bezdźwięczna" na podstawie wartości F_0:

$$\text{Dźwięczność} = \begin{cases} 1, & \text{jeśli } F_0 \geq \text{cut-off}, \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$$

- Przeskalowanie wartości F_0 i sygnału audio do zakresu $[0, 1]$ w celu łatwiejszej wizualizacji na wykresie.
- Graficzne przedstawienie wyników: wartość F_0, fragmenty sygnału sklasyfikowane jako dźwięczne lub bezdźwięczne oraz wartość progowa (*cut-off*).

Wizualizacja wyników

Wyniki analizy są przedstawiane na wykresie obejmującym:

- Przeskalowany przebieg sygnału audio (niebieski wykres) – umożliwia zrozumienie ogólnej struktury sygnału w czasie.
- Przeskalowane wartości F_0 (czerwony wykres) – pokazują zmienność tonu podstawowego w analizowanym nagraniu.
- Linię progową (*cut-off*) – zaznaczoną jako niebieską linię przerywaną, która wskazuje granicę między dźwięcznością a bezdźwięcznością.
- Fragmenty sygnału sklasyfikowane jako bezdźwięczne, gdzie F_0 przyjmuje wartości niższe niż wartość progowa.

Opis wyników dla różnych parametrów i sygnałów

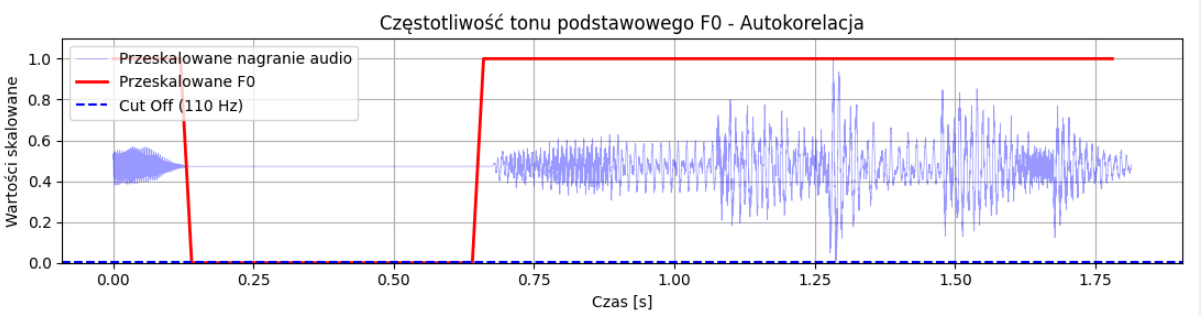
- **Przypadek 1: Nagranie muzyki z przerwami**

  Dla sygnału muzycznego, gdzie fragmenty ciszy występują w pauzach między frazami, zastosowano następujące ustawienia parametrów:

– Próg bezdźwięczności: 110 Hz,

– Rozmiar ramki: 882 próbek.

Na wykresie przedstawionym na Rysunek 5 widać, że algorytm poprawnie identyfikował pauzy między fragmentami melodii i zaznaczał je jako bezdźwięczne. Wysoki próg F_0 (110 Hz) umożliwił skuteczne wykluczenie tonów muzycznych o wyższych częstotliwościach z klasyfikacji jako bezdźwięczne.



Rysunek 5: Wykres z zaznaczonymi fragmentami bezdźwięcznymi dla nagrania z muzyką

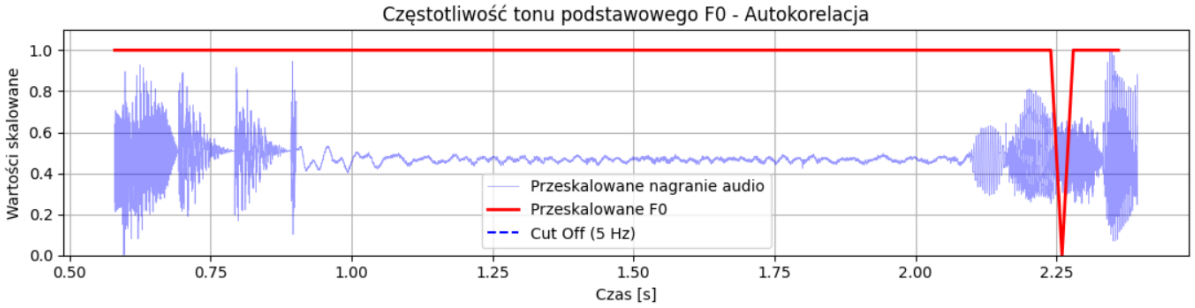
- **Przypadek 2: Nagranie z muzyką, fragmentem ciszy i mową**

