Ekstrakcja cech sygnałów dźwiękowych i synteza mowy

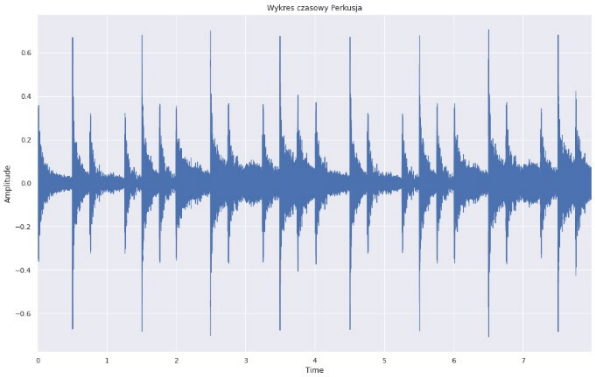
Jakub Robaczewski

Ekstrakcja cech sygnałów dźwiękowych:

Wybrałem plik audio „Perkusja”.

Zero-crossing rate

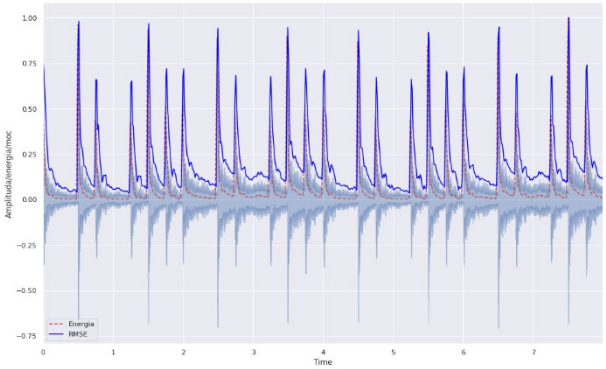
Szybkość, z jaką sygnał zmienia się z dodatniego na ujemny lub z ujemnego na dodatni.

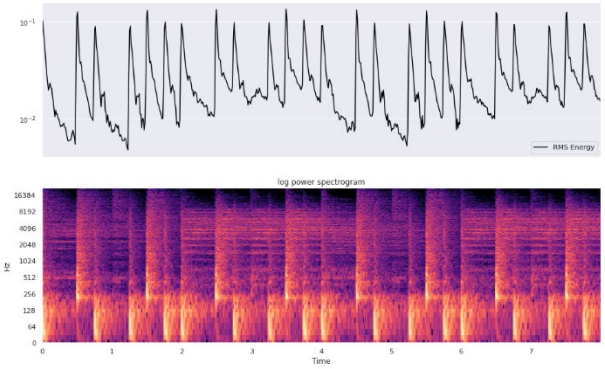


Zauważamy wysokie skoki w momentach, gdzie uderzamy w perkusję.

Parametry energetyczne

RMSE – średnia kwadratowa energii sygnału

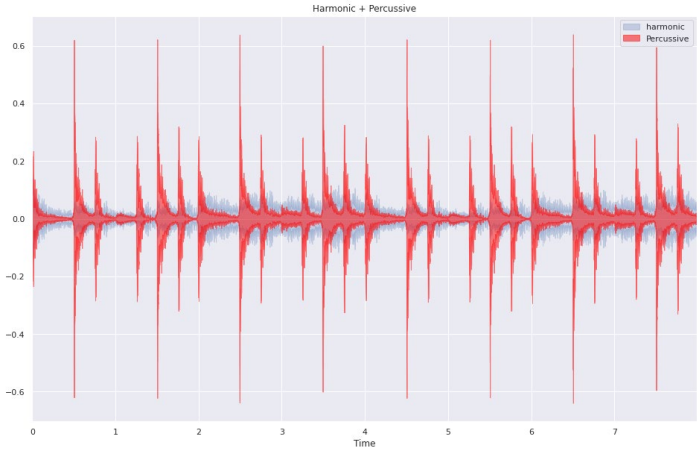




Energia nagrania jest powiązana z uderzeniami w perkusje i właśnie wtedy pojawiają się największe skoki.

