Politechnika Warszawska

WYDZIAŁ MATEMATYKI
I NAUK INFORMACYJNYCH

Analiza i przetwarzanie dźwięku

Cechy sygnału audio w dziedzinie czasu

Mikołaj Kida

Marzec 2022

1 Opis aplikacji

1.1 Wybór języka i frameworku

Aplikacja została wykonana w języku R przy użyciu framworku Shiny. Wybór ten został zmotywowany łatwości pracy z wektorami jaką oferuje R oraz prostotą tworzenia przyjemnej dla oka i w pełni responsywnej aplikacji graficznej za sprawą Shiny.

1.2 Biblioteki

Aplikacja korzysta z następujących bibliotek

- shiny
- dplyr
- DT
- ggplot2
- shinydashboard
- plotly
- tuneR

1.3 Interfejs graficzny

Aplikacja składa się z dwóch zakładek *Upload Data* oraz *Parameters*. Zakładka *Upload Data* pozwala na załadowania danych audio w formacie .wav, natomiast *Parameters* odpowiada za główną logikę programu.

W zakładce *Parameters* możemy wybrać które plik audio chcemy aktualnie analizować oraz zmienić wielkość klatek na jakie dzielimy nasz sygnał. Zakładka ta po załadowaniu danych daje nam podgląd do parametrów danego klipu, czy rozważanej klatki w postaci wykresów i wartości numerycznych.

2 Opis metod

Do analizy klipów audio użyto szeregu cech dzięki którym można je scharakteryzować. Cechy te dzielą się na krótkookresowe (na poziomie ramki) i długookresowe (na poziomie klipu). W tym projekcie skorzystano z następujących cech

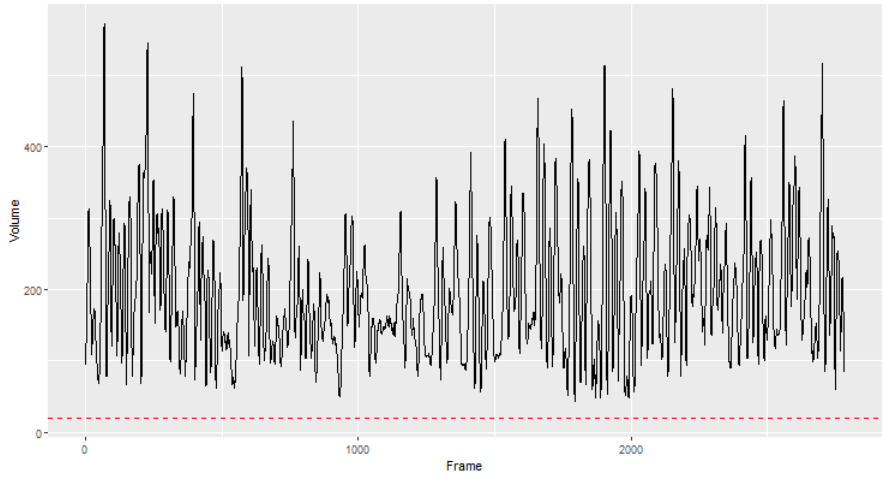
2.1 Cechy sygnału audio w dziedzinie czasu na poziomie ramki (Frame-Level)

1) Głośność (Volume) Miara ta wprowadza średnią głośność sygnału audio dla krótkich fragmentów i jest bardzo efektywna w detekcji ciszy i określania granic klipów audio.

