

# Cechy sygnału audio w dziedzinie czasu

## Projekt nr 1 - Analiza i Przetwarzanie Dźwięku

<https://github.com/szymon159/sound-time-params-evaluation>

Szymon Stasiak

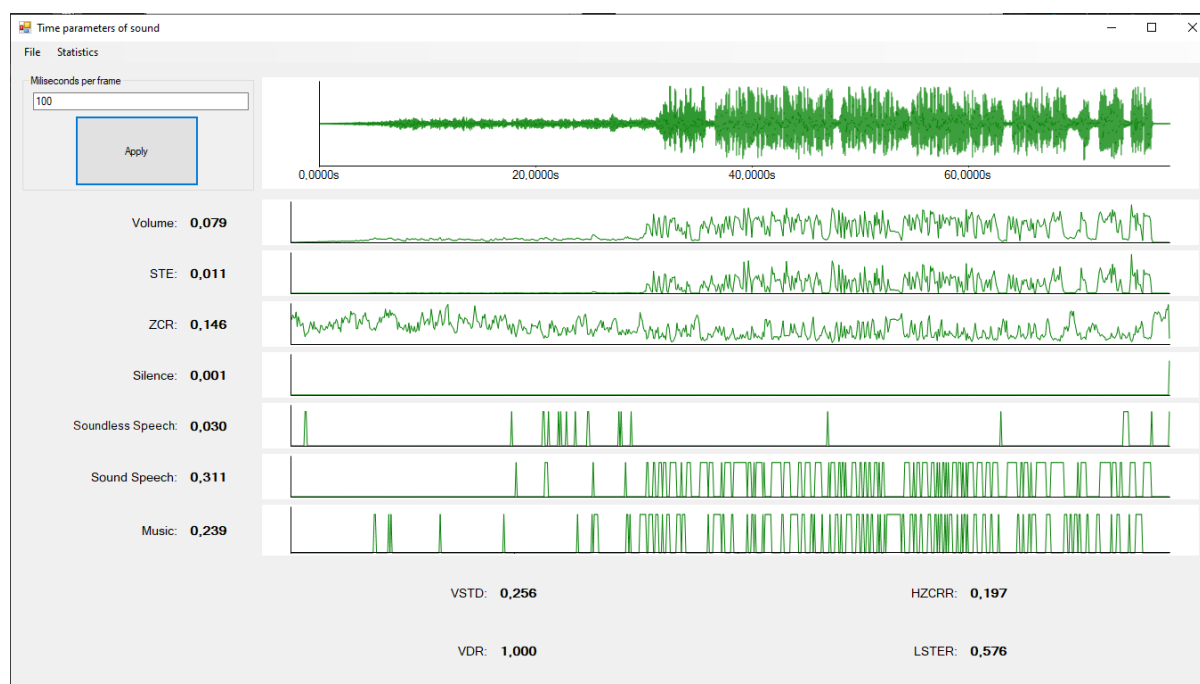
## 1 Opis aplikacji

Projekt stworzony został w .NET Framework, z warstwą prezentacji powstałą w Windows Forms i interfejsem w języku angielskim.

W celu uproszczenia i zoptymalizowania wczytywania i parsowania plików (\*.wav oraz \*.mp3), projekt korzysta także z jednej dodatkowej paczki nugetowej - NAudio.

Jako nagrania testowe posłużyły udostępnione transkrypcje (*Rysunki 2 i 3*), a ponadto: utwór *Portier!* Taco Hemingwaya (jako przykład rozmowy z tłem muzycznym) - *Rysunki 1, 4, 5*, fragment przemówienia Baracka Obamy (jako przykład mowy, pobrany ze strony <http://soundbible.com/>) - *Rysunek 6* oraz fragment utworu *Adventure* stworzony przez Bensound (jako przykład muzyki instrumentalnej, udostępniony za darmo na stronie <https://www.bensound.com/royalty-free-music/>) - *Rysunek 7*.

### 1.1 Główne okno programu



Rysunek 1: Interfejs programu

Parametry obok wykresów oznaczają średnią z wartości dla wszystkich ramek nagrania.

### 1.2 Statystyki

Przycisk *Statistics* w menu na górze okna otwiera menu kontekstowe pozwalające dla każdego dostępnego typu statystyki (wspierane: *cisza*, *mowa bezdźwięczna*, *mowa dźwięczna*, *muzyka*) wyświetlić lub wyeksportować statystyki.