  Nagranie zawiera fragment ciszy (nieidealnej), krótkie przerwy w muzyce oraz mowę. Parametry analizy były następujące:

  – Próg bezdźwięczności: 5 Hz,

  – Rozmiar ramki: 882 próbek.

  Na wykresie przedstawionym na Rysunek 6 można zauważyć, że algorytm nie rozpoznał fragmentów ciszy jako bezdźwięczne. Fragmenty mowy zostały jednak poprawnie zaznaczone jako bezdźwięczne, co pokazuje skuteczność algorytmu w identyfikacji dźwięcznych fragmentów nawet przy niewielkich zakłóceniach sygnału.



Rysunek 6: Wykres z zaznaczonymi fragmentami bezdźwięcznymi dla nagrania z muzyką, ciszą i mową

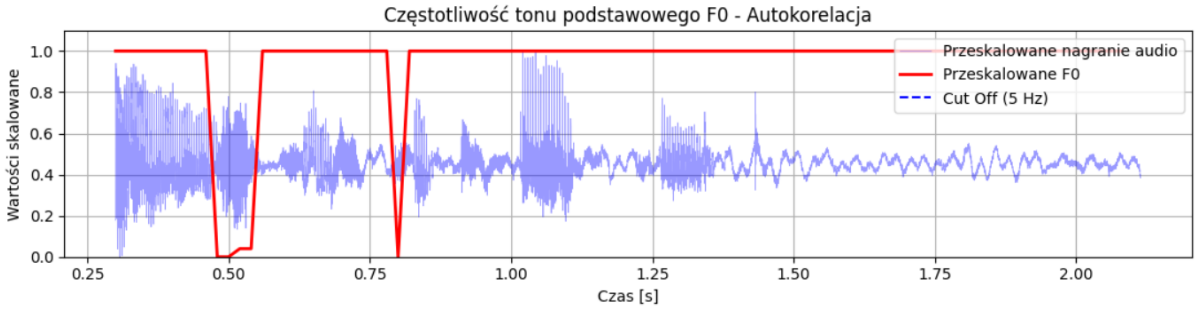
- **Przypadek 3: Nagranie mowy z hałasem tła**

  Dla nagrania mowy w obecności hałasu tła zastosowano następujące parametry:

  – Próg bezdźwięczności: 5 Hz,

  – Rozmiar ramki: 882 próbek.

  Jak pokazano na wykresie w Rysunek 7, fragmenty szumu po 1.5 sekundzie nie zostały zaklasyfikowane jako bezdźwięczne, co wskazuje, że szum został uznany za dźwięczny. Natomiast fragmenty mowy, pomimo zakłóceń hałasem, były poprawnie zaznaczane jako bezdźwięczne.



Rysunek 7: Wykres z zaznaczonymi fragmentami bezdźwięcznymi dla nagrania mowy z hałasem tła

Zastosowanie i wnioski

Metoda określania dźwięczności i bezdźwięczności znajduje zastosowanie w:

- Analizie mowy: identyfikacja spółgłosek bezdźwięcznych i dźwięcznych.
- Analizie sygnałów muzycznych: klasyfikacja tonów i pauz w muzyce.
- Detekcji ciszy w sygnałach złożonych: rozróżnienie ciszy od tonów o niskiej częstotliwości.

Możliwość dostosowania wartości progowej przez użytkownika umożliwia precyzyjne dopasowanie metody do różnych typów sygnałów.