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

Implementacja w kodzie

```
calculate_volume <- function(samples) {
     round(sqrt((1/length(samples))*sum(samples**2)), digits = 2)
}
```



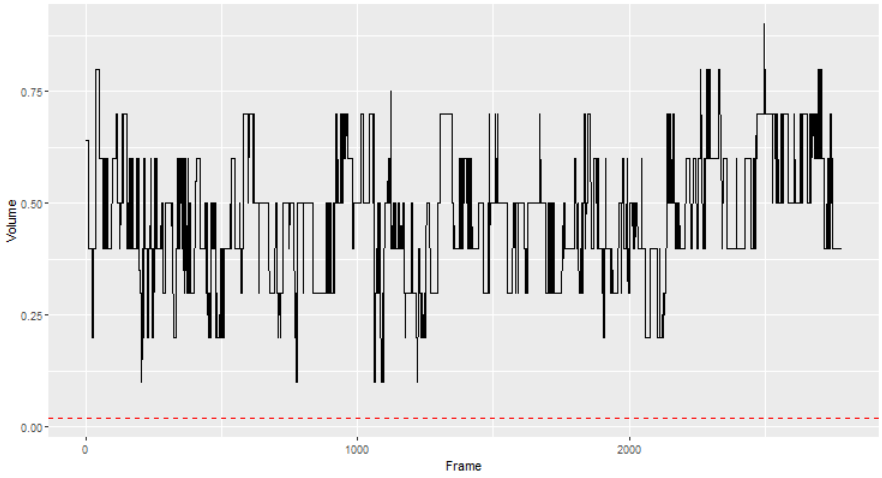
Rysunek 1: Wykres głośności dla przykładowego klipu

2) ZCR (Zero Crossing Rate) Liczba przejść amplitudy przez zero w jednej ramce sygnału audio.

$$Z(n) = \frac{1}{2N}\left(\sum_{i=0}^{N-1} |sgn(s_n(i)) - sgn(s_n(i-1))|\right)$$

Implementacja w kodzie

```
calculate_ZCR <- function(samples, f) {
      round(((1)/(2*length(samples)))*sum(abs(diff(sign(samples)))),
      digits = 2)
}
```



Rysunek 2: Wykres ZCR dla przykładowego klipu

3) SR (Silent Ratio) Miara ta jest wyliczana na podstawie dwóch poprzenich, głośności oraz ZCR. Gdy głośność oraz ZCR spadną poniżej pewnego progu (tu 0.02 dla głośności oraz 50 dla ZCR) to kwalifikujemy ramkę jako ciszę.

$$SR(n) = \mathbb{1}_{\{p(n)>0.02,\ ZCR(n)>50\}}$$

Implementacja w kodzie

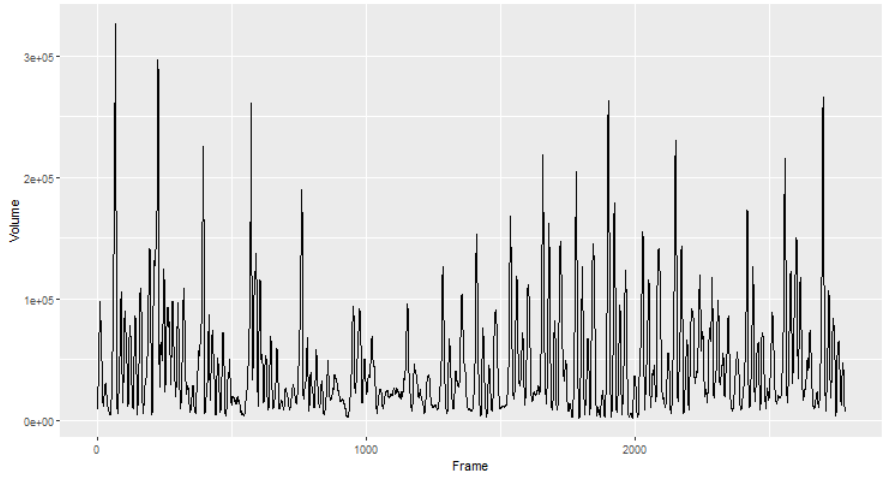
```
calculate_SR <- function(samples, f, volume_threshold, zcr_threshold)
{
      (calculate_volume(samples) > volume_threshold) |
      (calculate_ZCR(samples, f) > zcr_threshold)
}
```

4) STE (Short Time Energy) Kolejna użyteczna miara to energia, która definiowana jest jako kwadrat z głośności sygnału.