W przypadku decyzji o wyświetleniu, w pojawiającym się okienku modalnym, znajduje się lista interwałów, w których dane zjawisko zostało zaobserwowane oraz ich długość.

W przypadku wybrania eksportu, pojawia się okno zapisu pliku. W zapisanym pliku *\*.csv* każda linia oznacza jeden interwał. Numer interwału, czas początku, czas końca i długość są oddzielone przecinkiem. Dodatkowo pierwsza linia zawiera opisy kolumn.

## 2 Opis metod

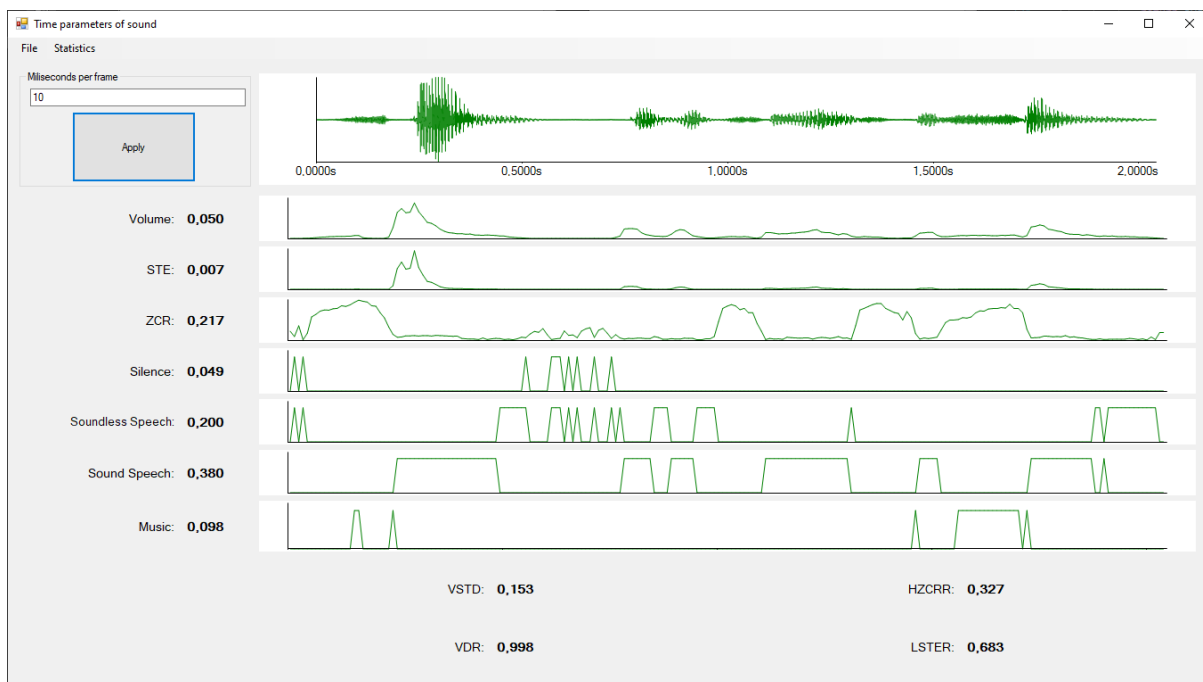
### 2.1 Cechy na poziomie ramki (Frame-level)

Ilość ramek wyliczana jest na podstawie podanej wartości milisekund na ramkę (*Milliseconds per frame*, lewy górny róg) oraz pobranej przy wczytywaniu pliku częstotliwości próbkowania.

Na wszystkich wykresach wartość dla ramki zaznaczona jest w jej środku. W przypadku nierównego podziału, ostatnia ramka może być dużo krótsza niż pozostałe (jednak punkt wciąż zaznaczony jest w jej środku).

We wszystkich wzorach z tego rozdziału zachowane zostało następujące znaczenie symboli:  $n$  - kolejny indeks ramki,  $M$  - długość ramki w próbkach,  $s_n(i)$  - amplituda  $i$ -tej próbki w  $n$ -tej ramce,  $sign(x)$  - funkcja signum.

Ponadto poniższe omówienie wszystkich cech może zawierać odwołania do przykładowych nagrań (zarówno nieznormalizowanego jak i znormalizowanego). W celu uproszczenia odbioru, *Rysunek 2* oraz *Rysunek 3* prezentują jedno przykładowe nagranie w obu tych wariantach.