Separacja części perkusyjnej i harmonicznej

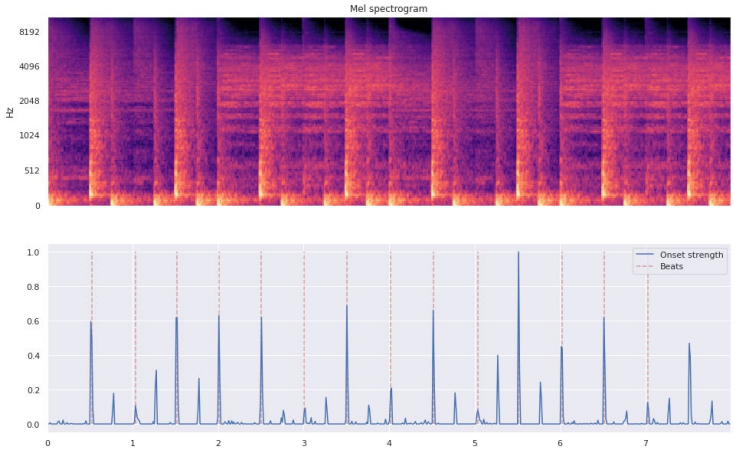
Rozdziela dźwięk na część perkusyjną i harmoniczną.



Głównym elementem nagrania perkusji jest część perkusyjna, ale możemy też zauważyć niewielką część hermoniczną.

Detekcja tempa

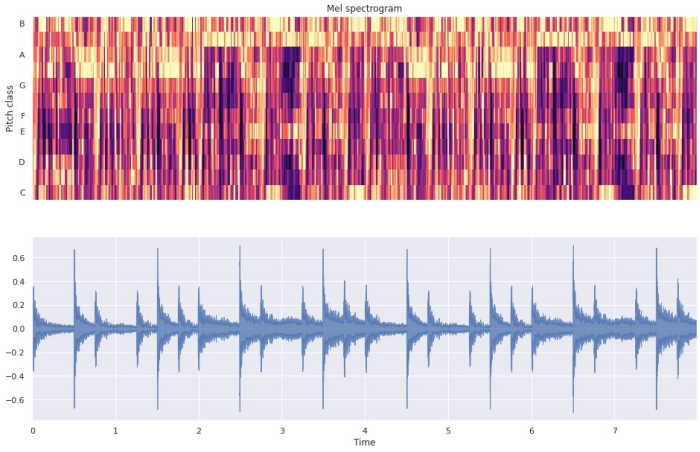
Wykrywa bit w nagraniu.



Zgodnie z przewidywaniami możemy zauważyć powtarzalne sekwencje powiązane z uderzeniami w odpowiednie instrumenty perkusyjne.

Chromogram

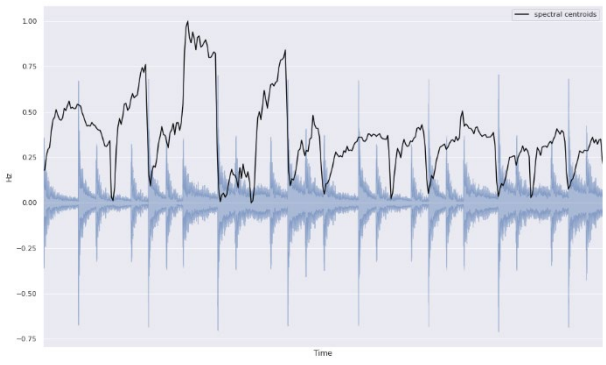
Rozdziela nagranie na tony, którymi posługują się ludzie przy tworzeniu muzyki.

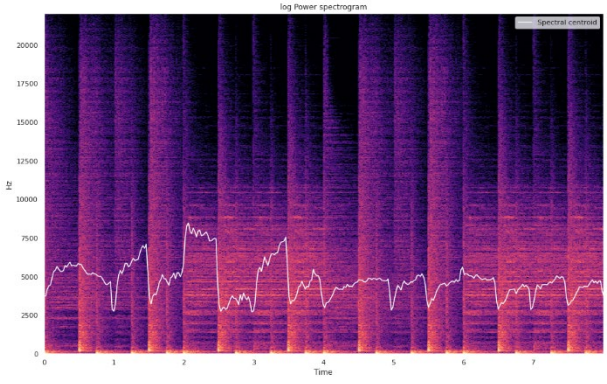


Możemy zauważyć, że nagranie składa się z powtarzających się zestawów tonów (E, F), (A, G) oraz (D, C).