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

Implementacja w kodzie

```
calculate_STE <- function(samples) {
     round((1/length(samples))*sum(samples**2), digits = 2)
}
```



Rysunek 3: Wykres energii dla przykładowego klipu

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

1) VSTD Jest to odchylenie standardowe głośności normalizowane przez głośność w całym klipie.

$$VSTD = \frac{\sigma}{max(p(n))}$$

Implementacja w kodzie

```
calculate_VSTD <- function(samples, number_of_frames) {
     round(suppressWarnings(sd(samples)/
     mean(unlist(lapply(split(samples, seq(1,number_of_frames)),
     FUN = calculate_volume)))), digits = 2) }
```

2) VDR Jest to stosunek zakresu głośności do maksymalnej głośności w klipie.

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

Implementacja w kodzie

```
calculate_VDR <- function(samples, number_of_frames) {
     volume <- suppressWarnings(unlist(lapply(split(samples, seq(1,number_of_frames)),
     FUN = calculate_volume)))
     round(1 - (min(volume))/(max(volume)), digits = 2) }
```

3) LSTER Odsetek liczby ramek w których STE są mniejsze niż 50% średniej STE w oknie jendo sekundowym.

$$LSTER = \frac{1}{2N} \sum_{n=0}^{N-1} \left[sgn(0.5\overline{STE} - STE(n)) + 1 \right]$$

Implementacja w kodzie

```
calculate_LSTER <- function(samples, number_of_frames) {
     ste_vector <- suppressWarnings(unlist(lapply(split(samples,
     seq(1, number_of_frames)),
     FUN = calculate_STE)))
     round(((1)/(2*number_of_frames))*
     (sum(sign(0.5*mean(ste_vector) - ste_vector) + 1)), digits = 2)
}
```

4) Entropia Cecha bazująca na energii. Użyteczna w detekcji klipów audio zawierających "wybuchowe" sceny, takie jak przemoc czy walka.

$$I = -\sum_{i=1}^{J} \sigma_i^2 log_2 \sigma_i^2$$

Implementacja w kodzie

```
calculate_Entropy <- function(samples, number_of_frames) {
     volume <- suppressWarnings(unlist(lapply(split(samples, seq(1,number_of_frames)),
     FUN = calculate_volume)))
     round(1 - (min(volume))/(max(volume)), digits = 2) }
```

5) HZCRR (High Zero Crossing Rate Ratio) Modyfikacja miary ZCR definiowana dla całego klipu.

$$HZCRR = \frac{1}{2N} \sum_{n=0}^{N-1} \left[sgn(ZCR - 1.5\overline{ZCR}) + 1 \right]$$

Implementacja w kodzie

```
calculate_HZCRR <- function(samples, number_of_frames) {
     zcr_vector <- suppressWarnings(unlist(lapply(split(samples,
     seq(1, number_of_frames)),FUN = calculate_ZCR)))
     round(((1)/(2*number_of_frames))*
     (sum(sign(zcr_vector - 1.5*mean(zcr_vector)) + 1)), digits = 2)
}
```

6) Funkcja autokorelacji Złóży do estymacji czasowej częstotliwości tonu podstawowego.

$$R_n(l) = \sum_{i=0}^{N-l-1} s_n(i) s_n(i+l)$$

3 Wyniki i wnioski

3.1 Testy metod

Aby zaprezentować wyniki działania posłużono się 3 klipami audio. Dwa z nich to klipy muzyczne, natomiast trzeci to nagranie lektorskie. Dla wszystkich nagrań porównano parametry opisane w poprzedniej sekcji.

ZCR w kombinacji z **głośnością** sprawdziły się zadziwiająco dobrze jak na swoją prostotę wykrywając początkową i końcową ciszę w klipach muzycznych oraz wykrywając większość przerw w mowie lektora. (Przy progach 0.02 dla głośności oraz 50 dla ZCR)

W przypadku **entropii** klipy muzyczne wykazywały się większą jej wartością od klipu z lektorem.