Rysunek 2: Nienormalizowane nagranie



Rysunek 3: Znormalizowane nagranie

### 2.1.1 Głośność (Volume)

Wartość głośności dla każdej klatki wyliczana jest zgodnie ze wzorem dostępnym w materiale źródłowym, czyli:

$$Volume(n) = \sqrt{\frac{1}{M} \sum_{i=0}^{M-1} s_n^2(i)} \quad (1)$$

Głośność jest wartością blisko powiązaną z amplitudą fali, co widać na wszystkich powyższych rysunkach. Dotyczy to zarówno formy znormalizowanej jak i nieznormalizowanej.

Zgodnie z oczekiwaniami, różnice między wykresami głośności dla obu tych wariantów ograniczają się do zakresu wartości. Widać to chociażby po mniejszej wartości średniej dla nagrania znormalizowanego. Różnica ta zależy od charakterystyki mikrofonu, którą proces normalizacji ma na celu zniwelować.

### 2.1.2 Energia (Short Time Energy)

Energia dla danej ramki jest niczym innym niż wartością głośności podniesioną do kwadratu:

$$STE(n) = \frac{1}{M} \sum_{i=0}^{M-1} s_n^2(i) = (Volume(n))^2 \quad (2)$$

Wartości energii, są niższe od wartości głośności. Jest to oczywiste biorąc pod uwagę fakt iż wartość bezwzględna amplitudy dla nagrania (zarówno znormalizowanego jak i nieznormalizowanego), jest mniejsza od 1.

### 2.1.3 Liczba przejść przez zero (Zero Crossing Rate)

Jak sama nazwa wskazuje, współczynnik ten określa liczbę przejść amplitudy przez zero w jednej ramce sygnału. Można go określić korzystając z następującego wzoru:

$$ZCR(n) = \frac{1}{2M} \sum_{i=0}^{M-1} |sign(s_n(i)) - sign(s_n(i-1))| \quad (3)$$

Zgodnie z oczekiwaniami - wartość ta jest jednakowa dla nagrania nieznormalizowanego i znormalizowanego.

Powyższe współczynniki są pomocne przy określaniu również innych parametrów dźwięku.

W przypadku kolejnych trzech parametrów, obliczenia opierają się na klasyfikacji danej ramki ze względu na wartości już obliczonych parametrów i w zależności od wyniku, nadaniu określanemu parametrowi wartości 0 lub 1.

#### 2.1.4 Cisza (Silent Ratio)

Miara określana na podstawie głośności i ZCR. Ramki o dostatecznie małych wartościach tych parametrów, można zaklasyfikować jako ciszę. W przypadku tego projektu, wartościami granicznymi okazały się 0.005 dla głośności oraz 0.1 dla ZCR.

#### 2.1.5 Mowa bezdźwięczna (Soundless Speech)

Miara określana na podstawie energii i ZCR. Aby ramka została zaklasyfikowana jako zawierająca mowę bezdźwięczną, energia oraz ZCR muszą być dostatecznie małe. Jest to zatem parametr określany podobnie do ciszy, w tym przypadku jednak należy zastosować inne progi.

Energia (która odróżnia mowę bezdźwięczną od ciszy) nie powinna przekraczać 0.001. Próg dla ZCR pozostaje niezmieniony i równy 0.1.

#### 2.1.6 Mowa dźwięczna (Sound Speech)

Określanie tej miary jest analogiczne do mowy bezdźwięcznej. Jedyną różnicą jest fakt iż mowa dźwięczna przenosi większą energię niż bezdźwięczna - stąd też w tym przypadku wartość energii powinna przekraczać określoną granicę 0.001.

#### 2.1.7 Muzyka (Music)

Istnieje wiele sposobów detekcji muzyki w nagraniu audio. Określanie jej na podstawie wyliczonych w tym projekcie parametrów jest możliwe, jednak obarczone dużym błędem. Zgodnie z koncepcją przedstawioną w opracowaniu dostępnym pod adresem <https://www.sciencedirect.com/science/article/pii/S101836391830850X>, muzyka różni się od mowy dźwięcznej wyższą wartością parametru ZCR. W przypadku tego projektu, za muzykę uznane są klatki dla których energia przekracza 0.001 a ZCR przekracza 0.1.