2.3.3 Określanie fragmentów muzyki i mowy

Opis metody

Określanie, czy dany fragment nagrania zawiera muzykę, czy mowę, opiera się na analizie parametrów LSTER (Loudness Spectral Temporal Energy Ratio) oraz ZSTD (Zero Crossing Rate Standard Deviation). Podczas gdy wcześniejsze sekcje raportu skupiały się na szczegółowym opisie tych parametrów, niniejsza część przedstawia, w jaki sposób wykorzystano ich wartości do klasyfikacji sygnału.

Proces klasyfikacji

Operacja przebiega według następujących kroków:

- Obliczenie wartości LSTER i ZSTD dla każdego okna sygnału, zgodnie z metodami przedstawionymi w sekcjach opisujących te parametry.
- Ustawienie progów rozdzielających wartości charakterystyczne dla mowy i muzyki:

$$\text{Klasyfikacja} = \begin{cases} \text{Mowa}, & \text{jeśli LSTER} \in [0.2, 0.5] \text{ i ZSTD} \in [0.1, 0.3], \\ \text{Muzyka}, & \text{jeśli LSTER} > 0.6 \text{ lub ZSTD} > 0.4. \end{cases}$$

- Dla każdego okna przypisanie etykiety „muzyka" lub „mowa" w zależności od spełnionych warunków progowych.
- Graficzne przedstawienie wyników: wartości LSTER i ZSTD.

Wizualizacja wyników

Wyniki analizy są przedstawiane na wykresie, który obejmuje:

- Przeskalowany przebieg sygnału audio (niebieski wykres) – pokazuje ogólną strukturę nagrania w czasie.
- Wartości LSTER (zielony wykres) – ilustrują dynamikę energii widmowej w czasie.
- Wartości ZSTD (czerwony wykres) – prezentują zmienność częstotliwości przejść przez wartość zerową.

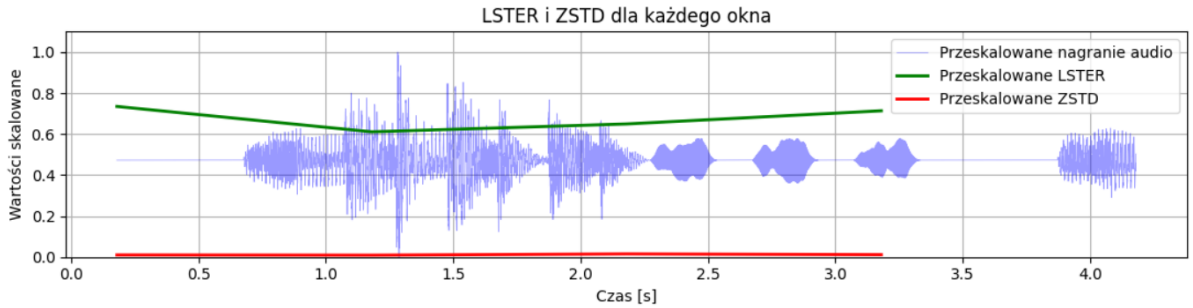
Przypadek 1: Nagranie z muzyką Na Rysunek 8 widać, że wartości LSTER są wysokie, podczas gdy ZSTD przyjmuje wartości niewielkie. Taki rozkład wyników nie pozwala jednoznacznie wskazać charakterystyki muzycznej sygnału. Na wykresie wartości LSTER dominują nad progami dla mowy, co sugeruje, że muzyka charakteryzuje się dużą zmiennością tonalną, której analiza może być trudna do rozróżnienia na podstawie tych parametrów.