Spectral centroid

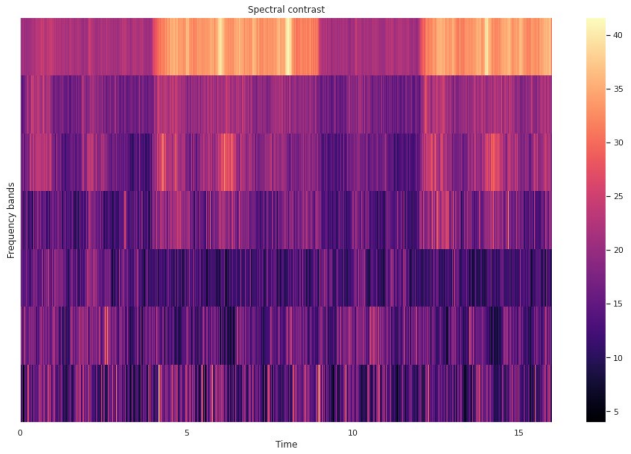
„Środek ciężkości” widma, odwzorowuje jasność brzmienia.





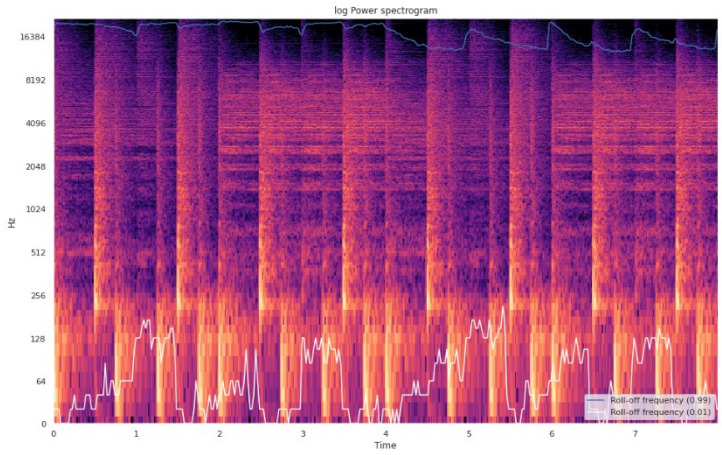
Spectral contrast

Kontrast energii wyliczany przez porównanie średniej energii w górnym kwantylu do dolnego kwantyla. Wysoki kontrast oznacza, że dźwięk występuje w szybkich „pikach”.



Spectral rolloff

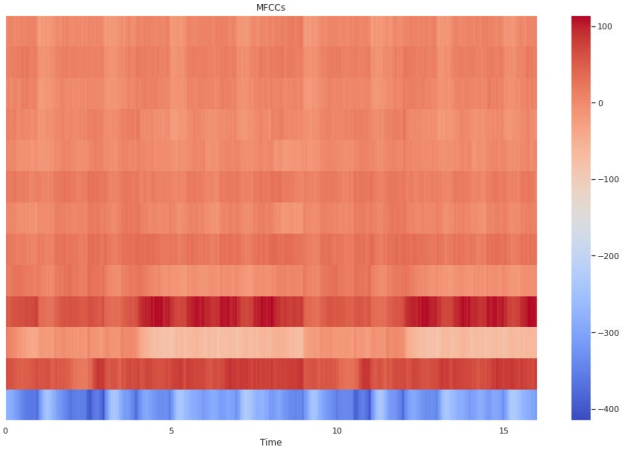
Częstotliwość poniżej której znajduje się określony procent całej energii, w tym przypadku 1%.



Możemy zauważyć, że w dźwięku przeważają wysokie dźwięki, a częstotliwości poniżej 128 Hz stanowią około 1%.

MFCC

Wektor parametrów, zawierających centrum sygnału przedstawione w skali melowej, mających odzwierciedlać naturalną odpowiedź układu słuchowego na pobudzenie dźwiękami mowy.



Synteza mowy - TTS:

Voice Cloning

Metoda ta daje zrozumiały tekst, jednak jest on częściowo zaszumiony i niewyraźny. Zasadniczą zaletą tego rozwiązania jest częściowa zdolność do naśladowania głosu, który został mu podany (w warunkach tego laboratorium to jedyna metoda, która pozwala uzyskać męski głos).