### 2.2 Cechy na poziomie klipu (Clip-level)

Określanie cech na poziomie ramki nie może dostarczyć wszystkich informacji - bierze ono pod uwagę krótki wycinek (ramkę) pliku audio. Dlatego też ogólna charakterystyka nagrania musi być opisywana innymi parametrami. Służą do tego parametry określone na poziomie klipu, wyliczane z użyciem opisanych powyżej parametrów Frame-level. W tej sekcji zostały opisane wszystkie cechy określone w projekcie, jednak można ich zdefiniować znacznie więcej.

#### 2.2.1 Znormalizowane odchylenie standardowe głośności (Volume Standard Deviation - VSTD)

Normalizacja zawarta w nazwie tego parametru polega na podzieleniu odchylenia standardowego głośności przez maksymalną głośność w klipie. Wartość parametru jest zatem dana wzorem:

$$VSTD = \frac{1}{Volume_{max}} \sqrt{\frac{\sum_{n=0}^{N-1} (Volume(n) - Volume_{avg})^2}{N}}. \quad (4)$$

W powyższym wzorze  $N$  oznacza liczbę ramek w nagraniu,  $Volume_{max}$  to najwyższa wartość głośności w nagraniu zaś  $Volume_{avg}$  to średnia wartość głośności w nagraniu.

#### 2.2.2 Rozpiętość tonalna (Volume Dynamic Range - VDR)

Różnica między najcichszym i najgłośniejszym fragmentem nagrania.

$$VDR = \frac{Volume_{max} - Volume_{min}}{Volume_{max}} \quad (5)$$

Dzięki normalizacji różnicy głośności przez wartość maksymalną, wartość VDR zazwyczaj jest zbliżona do 1. W przypadku sygnałów ciągłych o stałej wartości, zbliżona jest do 0.

### 2.2.3 High Zero Crossing Rate Ratio (HZCRR)

Jest to jeden ze współczynników bazujących na ZCR, pozwalających odróżnić nagranie zawierające muzykę od zawierającego mowę. Jego wartość obliczana jest według wzoru:

$$HZCRR = \frac{1}{2N} \sum_{n=0}^{N-1} [\text{sign}(ZCR(n) - 1.5 * \text{avg}_{ZCR}) + 1] \quad (6)$$

Symbol  $\text{avg}_{ZCR}$  w powyższym wzorze oznacza średnią wartość ZCR w 1-sekundowym oknie (a więc  $\text{avg}_{ZCR}$  jest dla danego nagrania niezależne od zadanej długości ramki).

### 2.2.4 Low Short Time Energy Ratio (LSTER)

Odsetek liczby ramek w których STE są mniejsze niż 50% średniej energii w 1-sekundowym oknie.

$$LSTER = \frac{1}{2N} \sum_{n=0}^{N-1} [\text{sign}(0.5 * \text{avg}_{STE} - STE(n)) + 1] \quad (7)$$

Symbol  $\text{avg}_{STE}$  oznacza średnią wartość energii w oknie 1-sekundowym.

Wartość współczynnika LSTER dla mowy jest zazwyczaj znacznie wyższa niż dla muzyki co pozwala z dużą skutecznością odróżnić muzykę od mowy.

