Cechy sygnału audio w dziedzinie częstotliwości

Projekt nr 3 - Analiza i Przetwarzanie Dźwięku

https://github.com/szymon159/sound-analysis

Szymon Stasiak

1 Opis aplikacji

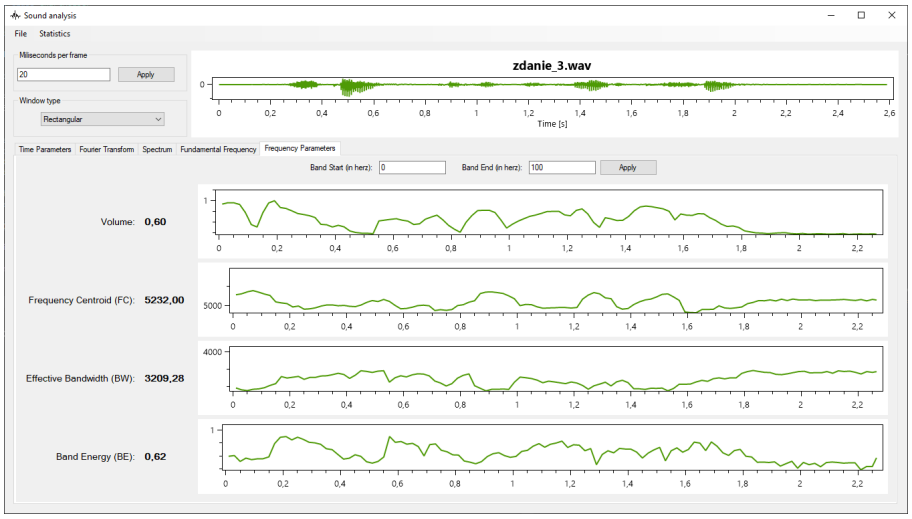
Projekt stworzony został w .NET Framework, z warstwą prezentacji powstałą w Windows Forms i interfejsem w języku angielskim. Analiza wspiera pliki **.wav* oraz **.mp3*.
Jest to rozwinięcie projektów *Cechy sygnału audio w dziedzinie czasu* oraz *Analiza częstotliwościowa dźwięku*, dodające do poprzednich wersji nową opcje widoku - parametry w dziedzinie częstotliwości **Frequency Parameters** oraz pola **Band Start** i **Band End** wykorzystywane w liczeniu *Band Energy*
W celu uproszczenia obliczeń używane są również zewnętrzne paczki (żadnych zmian w stosunku do poprzedniego projektu):

- *NAudio* - w celu wczytywania, parsowania i podstawowych operacji na plikach
- *MathNet.Numerics* - dostarcza narzędzia do obliczania transformaty Fouriera oraz stosowania funkcji okienkowych
- *Oxyplot* - usprawnienie wykresów

Jako nagrania testowe posłużyły nagrania stworzone na potrzeby przedmiotu (*Rysunki 1, 2 i 3*), a ponadto: fragment przemówienia Baracka Obamy (jako przykład mowy, pobrany ze strony `http://soundbible.com/`) - *Rysunki 4* oraz *6* oraz fragment utworu *Adventure* stworzony przez Bensound (jako przykład muzyki instrumentalnej udostępniony za darmo na stronie `https://www.bensound.com/royalty-free-music/`) - *Rysunki 5* oraz *7*.

1.1 Głowne okno programu

Główne okno programu uległo nieznacznej zmianie w stosunku do poprzedniego projektu. Dodana została zakładka *Frequency Parameters* z nowymi parametrami, pozostałe zakładki wciąż przedstawiają parametry obliczone w poprzednich projektach.



Rysunek 1: Interfejs programu

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

2 Opis metod

2.1 Cechy na poziomie ramki (Frame-level)

Ilość ramek wyliczana jest na podstawie podanej wartości milisekund na ramkę (*Miliseconds 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).
W odróżnieniu od obliczeń w dziedzinie czasu, punktem wyjścia do analizy częstotliwościowej jest wynik transformaty Fouriera. Z tego powodu możliwa jest również zmiana okna używanego przy stosowaniu transformaty (co z kolei wpływa na wyniki).
We wszystkich wzorach z tego rozdziału zachowane zostało następujące znaczenie symbol: n - kolejny indeks ramki, N - długość ramki w próbkach, $S_n(k)$ - wartość dla k-tej (obliczonej z pomocą transformaty Fouriera) częstotliwości w n-tej ramce.
Ponadto poniższe omówienie wszystkich cech zawiera 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: Nienznormalizowane 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:

$$Volume(n) = \frac{1}{N} \sum_{k=0}^{N-1} S_n^2(k) \tag{1}$$

Podobnie jak w przypadku analizy w dziedzinie czasu, widoczny jest wpływ normalizacji na głośność nagrania. W analizowanym przypadku, nagranie znormalizowane jest głośniejsze od nieznormalizowanego. Widać to jednak wyłącznie poprzez wartość średnią, gdyż sam kształt wykresu dla obu wersji jest niemal identyczny. Nieznaczne różnice wynikają częściowo z błędów zaokrągleń spowodowanych wielokrotną transformacją wczytanych wartości (w szczególności zastosowaną transformatą Fouriera).