Google TTS

Dobra metoda odwzorowania głosu, jednak zdarza się jej dziwnie i nienaturalnie akcentować wyrazy w dłuższych zdaniach.

Mozilla TTS

Metoda porównywalna do Google TTS, jednak wydaje się lepiej radzić z dłuższymi wyrażeniami.

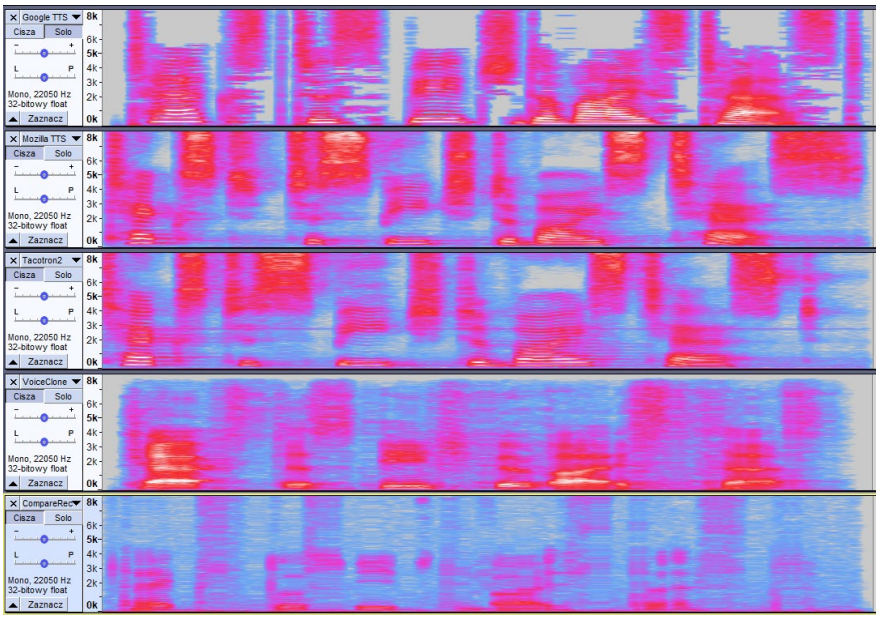
Tacotron2 + Waveglow

Wydaje się dawać najczystszy głos z podanych metod. Chociaż zmiana ta jest niewielka w porównaniu do algorytmów Google TTS i Mozilla TTS.

Synteza mowy - analiza spektrogramów

Przed rozpoczęciem analizy spektrogramu tempo wszystkie nagrania zostały wyrównane i docięte. Nagranie porównawcze (początkowo w wersji stereo) zostało znormalizowane do 1 kanału. Następnie wszystkie nagrania zostały przepróbowane do próbkowania 22050 Hz, które było używane przez większość nagrań. Do badań wykorzystałem nagrania otrzymane w poprzednim kroku („Text speech device test”).

Analizując spektrogramy możemy zauważyć, że nagranie porównawcze jest dużo mniej wyraźne od nagrań syntezowanych, wynika to z naturalnej ludzkiej dynamiki głosu. Możemy również zauważyć, że syntezowane nagrania (oprócz VoiceClone) charakteryzują się większą ilością wyższych składowych. Wynika to z tego, że są do nich wykorzystywane głosy kobiece, które mają więcej wyższych składowych niż męskie. Analizując spektrogramy bardzo łatwo jesteśmy w stanie odróżnić stosunkowo podobne do siebie syntetyzowane głosy od prawdziwego ludzkiego głosu.



Rozpoznawanie mowy - STT:

Oryginalny tekst	May the Force be with you
Sphinx	Made the phone spewing fuel
Google	**May the force be with you**
Mozilla	Made the fort beau

Jedynie algorytm Googla potrafił prawidłowo przetworzyć nagranie na tekst, pozostałe zwracały błędne wyniki (w różnym stopniu). Algorytm Sphinxa działa częściowo poprawnie, jednak popełnia błędy w podobnie brzmiących wyrazach (Made->May), zaś algorytm Mozilli zwraca słowa, które zupełnie nie pasują do użytego nagrania.