Przypadek 2: Nagranie z mową i muzyką

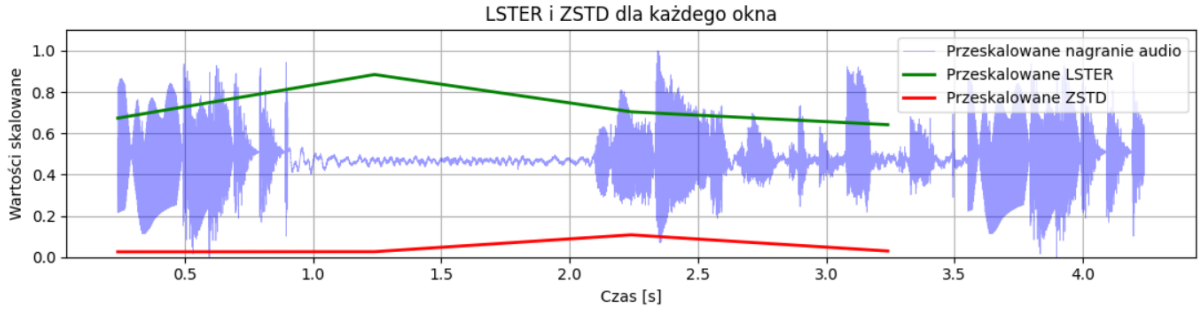
Na Rysunek 9 widać, że wartości LSTER osiągają najwyższe wartości podczas ciszy, co może sugerować, że ta miara jest bardziej czuła na zmiany w tle. Z kolei wartości ZSTD są wyraźnie wyższe dla fragmentów z mową, co potwierdza, że ta miara dobrze wychwytuje obszary z wyraźnymi zmianami akustycznymi, typowymi dla mowy.

Przypadek 3: Nagranie z tłem szumowym

Na wykresie Rysunek 10 widać, że ZSTD przyjmuje najwyższe wartości dla fragmentów z mową, a jego wartość spada w obecności szumów. Podobny trend występuje w przypadku

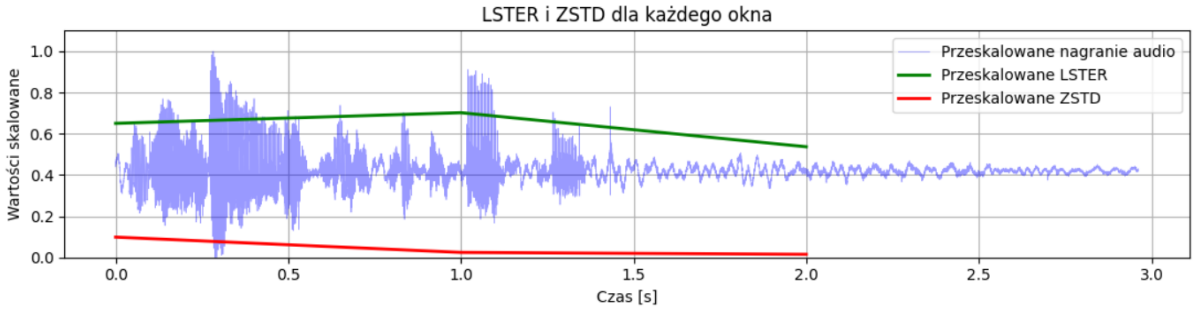


Rysunek 8: Wykres z widocznymi fragmentami muzycznymi w nagraniu zawierającym tylko muzykę



Rysunek 9: Wykres z widocznymi fragmentami mowy i muzyki w nagraniu zawierającym zarówno muzykę, ciszę, jak i mowę

LSTER, gdzie wartości również są wyższe dla mowy, a niższe dla szumu. Taki wynik może wskazywać, że zarówno ZSTD, jak i LSTER są czułe na różnice w dynamice sygnału, a ich wartość zmienia się w zależności od charakterystyki tła.



Rysunek 10: Wykres z widocznymi fragmentami mowy z tłem szumowym

Zastosowanie i wnioski

Metoda klasyfikacji fragmentów nagrania jako muzyki lub mowy ma szerokie zastosowanie w:

- Systemach rozpoznawania mowy,
- Analizie nagrań multimedialnych,

- Automatycznym indeksowaniu treści audio.

Ustalanie progów dla parametrów LSTER i ZSTD pozwala na dostosowanie metody do różnych rodzajów sygnałów, co zwiększa jej skuteczność w praktycznych zastosowaniach.

2.4 Inne

2.4.1 Wczytanie nagrania z pliku

Wczytanie pliku audio jest pierwszym krokiem w analizie sygnału dźwiękowego.

Idea i metoda

Operacja polega na załadowaniu danych z pliku w formacie WAV oraz wyznaczeniu kluczowych parametrów, takich jak częstotliwość próbkowania oraz wartości amplitudy próbek. Dane te są następnie wizualizowane w formie przebiegu czasowego. Operacja ta umożliwia użytkownikowi bezpośredni podgląd sygnału i jego charakterystyki przed dalszą analizą.