## 3 Wyniki działania

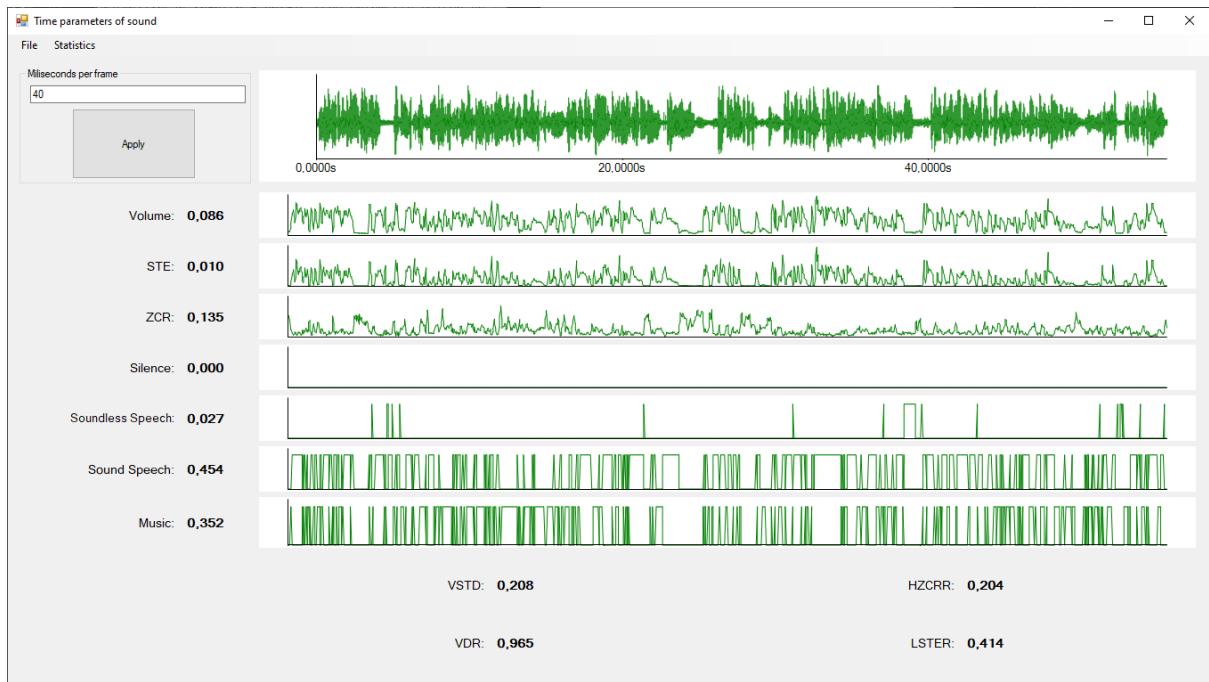
Szczególnie istotnymi aspektami projektu, na których warto się skupić jest odróżnianie muzyki od mowy (oraz podział mowy na dźwięczną i bezdźwięczną).

W tym celu już na etapie implementacji należało poświęcić dużo czasu na odpowiednie dobranie parametrów granicznej głośności, energii czy liczby przejść przez zero. W tym celu posłużono się dostępnymi już programami i starano się dostosować wyniki zwracane przez tworzony program do rezultatów poprawnych. Należy jednak pamiętać, że równie ważne (albo i ważniejsze) jest zachowanie zdrowego rozsądku - odróżnienie muzyki od mowy za pomocą słuchu jest prostsze i efektywniejsze.

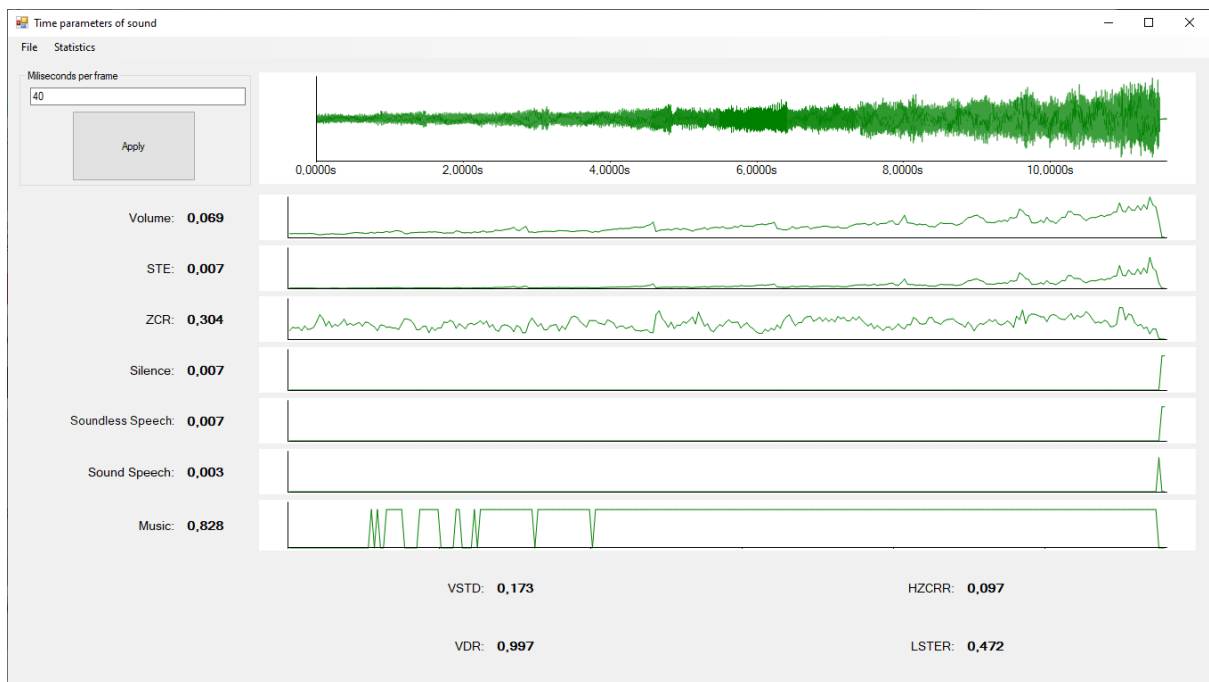
Testowymi nagraniami były już wspomniane nagrania korpusu tekstowego, jednak z powodu braku muzyki analizie zostały poddane również inne pliki źródłowe.