2.1.2 Centrum częstotliwości (Frequency Centroid)

Analiza komputerowa w dziedzinie częstotliwości jest analizą dyskretną. Z tego też powodu we wzorze służącym do obliczenia analizowanego parametru całka została zastąpiona sumą (a więc miara analityczna zastąpiona miarą numeryczną). Powstały wzór ma zatem postać:

$$FC(n) = \frac{\sum_{k=0}^{N-1} k * S_n(k)}{\sum_{k=0}^{N-1} S_n(k)} \tag{2}$$

Wartość tego parametru jest używana do odróżnienia mowy od ciszy oraz muzyki od mowy. Wyniki przedstawione są w następnej sekcji *3 Wyniki działania*.

2.1.3 Efektywne pasmo (Effective Bandwidth)

Podobnie jak w przypadku poprzedniego parametru, także w przypadku pasma należy przekształcić wzory do postaci dyskretnej. Dodatkowo warto odnotować iż wartość tego parametru w każdej ramce jest zależna od wyliczonego uprzednio centrum (oznaczonego przez $FC(n)$). Po podstawieniu odpowiednich symboli, używany wzór ma postać:

$$BW(n) = \sqrt{\frac{\sum_{k=0}^{N-1} [(k - FC(n)) * S_n(k)]^2}{\sum_{k=0}^{N-1} S_n^2(k)}} \tag{3}$$

Zastosowanie wyliczonej wartości pasma jest zbliżone do centroidu - tak jak on pozwala odróżniać mowę od muzyki oraz ciszę od mowy.

2.1.4 Energia (Band Energy)

Ostatnim z analizowanych parametrów jest energia określonego pasma częstotliwości. W tym przypadku, do analizy dochodzą dwa dodatkowe parametry f_0 oraz f_1 - odpowiednio oznaczające początek i koniec zakresu pasma. Są one konfigurowalne jako liczby całkowite z poziomu interfejsu aplikacji. Po dodaniu tych wartości i przejściu do postaci dyskretnej otrzymany wzór ma postać:

$$BE_{[f_0,f_1]}(n) = \frac{\sum_{k=k_0}^{k_1} S_n^2(k)}{\sum_{l=0}^{N-1} w(l)} \tag{4}$$

We wzorze pojawiają się jeszcze trzy nieopisane wcześniej symbole: k_0, k_1 oraz $w(l)$. Pierwsze dwa oznaczają indeksy próbek którym odpowiadają częstotliwości pasma: odpowiednio f_0 oraz f_1. Z kolei $w(l)$ oznacza wartość l-tej próbki wczytanego nagrania (przed przejściem do dziedziny częstotliwości).
Wartość energii również może być z powodzeniem używana w detekcji mowa-muzyka.