Proces wczytywania danych jest realizowany przez następujące kroki:

- Wybranie pliku przez użytkownika za pomocą interaktywnego okna dialogowego,
- Wczytanie danych audio z wybranego pliku,
- Konwersja sygnału stereo na mono, jeśli plik zawiera dane wielokanałowe (średnia wartości amplitud dla obu kanałów),
- Wyznaczenie długości sygnału (liczba próbek) oraz częstotliwości próbkowania,
- Wyświetlenie sygnału w formie przebiegu czasowego na wykresie. Oś X przedstawia czas w sekundach, a oś Y wartość amplitudy. Zakres przewijania sygnału jest dynamicznie dostosowywany do długości danych audio za pomocą suwaków.

Wady i ograniczenia

Operacja wczytywania danych audio ma swoje ograniczenia:

- Format obsługiwany przez funkcję ogranicza się do plików WAV – pliki w innych formatach, takich jak MP3, wymagają dodatkowej konwersji.
- Konwersja sygnału stereo na mono powoduje utratę informacji o przestrzennej strukturze dźwięku.

2.4.2 Odtworzenie nagrania

Odtworzenie nagrania audio jest podstawową operacją umożliwiającą użytkownikowi ocenę sygnału dźwiękowego w celu przeprowadzenia dalszej analizy.

Idea i metoda

Proces ten polega na wykorzystaniu próbek sygnału audio oraz odpowiedniej częstotliwości próbkowania w celu odtworzenia dźwięku przez urządzenie wyjściowe, takie jak głośniki lub słuchawki.

Operacja odtwarzania nagrania realizowana jest w następujących krokach:

- Sprawdzenie, czy dane audio i częstotliwość próbkowania zostały poprawnie wczytane. Jeśli dane są niedostępne, operacja zostaje przerwana.
- Odtworzenie sygnału za pomocą biblioteki `sounddevice`.

Wady i ograniczenia

Operacja odtwarzania nagrania posiada pewne ograniczenia:

- Odtwarzanie może być ograniczone przez zasoby sprzętowe lub konflikty z innymi aplikacjami wykorzystującymi urządzenie audio.
- Biblioteka `sounddevice` wymaga zainstalowanego środowiska wspierającego odpowiednie sterowniki audio.

3 Wnioski

Na podstawie wyników i implementacji można wyciągnąć następujące wnioski:

3.1 Problemy, które wystąpiły

W trakcie implementacji pojawiły się trudności z precyzyjnym określeniem progów dla określania bezdźwięczności i odróżniania muzyki od mowy oraz rozmiarem ramki, co mogło wpływać na dokładność wyników.

3.2 Czy metoda zawsze działa dobrze?

Metody wykrywania mowy, muzyki i ciszy działają dobrze w większości przypadków, ale mogą mieć trudności w sytuacjach, gdzie występuje złożone tło dźwiękowe lub zmienne dynamiki sygnału.

Dla czystych nagrań mowy i muzyki, metody działają zgodnie z oczekiwaniami, jednak dla nagrań z szumem, wyniki mogą być mniej dokładne.

3.3 Dlaczego są różne wersje metod? Czy któraś jest lepsza?

W aplikacji zastosowano różne wersje wyliczania częstotliwości tonu podstawowego (F0), zarówno oparte na autokorelacji, jak i AMDF. Obie metody mają swoje zalety i wady – autokorelacja działa lepiej w przypadku wyraźniejszych tonów, natomiast AMDF sprawdza się lepiej w bardziej szumowych sygnałach. W zależności od rodzaju sygnału, jedna metoda może być bardziej odpowiednia od drugiej.

4 Podsumowanie

Podsumowując, aplikacja stanowi skuteczne narzędzie do analizy sygnałów audio i może być rozwijana o dodatkowe funkcjonalności, takie jak lepsze algorytmy detekcji ciszy, adaptacyjne dostosowanie progów analizy, czy poprawienie dokładności detekcji w przypadku szumów.