### 3.1 Mowa z podkładem muzycznym

Przykładem jest już wspomniany utwór *Portier!* Taco Hemingwaya. Na poniższych zdjęciach fragment rozmowy z muzyką w tle (*Rysunek 4*) oraz samej narastającej muzyki (*Rysunek 5*).



Rysunek 4: Fragment rozmowy z utworu *Portier!* Taco Hemingwaya

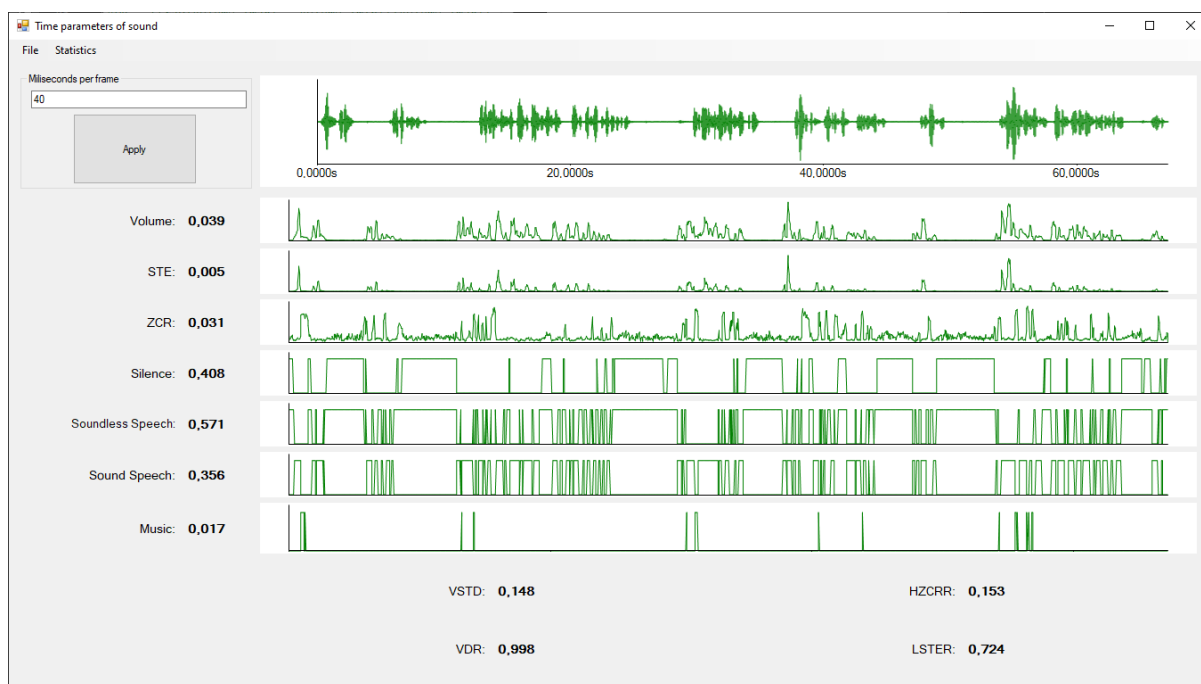


Rysunek 5: Fragment muzyki z utworu *Portier!* Taco Hemingwaya

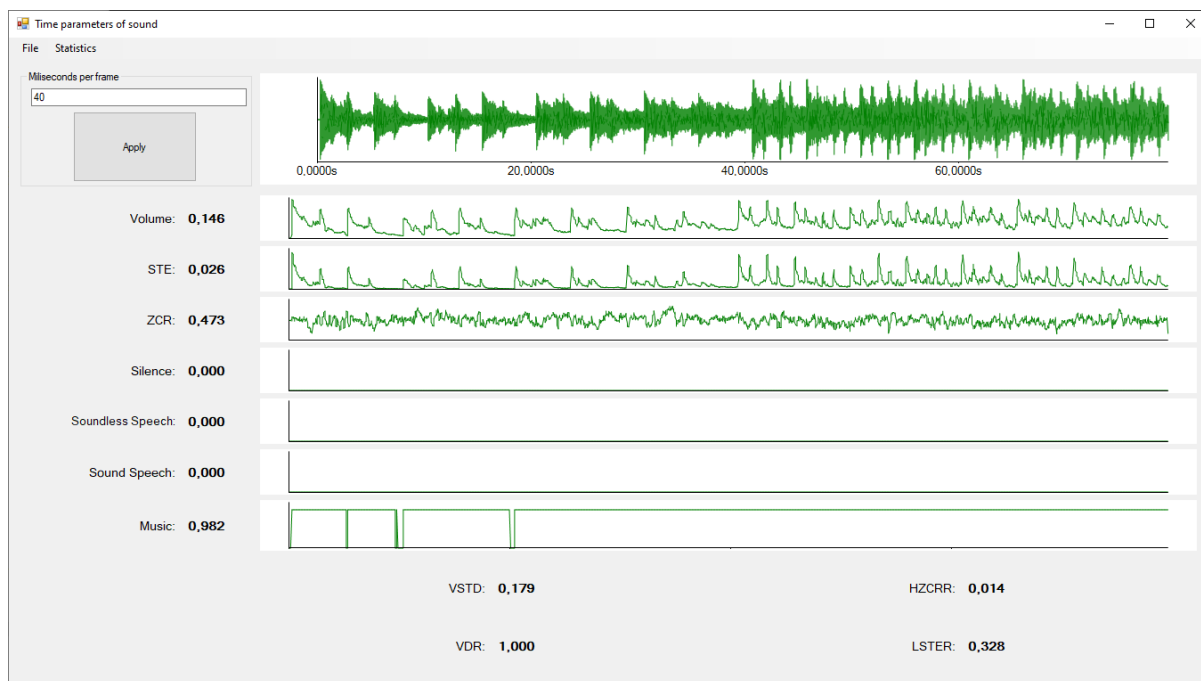
Jak widać na tym przykładzie, parametry Frame-level okazały się zawodne, a muzyka będąca podkładem do słów wyraźnie wybijała się w trakcie nagrania rozmowy. Niestety oznacza to, iż ciężko wyciągnąć wnioski analizując utwory zawierające mowę przeplataną z muzyką. Potwierdzać zdają się to parametry Clip-level. Przykładowo LSTER, który powinien być dla mowy znacznie wyższy, jest niewiele niższy.

### 3.2 Mowa a muzyka

Sytuacja prezentuje się inaczej przy osobnej analizie nagrania mowy oraz muzyki. W tym celu powołałam się ponownie na zapowiadany we wstępie fragment przemówienia Baracka Obamy (*Rysunek 6*) oraz utwór instrumentalny (*Rysunek 7*).



Rysunek 6: Fragment przemówienia Baracka Obamy