3 Wyniki działania

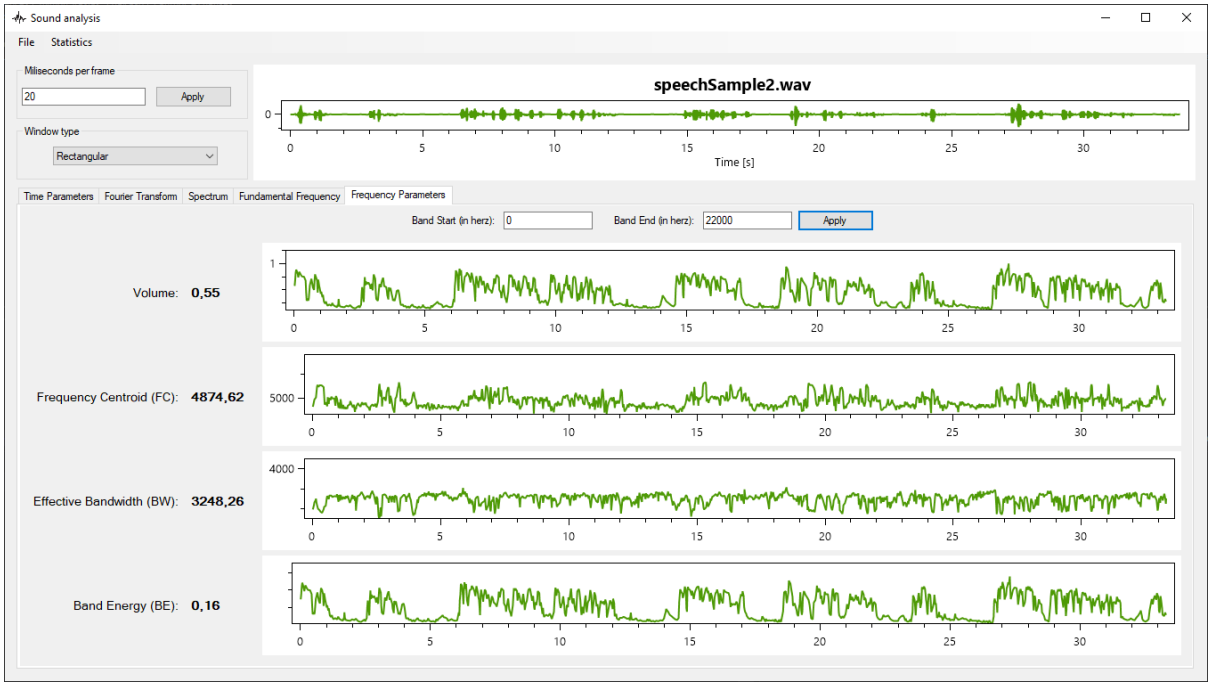
Szczególnie istotnymi aspektami projektu, na których warto się skupić jest, podobnie jak w przypadku parametrów w dziedzinie czasu, odróżnianie muzyki od mowy.
Testowymi nagraniami były już wspomniane nagrania korpusu tekstowego, jednak z powodu braku muzyki analizie zostały poddane również inne pliki źródłowe.

3.1 Normalizacja nagrań

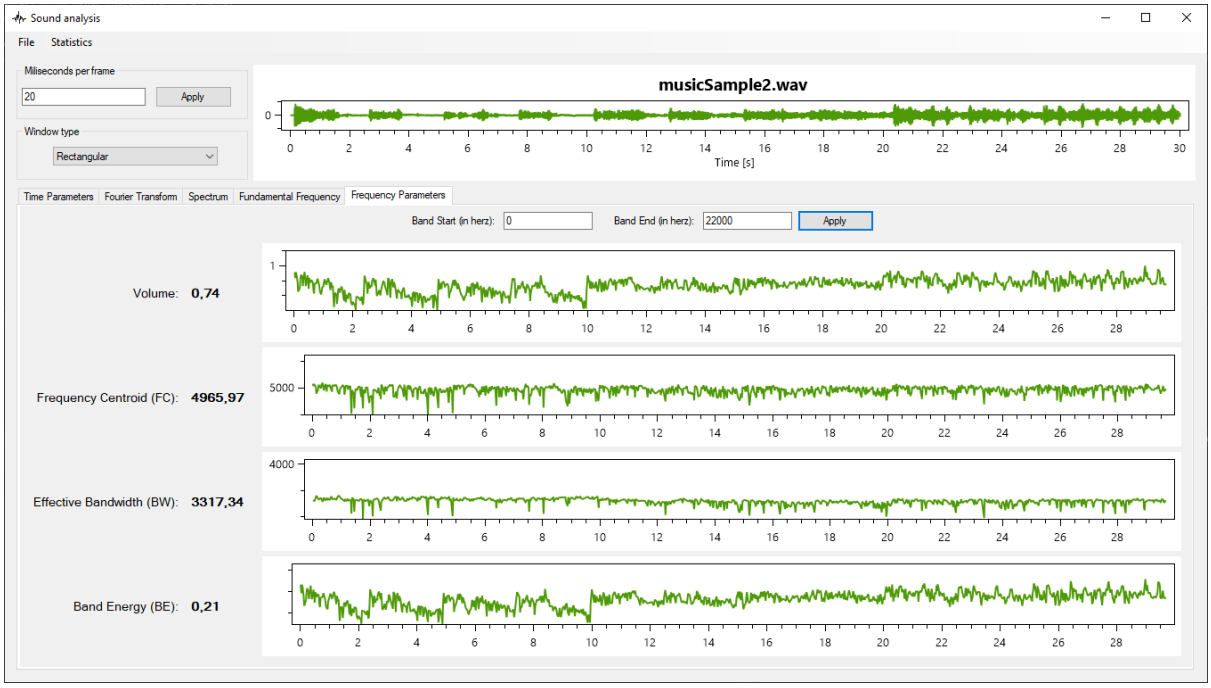
Na samym początku warto poświęcić chwilę analizie różnic między nagraniami nieznormalizowanymi a znormalizowanymi. Jako przykład może tutaj posłużyć porównanie *Rysunku 2* oraz *Rysunku 3* wspomnianych w sekcji *2 Opis metod*. Delikatne różnice między parametrami są widoczne szczególnie w zakresie wartości. Często wynika to z kumulacji błędów zaokrągleń - wartości znacznie mniejsze od 0 zostają poddane transformacie Fouriera a następnie wielokrotnie wykonywane są na nich obliczenia takie jak pierwiastek czy mnożenie.
Warto jednak zauważyć, że dla nagrania nieznormalizowanego wyłącznie głośność osiąga wartości mniejsze niż dla znormalizowanego. W przypadku pozostałych parametrów sytuacja jest odwrotna - wartości większe osiągane są przez nagranie znormalizowane.

3.2 Mowa a muzyka

Niestety nagrania stworzone na potrzeby przedmiotu nie dają możliwości porównania mowy i muzyki. Z tego powodu, analizie poddane zostaną: zapowiedziany we wstępie fragment przemówienia Baracka Obamy (*Rysunek 4* i *6*) oraz utwór instrumentalny (*Rysunek 5* i *7*).
Niestety, wyniki przedstawionej na następnej stronie analizy są rozczarowujące. Można było oczekiwać iż wartości *frequency centroid* oraz *effective bandwidth* będą znacznie mniejsze dla mowy niż dla muzyki, tym czasem wartości te są niemal równe.

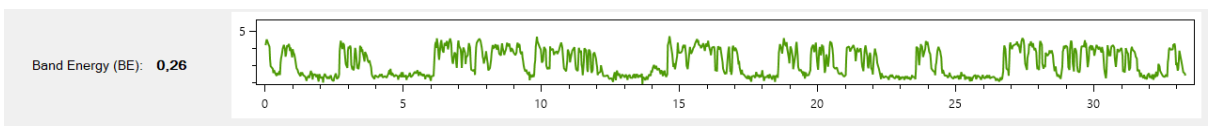


Rysunek 4: Fragment przemówienia Baracka Obamy

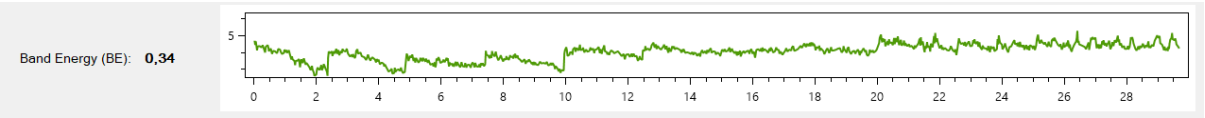


Rysunek 5: Fragment muzyki z utworu *Adventure* stworzonego przez Bensound

Sytuacja wygląda nieco lepiej dla analizy parametru *band energy*. W tym przypadku, zgodnie z oczekiwaniami, wartość dla muzyki jest większa niż dla mowy. Dodatkowo energia w pasmie $0 - 500\ Hz$, a więc charakterystycznym dla mowy, jest wyższa niż dla całego pasma.



Rysunek 6: Energia w pasmie $0 - 500\ Hz$ dla nagrania mowy



Rysunek 7: Energia w pasmie $0 - 500\,Hz$ dla nagrania muzyki

Kolejny aspekt przy którym wyznaczone parametry mogą okazać się pomocne, to odróżnianie fragmentów udźwiękowionych od ciszy. W tym przypadku efekty również są względnie zadowalające - dla przykładowej mowy fragmenty ciszy pokrywają się z niskimi wartościami parametrów *frequency centroid* czy *band energy*.

4 Wnioski

Parametry określane przez program niestety nie są wystarczająco dokładne aby jednoznacznie odróżnić na ich podstawie nagranie mowy i muzyki. Co prawda subtelne różnice między takimi nagraniami są widoczne, jednak wartości pozostawiają w tej kwestii wiele do życzenia. Z przedstawionej analizy wynika, iż najlepszym parametrem do tego typu klasyfikacji jest *band energy*, jednak różnice są zdecydowanie za małe aby metoda ta była uznana za bezbłędną czy nawet godną zaufania.
Pozytywnie można jednak ocenić odróżnianie fragmentów udźwiękowionych od ciszy, co jest jednak niewielkim pocieszeniem gdyż istnieje wiele prostszych, szybszych, a przede wszystkim skuteczniejszych metod.