Rysunek 7: Fragment muzyki z utworu *Adventure* stworzonego przez Bensound

Jak widać, w przypadku przemówienia ilość wykrytej muzyki jest niewielka, w szczególności w porównaniu z mową. Znaczna część przemówienia określona jest jednak jako cisza. Absolutnie nie jest to jednak błędem! Wystarczy spojrzeć na wykres amplitudy wczytanego nagrania na którym znajduje się bardzo

dużo przerw, co można potwierdzić odsłuchując faktyczne przemówienie.

Natomiast dla utworu instrumentalnego, ilość ciszy oraz mowy jest zerowa, mimo występowania fragmentów o niskiej głośności czy energii. Ponadto niemal cały utwór został zaklasyfikowany jako muzyka, co pozwala wysoko ocenić skuteczność parametrów Frame-level.

Dodatkowo wartość LSTER również pokrywa się z oczekiwaniami - wartość 0.724 dla przemówienia może zostać spokojnie zaklasyfikowana jako dużo wyższa niż 0.328 dla muzyki.

## 4 Wnioski

Parametry określone przez program są wystarczająco dokładne aby odróżnić nagranie muzyki od nagrania mowy. Jednak jeśli występują one jednocześnie, dokładność nie pozwala na określenie przedziałów gdzie muzyka/mowa dominuje.